Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

Curated and Asymmetric Exposure: A Case Study of Partisan Talk during COVID on Twitter

Zijian An, Jessica Breuhaus, Jason Niu, A. Erdem Sariyuce, Kenneth Joseph

Computer Science and Engineering Department, University at Buffalo, Buffalo, NY, USA
zijianan, jlbreuha, jasonniu, erdem, kjoseph@buffalo.edu

Abstract

Social media has been at the center of discussions about political polarization in the United States. However, scholars are actively debating both the scale of political polarization online, and how important online polarization is to the offline world. One question at the center of this debate is what interactions across parties look like online, and in particular 1) whether increasing the number of such interactions is likely to increase or reduce polarization, and 2) what technological affordances may make it more likely that these cross-party interactions benefit, rather than detract from, existing political challenges. The present work aims to provide insights into the latter; that is, we focus on providing a better understanding of how a set of 400,000 partisan users on a particular social media platform, Twitter, used the platform's affordances to interact within and across parties in a large dataset of tweets about COVID in 2021. Our findings suggest that Republican use of cross-party interaction were both more potent and potentially more strategic during COVID, that cross-party interaction was driven heavily by a small set of users and conversations, and that there exist non-obvious *indirect* pathways to cross-party exposure when different modes of interaction are chained together (especially retweets of quotes). These findings have implications beyond Twitter, we believe, in understanding how affordances of platforms can help to shape partisan exposure and interaction.

Introduction

One of the most well-established findings in computational social science is that Americans online are polarized along partisan lines. Studies have shown, for example, that partisan divides exist on Twitter (Conover et al. 2011), and that these trends are increasing over time (Garimella and Weber 2017). Other work has exposed similar patterns on Facebook (Bakshy, Messing, and Adamic 2015), on reddit (Guimaraes and Weikum 2021), and in web search data (Robertson et al. 2023). However, more recent work has suggested that this polarization may be restricted to a small set of highly active users (Weeks et al. 2017), and that some operationalizations of polarization are more empirically evident than others (Fraxanet et al. 2023). An ongoing debate thus exists over the extent of political polarization online in the U.S. (González-Bailón and Lelkes 2023).

Nearly all scholars would acknowledge, however, that *some* interaction exists across partisan lines. Indeed, even in early work on polarized debates online (Yardi and Boyd 2010), significant attention was paid to interactions across the partisan divide, where debate, discussion, and/or vitriol can emerge across parties. Others have similarly shown how interaction across partisan lines can emerge as online social movements co-opt (Jackson and Foucault Welles 2015a) and are co-opted by (Gallagher et al. 2018) counter-movements, how cross-party dialogue emerges in replies to particular tweets (Shugars and Beauchamp 2019), and the impacts of cross-party exposure on political attitudes (Bail et al. 2018). While cross-partisan talk may be limited relative to within-party interactions, these previous works show that careful study of them can provide insights into the dynamics of polarization online (Balietti et al. 2021).

More specifically, prior work has suggested the importance of considering both what we will call *direct* and *indirect* cross-party interaction. The majority of existent literature has focused on how (a lack of) *direct* cross-party interaction—e.g. a Republican (not) retweeting a Democrat—can shape affective polarization at the individual level (Bail et al. 2018) and the formation of "echo chambers" at the structural level (Garimella et al. 2018). Other work, such as the literature on counterpublics (Jackson and Foucault Welles 2015b) and theories of curated information flows (Thorson and Wells 2016), however, have instead noted that engagement through *indirect* cross-party interaction—e.g. a Republican seeing a reply that another Republican has sent to a Democrat—can also function to reshape interparty attitudes. But little is understood about the relative volume of direct versus indirect cross-party interaction, nor how these various forms of interactions are shaped by networks of elite actors. There is, consequently, a need for a deeper empirical investigation of these points.

To this end, the present work aims to characterize and understand direct and indirect partisan interactions, and the network structures underlying them, in a dataset of COVID-related tweets from roughly 400,000 Twitter users during 2021. While our work is thus a case study focusing on a single politicized setting, prior work has shown that COVID presents a critical political arena in which polarized discussions played out with significant implications for human life and social policy (Xue et al. 2020; Jiang et al. 2021; Muric,

Wu, and Ferrara 2021; Mønsted and Lehmann 2022), and thus is in and of itself a useful lens into online polarization and its consequences. We also are able to draw significant parallels to work in other domains, strengthening both our claims and those in the prior work. Of particular interest to us in this case study are 1) how users select particular technological affordances to engage in cross-party interaction and 2) how these decisions, when aggregated in different ways, expose patterns of polarization and partisan engagement. Specific to Twitter, our analysis focuses on differential uses and combinations of the quote, reply, and retweet functions, and how these decisions at the individual level result in different aggregate patterns of interaction.

To characterize the ways in which Twitter users engage in cross-party content, we develop a level-based formulation of social interaction on Twitter.In our formulation, at the first level (Level 0) are original tweets. Here, we refer to these original tweets as *OP*s as shorthand for original post. At the second level (Level 1) are any direct interactions with 1) the OP, or—because of the way in which data is returned by the Twitter API—2) any retweet of the OP. Beyond this are any interactions—retweet, quote, or reply— with Level 1 posts, and then recursively to the end of the interaction chain. Each OP thus constitutes its own starting point, or what we will refer to[1] as a single *conversation*.

Using this leveled formulation, we address the following three research questions:

- **RQ1:** How do Twitter users leverage the retweet, reply, and quote affordances of Twitter to interact across, relative to within, party at Level 1 of conversations (i.e. in *direct* response to OPs)?
- **RQ2:** (How) is cross-party talk amplified via downstream engagement (perhaps *indirectly*) beyond Level 1?
- **RQ3:** What can we learn about the *core* of partisan users who interact heavily both within and across party lines?

Methodologically, our work leverages two network-based methods. The first, which helps us to address all three of our questions, uses the retweet network to identify two distinct groups of users separated along partisan lines (Darwish et al. 2020). The second, which we use to address RQ3, is a recent method that finds cohesively polarized pair(s) of communities in a given signed network (Niu and Sarıyüce 2023).

Substantively, our work contributes several novel insights to the literature on polarization and cross-party talk online. With respect to RQ1, prior work has focused on both replies (Shugars and Beauchamp 2019; Hada et al. 2023) and quotes (Lorentzen 2020) individually. In recent work, Zade et al. (2023) has also focused on identifying motivational differences in the use of within versus cross-party replies and quotes. However, no work has yet explored how cross-party engagement differs in prevalence or topical content, across replies, quotes, and retweets. With respect to RQ2, our work is the first to show that, at least in the context of COVID, *a predominant vehicle through which users interacted with cross-party content was through a co-partisan lens*. This finding is important because it shows that nearly half of cross-party interactions occur not directly between individuals across parties, but rather indirectly through a co-partisan lenses. With respect to RQ3, we show that the core of the partisan interaction network during COVID, i.e. the users who interacted heavily both within *and* across party, were qualitatively distinct across parties both in who they were and how they interacted.

Our work therefore provides new empirical evidence for existing theories that demand a focus on an assymetric, indirect cross-party interactions, and the corresponding need for interventions that are sensitive to the affordances of modern social media platforms (Zade et al. 2023) and the many forms of highly curated content interactions they produce (Thorson and Wells 2016), as well as to the behaviors of a narrow set of antagonistic users that vary qualitatively across party lines (Guess et al. 2018). More narrowly, our work offers three advancements in our understanding of interaction dynamics and polarization on Twitter:

1. We show that in a large dataset on a topic that highlighted partisan differences in the U.S., replies were the predominant form of direct cross-party interaction in response to (retweets of) original tweets, and that Republicans interacted across party at higher rates than Democrats.
2. We show, however, that when including Level 2 interactions, *retweets of quotes* account for nearly a third of all observable cross-party interactions in our dataset, and in total nearly half of cross-party exposure occurs indirectly through a co-partisan frame.
3. Finally, we identify a core of less than 100 users who have both dense within-party and cross-party ties, and use it to highlight the asymmetric structure of partisan talk during COVID between the political left and right.

Related Work

Substantively, our work ties to literature on the study of political interaction and polarization online, and in particular to patterns in cross-party engagement. Methodologically, our work ties to the literature on measuring user ideology and identifying patterns in signed networks. Here, we briefly characterize connections to these literatures in separately.

Technological Affordances and (De)Polarized Interactions A vast (e.g. Yardi and Boyd 2010; Garimella et al. 2018; Barberá et al. 2015; Demszky et al. 2019) literature has explored patterns in polarized political discussions online. Much of this work focuses on empirically characterizing polarization via direct interactions within and across party, although more recent work has focused on expansions of theory (Kreiss and McGregor 2023; Törnberg 2022) and method (Fraxanet et al. 2023; Hada et al. 2023). Despite current debates, work in this area finds that direct interactions occur mostly between co-partisans. A second and growing consensus is that direct interactions across parties do not always lead to depolarization (Bail et al. 2018), in part because they are often hostile (Marchal 2022).

Of particular relevance to our work is the role that platform affordances play in cross-party interaction. Specific

[1] with slight abuse of the official use of the phrase from Twitter, who do not count retweets and quotes as part of a conversation, instead only replies

to the communicative affordances of Twitter, scholars have predominantly looked at how retweets, quotes, and replies are used in political settings. While some are adversarial (Guerra et al. 2017), retweets are generally understood to signal support (Metaxas et al. 2015; Joseph et al. 2019). Indeed, while we use retweets as a marker of support, we also observe a limited number of cross-party retweets that do not appear to take this form.

Most work on replies focuses on patterns of discussion that occur within threads associated with particular conversations. This work has assessed questions of exposure to different viewpoints (Hada et al. 2023), structural characteristics of the resulting conversation (Nishi et al. 2016; Cogan et al. 2012), and factors associated with user engagement (Shugars and Beauchamp 2019). Perhaps most notably here, this work suggests that reply threads are a source of cross-party interaction (Shugars and Beauchamp 2019), that interactions have the potential for both hostility and genuine dialogue (Lorentzen 2020), and that these interactions can be asymmetrical, with one side (the political right) more deeply engaging with the other in a hostile fashion (Hada et al. 2023). Finally, scholars looking at quote tweets have found that they act more like replies than retweets, serving as an indirect mechnism of cross-party exposure (Lai et al. 2019). However, relative to replies, quotes have the unique function of exposing the quoted tweet to a broader range of users, or what Gallagher (2022) calls an amplification effect.

Two recent works have focused on differences between replies and quotes. Garimella, Weber, and De Choudhury (2016) explore differences between quote tweets and replies and find that quote tweets likely to diffuse content more widely than replies. We extend this work by considering whether this holds across partisan lines, by looking into content differences between cross-party replies and quotes, and by proposing a level-based framework for interpreting these indirect exposures to cross-party content. More recently, Zade et al. (2023) conduct a qualitative study and introduce a novel codebook to help explain how Twitter users leverage quotes versus replies within versus across party lines. Pertinent to the present work, they find that cross-party quote tweets often aim to engage a broader audience by focusing on the topic of the OP, but twisting the words of the OP to emphasize a distinct viewpoint. In contrast, replies often directly engage the user posting the OP, but do so by shifting to a different topic. The present work complements the efforts of Zade et al. (2023) by looking at the prevalence of cross-party interactions (as opposed to just their contents), and by emphasizing the important role of elites.

Finally, because our case study focuses on COVID specifically, it is worth noting that a growing literature explores the use of Twitter during the pandemic (e.g. Xue et al. 2020; Jiang et al. 2021; Muric, Wu, and Ferrara 2021; Mønsted and Lehmann 2022). Of particular relevance, Crupi et al. (2022) used similar methods to those studied here and found that the Italian vaccine debate on Twitter remained highly polarized between pro-vaccine and anti-vaccine/hesitant groups throughout the COVID-19 pandemic. Their work and others highlight the multifaceted nature of Twitter discussions regarding COVID-19, emphasizing the crucial role of political ideologies in shaping public sentiment and perceptions about the virus and its vaccine. We build on this work but explore a novel component of the discussion process.

Detecting the Ideology of Twitter Users An extended literature considers the potentials and pitfalls of identifying the political ideology of Twitter users using behavioral data (Cohen and Ruths 2013; Barberá et al. 2015). A number of methods have been developed, but most common approaches rely on assessment of a user's social relationships (Volkova, Coppersmith, and Van Durme 2014; Demszky et al. 2019). Many of these works focus on the partisan leaning of accounts that a user follows, and using this as a proxy to assign the user's own partisan label. However, use of the follower network is prohibitive for large datasets. Previous research thus suggests that using retweets as a metric can yield comparable results with less resource investment (Magdy et al. 2016). More specifically, Darwish et al. (2020) proposed an unsupervised stance detection framework for identifying ideological stances on Twitter. The methodology employs dimensionality reduction techniques, in particular Uniform Manifold Approximation and Projection (UMAP) (McInnes et al. 2018), to project Twitter users into a low-dimensional space based on their retweet patterns. Subsequent clustering, using algorithms such as HDBSCAN, helps categorize these users into distinct ideological groups.

We have adopted the Unsupervised User Stance Detection approach outlined by Darwish et al. (2020) to define user ideology and extend it with various dimension reduction methodologies to ensure optimal results. Our research, however, also builds upon the work from Darwish et al. (2020) in our approach to validation of the clustering. Specifically, we integrate a pair of clustering approaches, leveraging their unique strengths and then identifying partisan users based on shared patterns across the two methods.

Finding polarized groups in signed networks Signed networks serve as a valuable tool for representing both positive and negative interactions, such as relationships of friendship versus enmity and trust versus distrust (Heider 1946; Cartwright and Harary 1956). One traditional method for identifying polarized groups within signed networks is through the concept of *balance*, which gauges stability based on the arrangement of positive and negative connections. Heider's definition of a balanced signed graph states that it should exhibit balance in all of its cycles, where a cycle is considered positive if it contains an even number of negative edges (Heider 1946). To assess partial balance, a common approach involves calculating the fractions of balanced triangles ($+++$ and $+--$) within the network (Aref and Wilson 2018; Cartwright and Harary 1956). In the context of polarized groups, one would expect that nodes within the same group are positively connected, while nodes from different groups are negatively connected. Recent research has suggested that identifying balanced subgraphs can serve as a useful proxy for discovering polarized communities within signed networks (Bonchi et al. 2019; Ordozgoiti, Matakos, and Gionis 2020; Tzeng, Ordozgoiti, and Gionis 2020; Xiao, Ordozgoiti, and Gionis 2020). However, a common drawback found in these studies is their reliance on a poorly-

L1→L0	QTs	RPs	RTs	Total
D→D	1,907,660	1,219,583	23,303,901	26,431,144
R→R	1,467,636	927,287	19,105,720	21,500,643
D→R	157,367	237,716	230,614	625,697
R→D	418,280	477,637	546,706	1,442,623

Table 1: Amount of Level 1 tweets for the different interaction types (QT = quotes, RP = replies, RT= retweets) for each combination of partisanship of Level 1/Level 0 users (e.g. D→D is for a Democrat QT/RP/RT of a Level 0 post from a Democrat).

defined metric, namely polarity, leading to the emergence of extensive subgraphs lacking a distinct sense of agreement or conflict (Niu and Sarıyüce 2023). This issue primarily stems from the fact that in real-world signed networks, triangles with all positive edges (+++) are notably more prevalent than those with two positive and one negative edge (+−−). Consequently, the dominance of +++ triangles in the resulting balanced subgraphs hinders capturing conflicts. Niu and Sarıyüce (2023) recently proposed the electron decomposition algorithm to remedy this issue (explained in detail in Section). The main idea is to find dense subgraphs with respect to the balanced triangles while watching out for the unbalanced triangles. In this work, we use electron decomposition to find polarized pairs of communities in the signed network of Twitter users with known political affiliation.

Data & Methods

Data Collection

This study begins with a dataset of 255,114,554 tweets from 18,237,593 users sent between March, 2021 and October, 2021. Tweets were collected using the Twitter v1.1 Streaming API using a small set of generic keywords relevant to the COVID-pandemic (covid, covid19, etc.) and a number of keywords to capture tweets related to vaccines (vaccine, vax) and specific vaccines (Johnson & Johnson, AstraZenica, etc.). The present work focuses on a subset of 392,165 highly active users that we can confidently identify a partisan affiliation for (details below). Our original dataset contains 177,119,306 (69.4%) retweets, 17,058,938 (6.7%) quote tweets, 37,331,819 (14.6%) replies, and 2,360,4494 (9.3%) original tweets. The highly active users we study here account for a substantial portion of these data—76,361,530 tweets, representing 30% of the total dataset, including 4,475,607 quotes (5.9%), and 6,907,777 replies (9.0%). While our dataset compromises a small proportion of all users in the dataset (only 2%), it therefore covers a proportionally larger sample of the content. For later reference, Table 1 also provides the subset of these tweets that are Level 1 interactions (as opposed to Level 0 or Level 2).

Given the focus of the present work on cross-party interactions between users, and in particular how these occur via replies and quotes, it is critical that we acknowledge the limitations inherent in studying these types of interactions via keyword-based samples. With respect to replies, keyword-based samples are limited in the extent to which we can observe chains of replies emanating from original tweets (Lorentzen and Nolin 2017). More specifically, if User A sends a tweet containing a keyword, and User B replies with a tweet that does not, a keyword-based dataset will not contain User B's reply. In contrast, the dataset will contain User B's response if it is a quote even if that quote does not contain the relevant keyword, as long as the OP does (Zade et al. 2023). To assess the limitations of our dataset in this context, we therefore collect an auxilliary dataset that we use to assess bias due to API limitations.

To construct our auxilliary dataset, we first sampled 1860 conversation IDs that had at least one reply and one quote from an account we labeled to be Democrat or Republican. We then used the Twitter v2 Historical Archive to collect all related quotes and Level 1 replies that were not deleted as of June, 2023. The resulting set of 3,285,348 replies and quotes, along with their metadata, were used to assess the possibility that our main results could be biased due to sampling biases, especially in the differences that exist between replies and quotes. Our robustness check involves a series of regressions and replication analyses; these are detailed in the Appendix in the section entitled *Analysis of bias due to missing data*. Overall, our robustness evaluation surfaces no reason to expect that the core findings of this paper are directly impacted by this sampling bias. However, where biases do exist that alter our results (in ways that do not impact the core substantive claims of the paper), we state so explicitly in the Results section of the main text.

Identifying Partisan Clusters of Users

Our approach to identifying Twitter users with left- or right-partisan leanings proceeds in a series of steps. First, as done in prior work (Darwish et al. 2020; Zhang et al. 2018), we filter the retweet network down to a more active set of users in order to obtain a more accurate clustering. We filter out from all users from the rows of the retweet matrix (the users we will cluster) anyone who sent less than ten retweets, and from the columns (used as the feature vector for clustering) anyone not retweeted at least 10 times. We further filtered the dataset by zeroing out cells of the matrix where the number of retweets between two users was less than five. This left us with a final matrix for the retweet network that consists of 650,972 rows (users who retweeted others), 125,063 columns (retweeted users), and that makes use of 46,804,138 retweets (18.3% of the original data).

We then applied two methods from prior work to this who-retweeted-whom matrix under a number of different hyperparameter settings. The two existing approaches to identifying clusters are 1) the "UMAP+HDBSCAN" approach from Darwish et al. (2019) introduced above, and 2) an approach that leverages VSP (Vintage Sparse PCA; Rohe and Zeng 2020). VSP has previously been applied to clustering follower relationships among Twitter users with considerable evidence of success, including (in part) to identify partisan leanings (Zhang, Chen, and Rohe 2022). Rohe and Zeng (2020) prove that under certain relatively weak conditions, estimating a VSP matrix decomposition is equivalent to a (fast) identification of the block structures identified by the commonly used stochastic block-model approach. While, to the best of our knowledge, VSP has not been ap-

plied to the retweet network, its prior use on Twitter data and its association with the widely-used stochastic block-model makes it a useful comparison point for the established UMAP-based approach.

The UMAP approach from Darwish et al. (2020) has several hyperparameters. Following prior work, we fix $k = 2$, the number of dimensions in the latent space, and use the default implementation of mean-shift clustering. We further found in initial analyses (as suggested in the documentation of the `umap-learn` python package we use; McInnes et al. 2018) that the only hyperparameter that had significant impacts on outputs was *n_neighbors*. For VSP, there is only one hyperparameter, the number of factors to estimate (i.e. the number of dimensions in the latent space), k.

To find a single clustering of users that is consistent across both methods, we run each method for a variety of hyperparameter settings and find the settings with the highest overlap. We operationalize overlap between the two clustering results using the Normalized Pointwise Mutual Information (NPMI) (Bouma 2009), a commonly-used metric to compare clusterings. More specifically, we run UMAP for values of *n_neighbors* between 5-100 (in increments of 5), and for VSP for k (the number of factor loadings) between 5 and 100 (in increments of 5). We then compare all pairs of clusterings across the two methods (i.e. compute the NPMI for UMAP with *n_neighbors*=5 and VSP with $k = 5$, and then UMAP with *n_neighbors*=5 and VSP with $k = 10$, etc.) and select the hyperparameters from both methods with the highest NPMI. We determined the optimal hyperparameters to be $k = 20$ for the VSP and *n_neighbors* = 15 in UMAP.

After finalizing clusters from both methods, we then manually inspected the ten largest resulting clusters from the UMAP+HDBSCAN approach. Of these, four clusters were significantly larger than the others, and qualitatively could be organized around four main identities: a US-based left- and right-leaning partisan identity, and UK-based left- and right-leaning identities, respectively. Given the US-centric focus of the present work, we restricted ourselves to an analysis of the two US-based groups. Having identified groups using the UMAP+HDBSCAN approach, we then find the clusters from the VSP-based method with the largest overlaps. The NMPI between the left-leaning cluster (which we will call Democrats) and the right-leaning cluster (Republicans) were 0.81 and 0.80, respectively. This indicates a significant overlap in the clusterings. We take as our final labeled users only those users who were identified by both methods, resulting in Democrat and Republican clusters containing 225,439 and 166,726 users, respectively.

Our final step is to conduct a number of validation checks on our clustering and its impact on results. First, we conducted a manual validation of the clusterings to ensure label precision; that is, to ensure that accounts labeled Democrats or Republicans appears to be correct to human coders. To this end, we randomly sampled 300 users identified as either Democratic or Republican by the final clustering. Two authors of the paper that were not involved in the clustering process then separately annotated these users based on 1) their profile description and 2) a tweet they had retweeted. The two annotators agreed on 277 of the 300 users, resulting in a Krippendorf's alpha of 0.85, which signals high agreement. A third author resolved discrepancies for the 23 users where the initial two annotators disagreed. The resulting labels were compared to output from our clustering. In total, 96% of the 300 annotated users were correctly labeled by our automated approach, giving us confidence in the precision of labels assigned using our automated method.

Second, we conducted a validation study to assess label recall; that is, to assess the extend to which our clustering leaves out accounts that should have been labeled Democrats or Republicans. To do so, five authors of the paper labeled 200 randomly sampled accounts that were *not* identified as U.S.-based Democrats or Republicans by our method. Two of the five annotators then labeled each account as being 1) a (U.S.-based) Republican, 2) a Democrat, or 3) neither. A third annotator resolved disagreements. Overall, 87% (174 of 200) accounts were correctly excluded—that is, they were not labeled as Republicans or Democrats by either our algorithm or human annotators. As expected, the majority of these accounts that were correctly excluded were from Great Britain. However, a number of others from India and Australia were also observed. With a Krippendorf's alpha of 0.54, we note that task was more difficult than the precision-centered analysis, but in line with other difficult annotation tasks on social media (e.g. hate speech and image annotation Du, Masood, and Joseph 2020)). Finally, of the remaining accounts, 17 were Republicans and 9 were Democrats, suggesting a slight bias towards excluding Republicans. However, given the high level of recall overall, and the fact that we do not observe any clear qualitative differences between these Republicans and those included in the sample, we do not believe this bias is likely to impact our findings.

Finally, a limitation of our methodology (and those we base it on) is that users who do not actively interact themselves, but are heavily interacted *with*, are not contained in our sample. This means, for example, that FoxNews is not identified as a Republican account in our analysis, because while it is often retweeted, it rarely retweeted other users in our sample. This presents a potential bias in our results; to address this possible bias, we therefore conduct a final robustness check where we develop a heuristic approach to identifying such users and labeling them. Full details of our methodology for this robustness check are provided in the appendix, in the section entitled *Robustness Check Using Additional Labels for Heavily Retweeted Accounts*. Briefly, the intuition behind our approach is to 1) determine the number of times that each account in the dataset was retweeted by accounts we currently had labeled as Republican or Democrat, 2) to compute the weighted log-odds (Monroe, Colaresi, and Quinn 2008) of each account being retweeted by a Democrat versus a Republican, and then 3) label as Democrat any account predominantly retweeted by Democrats, and as Republican any account predominantly retweeted by Republicans. Using this methodology, we replicate the main results of the paper with an additional 90,650 labeled users. Because this methodology incorporates heuristics not based on prior work and not as extensively validated as our original labels, however, we opt to present in the main text results with the smaller set of well-

validated labels from the original set of 392,165 users.

Content Analyses

To extend our understanding of how cross-party interactions are used, we conduct various forms of content analysis. For RQ2 and RQ3, this analysis takes the form of a limited qualitative investigation of heavily retweeted quote tweets (for RQ2) and a curated set of interactions at the core of the network (for RQ3). For RQ1, given the larger set of interactions to analyze, we also leverage quantitative methods for content analysis. Specifically, we leverage both basic statistical methods as well as two more complex content-based methods. We describe the latter two methods here.[2] The first explores the extent to which linguistic differences can be identified across different types of interactions at Level 1, specifically 1) for replies versus quotes, and 2) for in-party versus cross-party interactions. To determine whether these linguistic differences exist, we employ both a term-based approach and a tweet-level approach. For the tweet-level approach, we fine-tuned a RoBERTa (Liu et al. 2019) model to perform classification on 1) whether a Level 1 tweet is a quote or a reply, 2) whether a Level 1 tweet is from a Democrat or Republican, and 3) the intersection (i.e. differentiating between Republican replies and Democrat quotes, and vice versa). In order to avoid issues with imbalanced training and/or evaluation, we implemented an under-sampling technique to balance the number of samples across classes. For the term-based approach, we use the Leave-Out Estimator, proposed by Gentzkow, Shapiro, and Taddy (2016) and used by Demszky et al. (2019), which provides a statistical method to estimate the salience of differences in unigram usage across (e.g.) partisan lines. We apply the Leave-Out Estimator to the same three different groupings that we train the roBERTa models on. For both approaches, we use a five-fold cross-validation strategy for evaluation.

The second content-based analyses we conduct for RQ1 moves beyond whether linguistic differences exist across these dimensions to the topical contents of cross-party interactions specifically. To do so, we employed BERTopic (Grootendorst 2021), which identifies topics using a three-step procedure: 1) a transformer-based language model is used to produce document embeddings, 2) these embeddings are subsequently clustered (e.g. with HDBSCAN), and 3) a procedure is used to generate a topic description from a list of documents clustered into each topic. In our case, we used the default approach in the BERTTopic package, which involves 1) a BERT-based embedding, 2) HDBSCAN for clustering, and 3) a GPT-3.5-turbo-based approach to provide salient terms for each topic. With this approach, we conducted two topic models- one on all cross-party replies, and one on all cross-party quotes. Quotes and replies are analyzed separately because models trained on both produced results that were too heavily dependent on affordance (e.g. the heavy use of usernames in replies) and ultimately proved less informative.

[2] All work in this section was conducted on a single server with 64 CPU cores and two NVIDIA A100 GPUs.

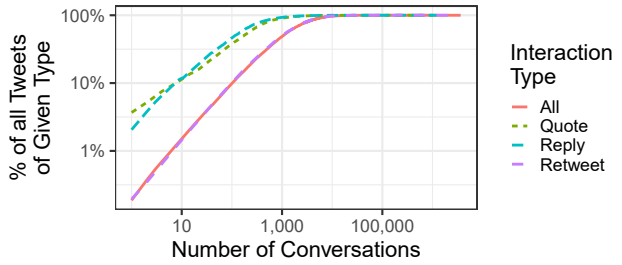

Figure 1: The Empirical Cumulative Distribution function (eCDF) for the proportion of all interactions (y-axis) of a particular type (line type and coloring) contained within a given number of conversations (x-axis).

Electron Decomposition to Find Polarized Groups

Given prior work suggesting that political polarization may be restricted to a small subset of users (Guess et al. 2018), RQ3 aims to understand what the core of polarized users in our dataset might look like. Here the core of polarized users refer to the two groups where the number of positive interactions within each group is high as well as the number of negative edges across two groups is large, which is inspired by the concept of balance (Heider 1946). To find the core of polarized users, we select electron decomposition method (Niu and Sarıyüce 2023), explained above, that is able to identify a subset of users who interact heavily both within and across party lines, across all conversations.

As our dataset is node-labeled, directed, and unsigned but electron decomposition works on unlabeled, undirected, and signed graph, we employ the following transformation. We put a positive edge between two nodes if they ever have an interaction (quote, reply, or retweet) and both nodes are from the same party. Likewise we put a negative edge between two nodes if they ever have an interaction and the nodes are from different parties. We then use electron decomposition to find the most polarized communities (Niu and Sarıyüce 2023). Electron decomposition captures the cohesion along with the polarity to better model the agreement within communities and the conflict across communities. In a given signed network, electron decomposition first removes the nodes that are part of many unbalanced triangles ($++-$, $---$) and then finds cohesive subgraphs in the remaining graph that are abundant with $+--$ signed triangle. Electron decomposition is inspired by the truss decomposition (Cohen 2008) which finds triangle-rich cohesive regions with hierarchical relations in simple unsigned network. The time and space complexities of electron decomposition are the same as truss decomposition: $O(|V| + |E|)$ space and $O(\sum_{v \in V} |N(v)|^2)$ time, where V and E are the number of nodes and edges, and $|N(v)|$ is the degree of node v (further details are available in Niu and Sarıyüce 2023).

Results

Patterns in Level 1 Interactions (RQ1)

Replies and quotes at Level 1 are concentrated on only a few conversations. Figure 1 shows, for example, that over 95%

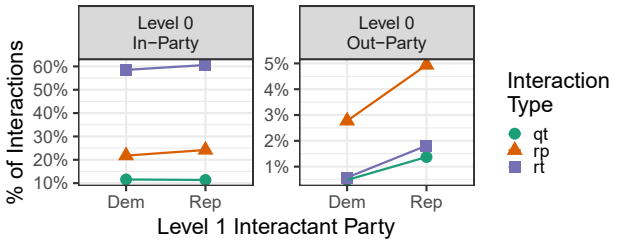

Figure 2: Marginal effects estimated from a binomial regression on the proportion of Level 1 interactions for a given conversation (y-axis) that are of a given type (color) given the partisanship of the user of the Level 1 tweet (x-axis) and whether the Level 0 user was the same party.

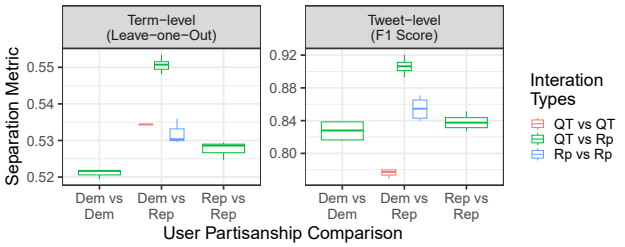

Figure 3: A (left subplot); Results using the leave-one-out metric (y-axis) for content differences across tweets from users with particular partisanships (x-axis) for combinations of interaction types (point color). A higher score represents a greater degree of polarization; results are shown as a boxplot summarizing cross-validation runs.
B (right subplot); The same as A), except results for the tweet-level predictive task results based on RoBERTa. A higher F1 score denotes greater classification accuracy, indicating more polarized language.

of both quotes and replies are contained within the top 1000 conversations in our dataset. These top 1000 conversations account for only 0.01% of all conversations in our dataset. Retweets are less concentrated, with the top 1000 conversations accounting for around only 60% of all retweets.

With respect to the cross versus within-party interactions that occurred within conversations, we find that on average, for a given conversation 1) the majority of all retweets, quotes, and replies are in-party (note the difference in the y-axes between the two subplots in Figure 2), 2) most interactions across party lines are replies (despite replies being limited in our dataset due to API restrictions), and 3) conversations with a Democrat as the OP are more likely to have direct cross-party interactions for retweets and replies. While this is also true of quotes in the regression presented in Figure 2, our validation study suggests that this may be impacted by the biases in our keyword sampling for quotes (only). However, caution should be used in any case for intepreting differences in cross-party interactions for quotes, as the effect size, while statistically significant, was practically small (less than one percent).. These findings are displayed in Figure 2, which shows marginal effect estimates from a binomial regression model estimated on all Level 1 interactions in our dataset. Independent variables in the regression were the Level 0 user's partisanship, the Level 1 user's partisanship, and the type of interaction (reply, quote, or retweet). We estimate a full interaction model from these three variables. The dependent variable was the proportion of interactions that were a given interaction type, for each conversation with at least one Level 1 interaction.

With respect to content differences in Level 1 interactions, we first find that at both the term-level and the tweet-level, there are salient differences 1) across parties and 2) across interaction types. After basic text pre-processing, we computed the leave-out estimate for each user group's text. Figure 3A) shows that the text of replies and quotes, whether from Republicans or Democrats, is highly polarized, with values ranging from .548 to .553. Notably, Democrats tend to use more similar terms across different interaction types. Our roBERTa-based approach (Figure 3B) reveals similar findings when comparing Democrats versus Republicans and quotes versus replies. The highest average classification score ranged from .89 to .92, indicating salient linguistic differences both across parties and across interaction types. Note in both figures that results are not shown for within-party, within-affordance comparisons, as there is no comparison to be made.

Figure 4 presents results from our topic analysis of cross-party Level 1 interactions. The figure shows the top 15 topics that were 1) associated with more than 1,000 cross-party tweets and 2) had the highest absolute weighted log-odds of being shared by a Democrat vs. a Republican, or vice versa. In analyzing these topics, three notable points arise. First, perhaps surprisingly, topics frequently tweeted about in replies by Republicans (but not Democrats) tended to focus on (sometimes misinformed) appeals to science, such as challenges to FDA approval of the vaccine, interpretations of a study of COVID spread in Israel, questions about the transmissibility of the virus even after vaccination, and complications from the vaccine (in particular myocarditis). Relevant to the findings of Zade et al. (2023), cross-party replies in our dataset by Republicans thus often aimed to use *appeals to evidence* to substantiate challenges to government mandates. Second, topics frequently quote tweeted about by Republicans instead tended to challenge not the science, but rather to directly challenges directly the ways in which the U.S. government enacted rules and restrictions on American citizens during COVID, and "Big Pharma's" (sometimes sensationalized) role in this. Finally, consistent with the results in Figure 3, Democrat cross-party interactions did not show as clearly a strategic divide between replies and quotes; instead, both were themed around a variety of personal attacks (e.g. towards Texas Governor Abbott, Florida Governor Ron Desantis, and Donald Trump) and policy issues (e.g. vaccine mandates in the military).

Amplification Beyond Level 1 (RQ2)

Interactions at Level 2 and beyond significantly amplify content from Level 1 interactions, but this amplification process

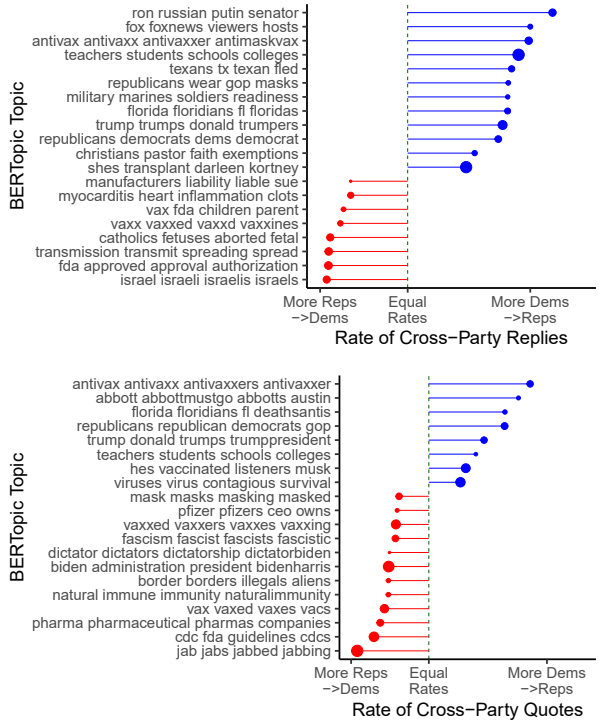

Figure 4: The fifteen topics (x-axis) for cross-party replies (top plot) and quotes (bottom plot) with the highest absolute weighted log-odds (y-axis) of a Democrat interacting with a Republican (or vice versa). Point size indicates the number of tweets assigned to the topic.

varies for quotes and replies. Figure 5 shows that for quotes, the predominant form of amplification is through retweets - there are nearly 2.5 times more retweets of quotes than quotes themselves in our dataset. In contrast, amplification of replies is more muted, in that Level 2+ interactions with Level 1 replies are lower than compared to tweets. Further, Level 2 interactions with replies come more heavily in the form of replies to replies.

These forms of amplification amount to a significant proportion of cross-party interactions via the leap from Level 0 to Levels 2 and beyond, and amongst distinct forms of interaction at Level 1 and beyond. Figure 6 shows, more specifically, that cross-party retweets of quotes make up nearly a third of all cross-party interactions in our data, and cumulatively, indirect cross-party interactions amount to almost half (48.6%) of all cross-party interactions. That is, 30% of all cross-party interactions in our dataset result from situations where User A retweets a quote tweet from User B, who has quoted User C, and User A and User C belong to different partisan groups. Importantly, results here may not reflect the importance of longer reply chains that are not captured in our sample. However, as we discuss further in the appendix, and as noted in prior work (Shugars and Beauchamp 2019), such long chains are relatively rare, and thus we expect that cross-party exposure through a co-partisan frame is still a critical vector of cross-party interaction.

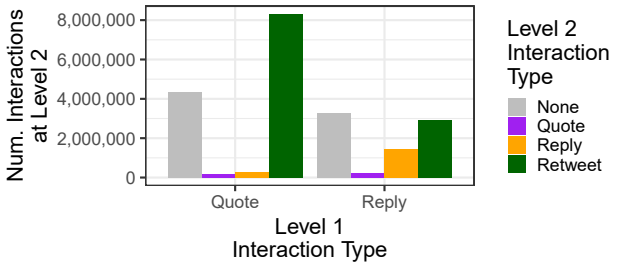

Figure 5: The number of tweets (y-axis) that interact with the two kinds of Level 1 interactions (Quotes and Replies; x-axis) for each kind of Level 2 interaction (different color bars). Provided for comparison is the number of Level 1 interactions of each type (first bar, entitled "None" for Level 2 interactions)

Note that this observation is not necessarily as obvious as it might seem at first glance. It is true that *if* a cross-party quote tweet is retweeted, then there must be at least as many retweets as originating quote tweets. However, the vast majority of quote tweets are not retweeted. Ultimately, then, in our dataset, most observable cross-party interactions are situations where users are retweeting content from the opposing party *that has already been filtered by someone from their own party*. Put another way, observable cross-party interactions in our dataset are primarily driven by users who are seeing things through a co-partisan lens.

To better understand the content of these interactions, we analyzed the 1,471 cross-party quote tweets that were retweeted at least 100 times in our dataset. With respect to content, we find that relative to the topical foci of quote tweets presented in Figure 4, highly retweeted cross-party quote tweets from both parties were almost exclusively centered on vaccines, vaccine mandates, and the Biden presidency. We also see, consistent with findings from Zade et al. (2023), that cross-party quote tweets were predominantly used 1) to speak to one's own audience (relative to the original tweeter), 2) "to relay a sense of antagonism to the politically opposed" (pg. 20). For example, a tweet from a widely followed liberal Twitter account Acyn about a report identifying accounts spreadking vaccine misinformation on Twitter was quote tweeted by Jack Posobiec, a right-leaning media personality, who says "I told you the Biden Admin was working on lists."

What differs between the prior work and our study, however, is with respect to the "who." While Zade et al. (2023) explicitly select for non-elite accounts, we instead aim here to emphasize their importance. Indeed, we observe that highly retweeted cross-party quotes were almost exclusively quotes of elite accounts, by elite accounts. We find, for example, that of the top 50 most retweeted cross-party quote tweets, only one quoted or quoting account had fewer than 10,000 followers. Similarly, half of the 1,471 cross-party quote tweets that were retweeted 100 or more times were quotes of only 32 accounts, which constituted politicians (e.g. Marjorie Taylor Green, Joe Biden, and Rand Paul)and

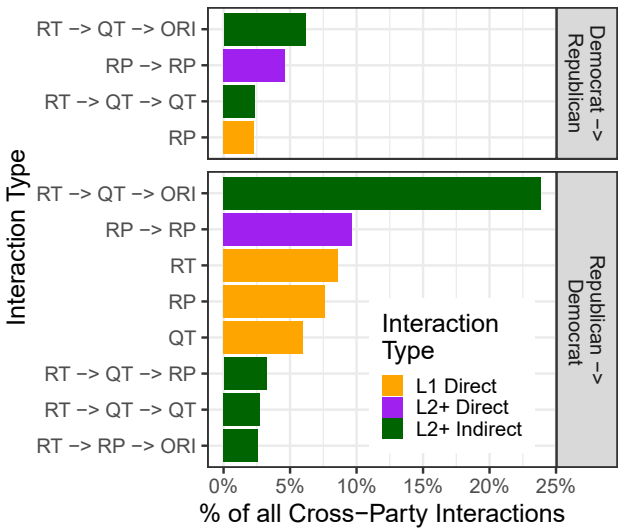

Figure 6: The percentage of all cross-party interactions in our dataset (x-axis) from different types (y-axis) of direct Level 1 interactions (QT, RT, RP; orange bars); direct Level 2 and beyond interactions (e.g. replies to replies, or RP -> RP), and Indirect Level 2 (e.g. RT -> QT -> ORI is a retweet of a quote of an original tweet). The two subplots define the direction of the interaction; we show only interaction types that account for more than 2% of all cross-party interactions.

media accounts (e.g. CNN and Brian Stetler), as well as the account for Pfizer.

The Core of Polarized Users Across All Conversations (RQ3)

Using electron decomposition, we identify a set of 38 Democrats and 57 Republicans that represent a dense core of users that interact heavily within and across parties. The network consisting of these 95 core users has 2706 edges, 1326 of which are cross-party and 1380 of which are within-party. All triangles among these users are balanced, i.e., either +++ or +−−, hence there is a perfect relative balance.

The 57 Republicans in the core represent a diverse array of actors, including authors (e.g., annbauerwriter, Zigmanfreud), political commentators (e.g., YossiGestetner, benshapiro) and entrepreneurs (e.g., aginnt). The Democratic group is largely pro-regulation and consists of news sources (e.g., CNN, washingtonpost, nytimes, AP), journalists (e.g., NateSilver538, apoorva_nyc), scientists (e.g., EricTopol, ashishkjha), and doctors (e.g., walidgellad, _stah, PeterHotez). All users in the two groups have more than 10,000 followers, and thus we provide their de-anonymized user names in Table 3 in the Appendix. With respect to the core set of users, the most notable difference between Democrats and Republicans is that major right-leaning news sources (e.g. Fox News and Newsmax) are not included. This, we find, occurs even when we extend our labeling to include users that do not often interact across party, see details in the robustness checks in the appendix.

L1→L0	QTs	RPs	RTs	Total
D→D	962	306	1,729	2,997
R→R	892	1,354	6,357	8,603
D→R	25	597	7	629
R→D	1,967	2,449	1,793	6,209

Table 2: Statistics of the most cohesively polarized pair of groups, 38 Democrats and 57 Republicans, on the number of Level 1 tweets for the different interaction types (QT = quotes, RP = replies, RT= retweets) for each combination of partisanship of Level 1 and Level 0 users (e.g. D→D is for a Democrat replying to, retweeting, or quoting a Level 0 post from a Democrat)

With respect to interactions amongst these users, Table 2 shows that the vast majority of interactions are within or cross-party interactions originating with Republicans. When compared to the entire set of Level 1 interactions given in Table 1, we observe that the fraction of retweets is far fewer in the polarized pair of communities (53.6% vs. 82.5%) whereas the fractions of quotes and replies are larger (20.9% vs. 8.5% and 25.5% vs 9.0%). This is consistent with our focus on users that interact heavily both within and across party. Regarding cross-party interactions, Republican users in the core leverage quotes, replies, and retweets at approximately the same rate as in the overall data. However, regarding the interactions from Democrats to Republicans, however, there is a striking difference. The fraction of quotes, replies, and retweets from Democrats to Republicans are 29.1%, 42.9%, 28.0% in Table 1, but those numbers change to 4.0%, 94.9%, 1.1% when analyzing the core. At the core of the network, Democrats thus make almost no quotes nor retweets of Republicans.

Indeed, core Democrats retweet core Republicans only seven times. These do, however, appear to be legitimate cases in which left-leaning users could have reasonably agreed with right-leaning users, e.g. on criticisms of mask mandates (*"So, to be clear, the CDC is now pushing masking for the vaccinated nationally because there were some 882 cases of covid... a grand total of seven were hospitalized and 0 died"*), and the ineffectiveness of government institutions (*"It was private industry and the states that basically ended the pandemic. The CDC and FDA completely blew it..."*). Finally, we see that interactions are in general asymmetric, in that 33.6% of all interactions are Republicans interacting with Democrats, compared to only 3.2% for the whole data.

In summary, our analysis of the network core is consistent with our findings for RQ1 and RQ2, in that most cross-party interactions originate with Republicans. However, we find that at the core of the interaction network, this distinction is much more prominent: the core left-leaning users are largely either established media accounts or scientists that appear uninterested in cross-party dialogue, whereas the core of the right-leaning network relies heavily on both bringing content from the left into their networks with comments (i.e. quoting) and with direct replies. We do see, however, some evidence of legitimate cross-party support.

Conclusion

The present work provides insight into patterns in cross-party interactions in a large dataset tweets about COVID. At the highest level, we present a number of findings showing that 1) cross-party interactions were assymetric, in that Republicans more heavily interacted across party lines and in more distinct ways across replies versus quotes, 2) that these cross-party interactions were centered on a very small number of conversations, and 3) that *indirect* cross-party interactions—especially in the form of retweets of cross-party quotes of and by political elites—account for nearly half of all forms of cross-party interaction.

Taken together, these findings have both theoretical and practical implications. Theoretically, our work provides additional support for the growing body of work that aims to center elite discourse (Green 2021) and its amplification (Gallagher 2022) in the study of online political polarization, relative to work that aims to measure polarization of ordinary users. More concretely, we suggest that the combination of quote tweets, elite discourse, and its amplification produces an information environment where non-traditional right-wing political elites can enter into a one-way, faux "discussion" with left-wing traditional elites via the quote feature. This is a faux discussion, in that while the quote represents a cross-party interaction, it is instead used as a means to frame the out-party for one's own (politically congruent) followers. Moreover, we find that cross-party interaction was centered within a limited number of conversations, perhaps further centralizing the importance of elite discussion. Future theorizing is needed to help generalize this complex interplay between asymmetric political behavior of elites, platform affordances, and the long-tail of attention on social media.

Practically, and related, our work suggests that interventions to address political polarization online must account for the importance of elites and potential assymetries in behavior across the partisan spectrum and platform affordance. More specifically, our work points to the need for interventions that emphasize the ways in which co-partisan elites selectively sample examples from the out-party to make a point, and aim to encourage users to see (potentially explicitly) examples of out-party users who may not have, e.g., as extreme political views as those selected to be quoted.

There are, however, several limitations to our work that others should take care to note. First, we focus on a specific case study, and findings may not generalize beyond it. However, we note here that 1) many of our findings are consistent with prior work on other data (Zade et al. 2023; Garimella, Weber, and De Choudhury 2016), 2) that COVID represents a particularly important case study for modern American politics, and 3) that even without generalizing beyond COVID, our findings present an important caveat to traditional considerations of how partisan discussions play out online. Second, our work is based on a classification scheme that labels nearly 400,000 users. While we found our method to have high precision and recall, and that our results hold across a number of robustness checks, it is therefore possible that our results are driven by misclassifications. Third and similarly, while we do our best to address concerns about API biases in the appendix, our results could be driven by biases in the API that we do not consider here.

Finally, our empirical findings are specific to Twitter and more particularly to the use of Twitter surrounding one particularly polarizing topic. However, as Gallagher (2022) notes, we can with care generalize the core technological affordances of retweeting, quoting, and replying to other social media platforms. In particular, the ways in which we can amplify content from the opposing party with our own comments is available on a number of platforms, e.g. on TikTok as a form of video remixing. Zade et al. (2023) make a similar point, tying their findings on the use of reply versus quote usage to decisions that other platforms (e.g. Mastadon in their case) are facing, and how findings on one platform can, with care, inform decisions on other platforms. We therefore believe that our findings may provide new insights beyond Twitter into how the different ways in which users can share content can, when combined, lead to new and pernicious forms of mediated exposure across parties.

Acknowledgements

Z. An and K. Joseph were supported by an ONR MURI N00014-20-S-F003. J. Niu, J. Breuhaus, and A. E. Sariyuce were supported by NSF awards OAC-2107089 and IIS-2236789.

References

Aref, S.; and Wilson, M. C. 2018. Measuring partial balance in signed networks. *Journal of Complex Networks*, 6(4): 566–595.

Bail, C. A.; Argyle, L. P.; Brown, T. W.; Bumpus, J. P.; Chen, H.; Hunzaker, M. B. F.; Lee, J.; Mann, M.; Merhout, F.; and Volfovsky, A. 2018. Exposure to Opposing Views on Social Media Can Increase Political Polarization. *Proceedings of the National Academy of Sciences*, 115(37): 9216–9221.

Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239): 1130–1132.

Balietti, S.; Getoor, L.; Goldstein, D. G.; and Watts, D. J. 2021. Reducing Opinion Polarization: Effects of Exposure to Similar People with Differing Political Views. *Proceedings of the National Academy of Sciences*, 118(52): e2112552118.

Barberá, P.; Jost, J. T.; Nagler, J.; Tucker, J. A.; and Bonneau, R. 2015. Tweeting from Left to Right: Is Online Political Communication More than an Echo Chamber? *Psychological science*, 26(10): 1531–1542.

Bonchi, F.; Galimberti, E.; Gionis, A.; Ordozgoiti, B.; and Ruffo, G. 2019. Discovering polarized communities in signed networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 961–970.

Bouma, G. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30: 31–40.

Cartwright, D.; and Harary, F. 1956. Structural balance: a generalization of Heider's theory. *Psychological review*, 63(5): 277.

Cogan, P.; Andrews, M.; Bradonjic, M.; Kennedy, W. S.; Sala, A.; and Tucci, G. 2012. Reconstruction and Analysis of Twitter Conversation Graphs. In *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, 25–31.

Cohen, J. 2008. Trusses: Cohesive subgraphs for social network analysis. *National Security Agency Technical Report*, 16(3.1).

Cohen, R.; and Ruths, D. 2013. Classifying political orientation on Twitter: It's not easy! In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, 91–99.

Conover, M.; Ratkiewicz, J.; Francisco, M.; Gonçalves, B.; Menczer, F.; and Flammini, A. 2011. Political polarization on twitter. In *Proceedings of the international aaai conference on web and social media*, volume 5, 89–96.

Crupi, G.; Mejova, Y.; Tizzani, M.; Paolotti, D.; and Panisson, A. 2022. Echoes through Time: Evolution of the Italian COVID-19 Vaccination Debate. Technical Report arXiv:2204.12943, arXiv. ArXiv:2204.12943 [physics] type: article.

Darwish, K.; Stefanov, P.; Aupetit, M.; and Nakov, P. 2020. Unsupervised User Stance Detection on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 14: 141–152.

Darwish, K. M.; Weber, I.; Wagner, C.; Zagheni, E.; Nelson, L.; Aref, S.; and Flöck, F., eds. 2019. *Social Informatics: 11th International Conference, SocInfo 2019, Doha, Qatar, November 18–21, 2019, Proceedings*, volume 11864 of *Lecture Notes in Computer Science*. Cham: Springer International Publishing. ISBN 978-3-030-34970-7 978-3-030-34971-4.

Demszky, D.; Garg, N.; Voigt, R.; Zou, J.; Gentzkow, M.; Shapiro, J.; and Jurafsky, D. 2019. Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings. ArXiv:1904.01596 [cs].

Du, Y.; Masood, M. A.; and Joseph, K. 2020. Understanding Visual Memes: An Empirical Analysis of Text Superimposed on Memes Shared on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 153–164.

Fraxanet, E.; Pellert, M.; Schweighofer, S.; Gómez, V.; and Garcia, D. 2023. Unpacking Polarization: Antagonism and Alignment in Signed Networks of Online Interaction. arxiv:2307.06571.

Gallagher, R. J. 2022. *The Network Structure of Online Amplification*. Ph.D. thesis, Northeastern University.

Gallagher, R. J.; Reagan, A. J.; Danforth, C. M.; and Dodds, P. S. 2018. Divergent Discourse between Protests and Counter-Protests: #BlackLivesMatter and #AllLivesMatter. *PLOS ONE*, 13(4): e0195644.

Garimella, K.; Morales, G. D. F.; Gionis, A.; and Mathioudakis, M. 2018. Quantifying Controversy on Social Media. *ACM Transactions on Social Computing*, 1(1): 1–27.

Garimella, K.; Weber, I.; and De Choudhury, M. 2016. Quote RTs on Twitter: Usage of the New Feature for Political Discourse. In *Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, 200–204. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-4208-7.

Garimella, V. R. K.; and Weber, I. 2017. A long-term analysis of polarization on Twitter. In *Proceedings of the International AAAI Conference on Web and social media*, volume 11, 528–531.

Gentzkow, M.; Shapiro, J. M.; and Taddy, M. 2016. Measuring Polarization in High-Dimensional Data: Method and Application to Congressional Speech.

González-Bailón, S.; and Lelkes, Y. 2023. Do Social Media Undermine Social Cohesion? A Critical Review. *Social Issues and Policy Review*, 17(1): 155–180.

Green, J. 2021. Belief Systems in Theory and Practice: Evidence from Political Pundits.

Grootendorst, M. 2021. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Guerra, P.; Nalon, R.; Assunção, R.; and Meira Jr., W. 2017. Antagonism Also Flows Through Retweets: The Impact of Out-of-Context Quotes in Opinion Polarization Analysis. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1): 536–539.

Guess, A.; Lyons, B.; Nyhan, B.; and Reifler, J. 2018. *Avoiding the echo chamber about echo chambers: Why selective exposure to like-minded political news is less prevalent than you think*.

Guimaraes, A.; and Weikum, G. 2021. X-posts explained: Analyzing and predicting controversial contributions in thematically diverse reddit forums. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 163–172.

Hada, R.; Ebrahimi Fard, A.; Shugars, S.; Bianchi, F.; Rossini, P.; Hovy, D.; Tromble, R.; and Tintarev, N. 2023. Beyond Digital "Echo Chambers": The Role of Viewpoint Diversity in Political Discussion. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, 33–41. ACM.

Heider, F. 1946. Attitudes and cognitive organization. *The Journal of Psychology*, 21(1): 107–112.

Jackson, S. J.; and Foucault Welles, B. 2015a. Hijacking# myNYPD: Social Media Dissent and Networked Counterpublics. *Journal of Communication*, 65(6): 932–952.

Jackson, S. J.; and Foucault Welles, B. 2015b. Hijacking #myNYPD: Social Media Dissent and Networked Counterpublics: Hijacking #myNYPD. *Journal of Communication*, 65(6): 932–952.

Jiang, X.; Su, M.-H.; Hwang, J.; Lian, R.; Brauer, M.; Kim, S.; and Shah, D. 2021. Polarization Over Vaccination: Ideological Differences in Twitter Expression About COVID-19 Vaccine Favorability and Specific Hesitancy Concerns. *Social Media + Society*, 7(3): 205630512110484.

Joseph, K.; Swire-Thompson, B.; Masuga, H.; Baum, M. A.; and Lazer, D. 2019. Polarized, together: Comparing partisan support for Trump's tweets using survey and platform-based measures. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 290–301.

Kreiss, D.; and McGregor, S. C. 2023. A Review and Provocation: On Polarization and Platforms. *New Media & Society*, 14614448231161880.

Lai, M.; Tambuscio, M.; Patti, V.; Ruffo, G.; and Rosso, P. 2019. Stance polarity in political debates: A diachronic perspective of network homophily and conversations on Twitter. *Data & Knowledge Engineering*, 124: 101738.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv:1907.11692 [cs] version: 1.

Lorentzen, D. G. 2020. Bridging Polarised Twitter Discussions: The Interactions of the Users in the Middle. *Aslib Journal of Information Management*, 73(2): 129–143.

Lorentzen, D. G.; and Nolin, J. 2017. Approaching Completeness: Capturing a Hashtagged Twitter Conversation and Its Follow-On Conversation. *Social Science Computer Review*, 35(2): 277–286.

Magdy, W.; Darwish, K.; Abokhodair, N.; Rahimi, A.; and Baldwin, T. 2016. #ISISisNotIslam or #DeportAllMuslims?: predicting unspoken views. In *Proceedings of the 8th ACM Conference on Web Science*, 95–106. Hannover Germany: ACM. ISBN 978-1-4503-4208-7.

Marchal, N. 2022. "Be Nice or Leave Me Alone": An Intergroup Perspective on Affective Polarization in Online Political Discussions. *Communication Research*, 49(3): 376–398.

McInnes, L.; Healy, J.; Saul, N.; and Grossberger, L. 2018. UMAP: Uniform Manifold Approximation and Projection. *The Journal of Open Source Software*, 3(29): 861.

Metaxas, P.; Mustafaraj, E.; Wong, K.; Zeng, L.; O'Keefe, M.; and Finn, S. 2015. What do retweets indicate? Results from user survey and meta-review of research. In *Proceedings of the international AAAI conference on web and social media*, volume 9, 658–661.

Monroe, B. L.; Colaresi, M. P.; and Quinn, K. M. 2008. Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis*, 16(4): 372–403.

Mønsted, B.; and Lehmann, S. 2022. Characterizing Polarization in Online Vaccine Discourse—A Large-Scale Study. *PLOS ONE*, 17(2): e0263746.

Muric, G.; Wu, Y.; and Ferrara, E. 2021. COVID-19 vaccine hesitancy on social media: building a public Twitter data set of antivaccine content, vaccine misinformation, and conspiracies. *JMIR public health and surveillance*, 7(11): e30642.

Nishi, R.; Takaguchi, T.; Oka, K.; Maehara, T.; Toyoda, M.; Kawarabayashi, K.-i.; and Masuda, N. 2016. Reply Trees in Twitter: Data Analysis and Branching Process Models. *Social Network Analysis and Mining*, 6(1): 26.

Niu, J.; and Sarıyüce, A. E. 2023. On Cohesively Polarized Communities in Signed Networks. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, 1339–1347. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394192.

Ordozgoiti, B.; Matakos, A.; and Gionis, A. 2020. Finding large balanced subgraphs in signed networks. In *Proceedings of The Web Conference 2020*, 1378–1388.

Robertson, R. E.; Green, J.; Ruck, D. J.; Ognyanova, K.; Wilson, C.; and Lazer, D. 2023. Users choose to engage with more partisan news than they are exposed to on Google Search. *Nature*, 1–7.

Rohe, K.; and Zeng, M. 2020. Vintage Factor Analysis with Varimax Performs Statistical Inference. ArXiv:2004.05387 [math, stat].

Schnoebelen, T.; Silge, J.; and Hayes, A. 2022. *tidylo: Weighted Tidy Log Odds Ratio*. R package version 0.2.0.

Shugars, S.; and Beauchamp, N. 2019. Why Keep Arguing? Predicting Engagement in Political Conversations Online. *SAGE Open*, 9(1): 2158244019828850.

Thorson, K.; and Wells, C. 2016. Curated Flows: A Framework for Mapping Media Exposure in the Digital Age. *Communication Theory*, 26(3): 309–328.

Tzeng, R.-C.; Ordozgoiti, B.; and Gionis, A. 2020. Discovering conflicting groups in signed networks. *Advances in Neural Information Processing Systems*, 33.

Törnberg, P. 2022. How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences*, 119(42): e2207159119.

Volkova, S.; Coppersmith, G.; and Van Durme, B. 2014. Inferring User Political Preferences from Streaming Communications. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 186–196. Baltimore, Maryland: Association for Computational Linguistics.

Weeks, B. E.; Lane, D. S.; Kim, D. H.; Lee, S. S.; and Kwak, N. 2017. Incidental exposure, selective exposure, and political information sharing: Integrating online exposure patterns and expression on social media. *Journal of computer-mediated communication*, 22(6): 363–379.

Xiao, H.; Ordozgoiti, B.; and Gionis, A. 2020. Searching for polarization in signed graphs: a local spectral approach. In *Proceedings of The Web Conference 2020*, 362–372.

Xue, J.; Chen, J.; Hu, R.; Chen, C.; Zheng, C.; Su, Y.; and Zhu, T. 2020. Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach. *Journal of Medical Internet Research*, 22(11): e20550.

Yardi, S.; and Boyd, D. 2010. Dynamic Debates: An Analysis of Group Polarization Over Time on Twitter. *Bulletin of Science, Technology & Society*, 30(5): 316–327. Publisher: SAGE Publications Inc.

Zade, H.; Williams, S.; Tran, T. T.; Smith, C.; Venkatagiri, S.; Hsieh, G.; and Starbird, K. 2023. To Reply or to Quote: Comparing Conversational Framing Strategies on Twitter. *ACM Journal on Computing and Sustainable Societies*.

Zhang, J.; Danescu-Niculescu-Mizil, C.; Sauper, C.; and Taylor, S. J. 2018. Characterizing Online Public Discussions through Patterns of Participant Interactions. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW): 1–27.

Zhang, Y.; Chen, F.; and Rohe, K. 2022. Social Media Public Opinion as Flocks in a Murmuration: Conceptualizing and Measuring Opinion Expression on Social Media. *Journal of Computer-Mediated Communication*, 27(1): zmab021.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, we believe our work contributes to our understanding of how existing social contracts are publicly debated, and that we do so in a respectful way.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? We have aimed not to over-state our findings in these sections
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? See the Data & Methods section, where we try to justify our approach especially via the fact that it draws heavily from prior well-validated approaches
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? See the Data & Methods section, in particular, where we describe the limitations of our datasets, and in the Conclusion, where we address methodological limitations
   (e) Did you describe the limitations of your work? See the Conclusion, where we note several limitations
   (f) Did you discuss any potential negative societal impacts of your work? See the Conclusion
   (g) Did you discuss any potential misuse of your work? See the Conclusion
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? See the Conclusion, where we discuss replication materials in particular.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? We have!
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Our work is exploratory in nature, we correspondingly state the limitations of such in the Conclusion (relative to hypothesis-driven work).
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? We have not. We plan to provide this in a de-anonymized version of the final publication, if accepted
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? We believe we have provided the necessary information to replicate this work
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, where necessary, or used other methods of variation.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, We discuss our various methods for validating our approach to identifying partisanship of users.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, in the conclusion we discuss limitations associated with misclassifications for our method.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? No, as citing our original work with this dataset would, we believe, violate anomynity. We will provide such citations in a de-anonymized version of the paper
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? NA
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No. We believe our approach to data collection and use would be considered standard in the ICWSM community and thus have chosen not to address the well-known concerns that arise in the use of Twitter data
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? No, for similar reasons as above
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? NA

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see **?**)? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA, our university does not consider work on social media to be human subjects work and thus will not entertain any IRB in this vein.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

Appendix

38 Democrats		57 Republicans	
@DrEricDing	@DrLeanaWen	@benshapiro	@stopthistrain95
@_stah	@sailorrooscout	@aginnt	@MissingLaptop
@ASlavitt	@MarkLevineNYC	@ianmSC	@rTIKId
@nytimes	@MonicaGandhi9	@ConceptualJames	@AveMaria2018
@EricTopol	@NateSilver538	@justin_hart	@JVolushin
@ashishkjha	@CDCDirector	@kerpen	@neo_mexicanus
@zeynep	@VincentRK	@DanielKotzin	@CML915
@CNN	@celinegounder	@EWoodhouse7	@jonathanwsabin
@PeterHotez	@JamesSurowiecki	@YossiGestetner	@HockeyNow65
@CDCgov	@washingtonpost	@ITGuy1959	@PapaBeaver2023
@AP	@Bob_Wachter	@Zigmanfreud	@ceb217
@therecount	@JeromeAdamsMD	@tlowdon	@AWokeZombie
@RidleyDM	@drlucymcbride	@ontheasternsea	@NormalcyNow
@apoorva_nyc		@AJKayWriter	@HumphreyPT
@AGHamilton29		@txsalth2o	@peterphan
@ScottGottliebMD		@annbauerwriter	@HeckofaLiberal
@Jusrangers		@PhilHollowayEsq	@reopenpa
@ProfEmilyOster		@districtai	@AConcernedPare2
@DrTomFrieden		@rfsquared	@AmyA1A
@JenniferNuzzo		@E_got_tweets	@TheRationalMD
@rweingarten		@LaffersNapkin	@HjgTweet
@DLeonhardt		@ifihadastick	@MomOnAMission30
@j_g_allen		@erichhartmann	@boutros555
@walidgellad		@el_s00nd	

Table 3: The most cohesive pair of communities found by electron decomposition. There are 38 Democrats, 57 Republicans, with 2706 total edges among them, 1326 of which are cross-party and 1380 are intra-party.

Robustness Check Using Additional Labels for Heavily Retweeted Accounts

In this section, we provide a robustness check that shows that our main results hold when we include an additional 91,322 accounts that were retweeted more than ten times by labeled users but that themselves did not interact heavily enough to be assigned a label in our original methodology.

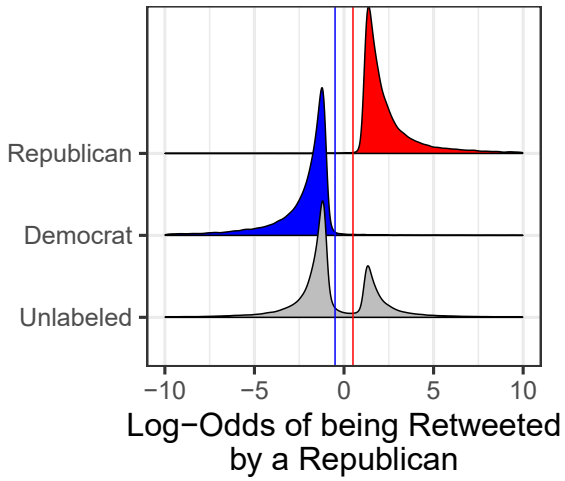

Figure 7: The weighted log-odds of an account being retweeted by a labeled Republican (vs. a Democrat) for accounts labeled in a particular way (y-axis). 5,228 accounts with extreme values are not included for visual clarity.

To identify these users, we first subset the 759,654 unlabeled users who were retweeted at least once by labeled users in our dataset to the 12.3% of them (93,650) were retweeted more than 10 times. We do so, again, to ensure we have enough data for these accounts to provide a valid estimate of how often they were retweeted by Republicans vs. Democrats. As a point of comparison, there were 35,892 accounts we labeled as Democrats that were retweeted at least 10 times, and 24,326 accounts we labeled as Republican.

With these users, we then compute the log-odds log-odds of each account being retweeted by a Republican, relative to a Democrat. To do so, we use the empirical Bayesian approach for log-odds computation from Monroe, Colaresi, and Quinn (2008), as implemented in the R package `tidylo` (Schnoebelen, Silge, and Hayes 2022).

Using this method, we find that on average, an account labeled as a Democrat was around 23 times more likely to be retweeted by a Democrat than a Republican, and a labeled Republican was 32 times more likely to be retweeted by a Republican. Figure 7 shows density functions which make this point visually, and also show that unlabeled accounts have a strongly bimodal distribution of this weighted log-odds measure that clearly separates these accounts into those heavily retweeted by Democrats, and those heavily retweeted by Republicans.

Using this figure as a guide, our robustness check therefore heuristically splits unlabeled accounts retweeted more than ten times into three bins: those with a weighted log-odds of 1) less than -0.5, which we label as Democrats for our robustness check 2) greater than 0.5, which we label as Republicans, and 3) on the interval [-0.5,0.5], which we keep as unlabeled. These three bins represent 65% (60,617 accounts), 33% (30,705 accounts), and 2% (2,328 accounts) of the 93,650 unlabeled accounts in our dataset that were retweeted more than 10 times. As a comparison, 98.6%

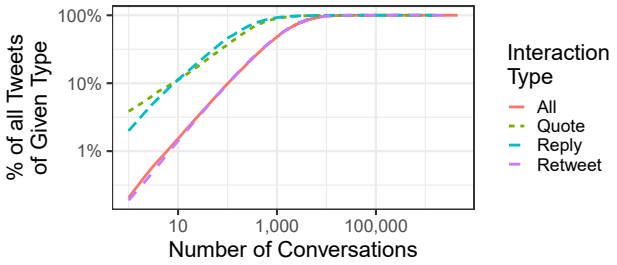

Figure 8: Replication of Figure 1 with additional user party labels.

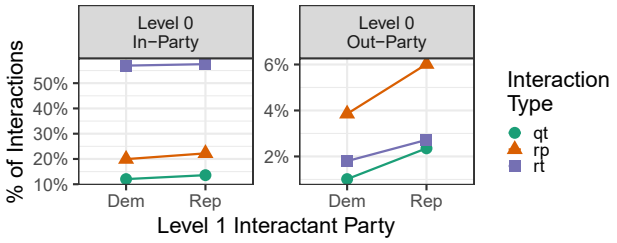

Figure 9: Replication of Figure 2 with additional user party labels.

(99.7%) of accounts we labeled Democrat (Republican) that were retweeted more than 10 times are labeled as Democrat (Republican) using this same heuristic. As a result, we use for our robustness check a set of 60,617 additional labeled Democrats and 30,705 labeled Republicans.

Figures 8-10 replicate the core findings of the paper with this data, showing results that are consistent, qualitatively, with our claims in the main text. Figure 11 differs from Figure 6 in that the percentage of cross-party interactions that are replies-to-replies drops below direct L1 interactions with the expanded data. However, our core argument stemming from Figure 6 is that the majority of cross-party interactions take the form of retweets of quotes of original posts, and thus do not believe this particular difference is qualitatively salient for the paper's main conclusions.

We also repeat the experiments for RQ3 with thousands of additional user party labels. Compared to the

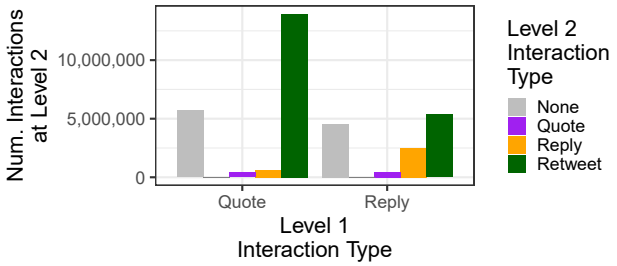

Figure 10: Replication of Figure 5 with additional user party labels.

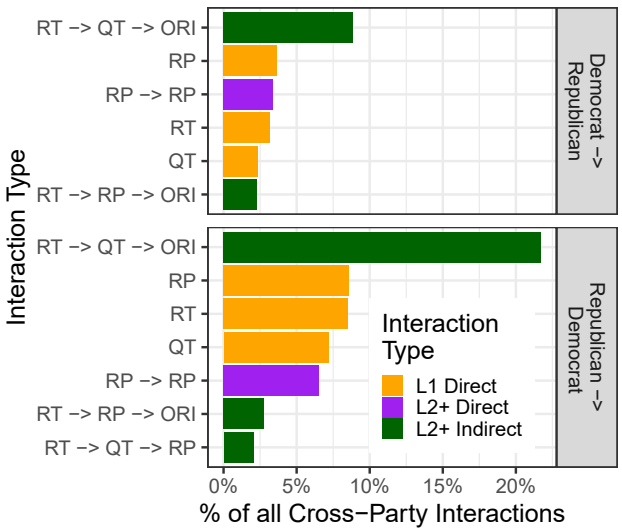

Figure 11: Replication of Figure 6 with additional user party labels.

L1→L0	QTs	RPs	RTs	Total
D→D	917	218	1,548	2,683
R→R	1,118	1,436	6,780	9,334
D→R	20	482	6	508
R→D	2,206	2,533	1,921	6,660

Table 4: Updated Table 2 statistics on expanded party labels. 40 Democrats and 57 Republicans participate in this dense core of users with 2834 total edges among them, 1430 of which are cross-party and 1404 are intra-party.

original subgraph listed in Table 3, this updated subgraph contains nine additional users (@ChristinaPushaw [R], @BenMarten [R], @schraderism [R], @LilithAssyria [R], @lcarb_4u [R], @thehill [D], @AstorAaron [D], @apsmunro [D], @POTUS [D]) while removing seven original members (@RidleyDM, @VincentRK, @GreatNickDix, @Melinda01212917, @OHcs2021, @rfsquared, @ontheasternsea). There are a total of 40 Democrats and 57 Republicans connected by 2834 edges with 1430 edges across party lines and 1404 edges within each party. We give the updated statistics in Table 4.

It is evident that there is not a significant difference between the original statistics in Table 2 and the new statistics in Table 4. These results confirm our findings that most cross-party interactions originate with Republicans and is much more prominent at the core. As such, core left-leaning users, composed of mainly established media accounts or scientists, have a tendency to avoid cross-party dialogue. However, core right-leaning users often participate in cross-party interactions through quotes, replies, and retweets.

Analysis of biases due to missing replies

In our sample of 1,860 conversations in which there was at least one reply and one quote, we find that on average per

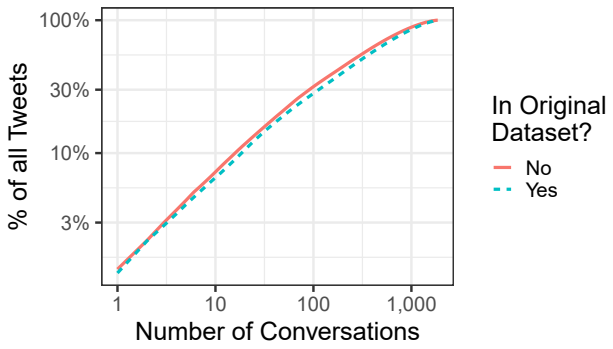

Figure 12: Replication of Figure 1 comparing concentration of replies in the validation data that were or were not in the original sample.

conversation, our original 255M tweet keyword-based sample captured 76% of all quote tweets, averaging roughly 562 missed quotes per conversation. For replies, conversations in our dataset were on average missing 901 replies, meaning that our original dataset on average captured 22% of all replies to a conversation. However, the distribution of missing quotes was bimodal, and was based on whether or not the OP contained a keyword we used for data collection: in cases where it contained the keyword, we obtain almost all (97%) of quotes; in cases where it did not, we capture only around 13.5% of all quotes.

Using these data, we consider evidence we have for the potential impact of these biases in collections of quotes and replies on our main results:

- We find that *at L1, retweets, replies, and quotes are centered on a small proportion of conversations.* Data from our validity check suggests this result is not likely to be biased- the concentration of replies and quotes in conversations in the validity check is almost identical to the concentration for replies/quotes that are vs. those that are not contained in our original sample; see Figure 12.
- We find that *at L1, retweets, replies, and quotes are predominantly in-party, on average for a given tweet.* Analysis of data from our validity check suggests this result is not likely to be biased. Specifically, we conduct a logistic regression where the dependent variable is the probability of an interaction being a within-party interaction, and the dependent variables are 1) the interaction type (reply or quote), 2) the party of the user at Level 0, and 3) whether or not the interaction is contained in our original sample. We find no evidence (p =.23) that a model which has interaction terms between these three dependent variables fits better than a model with only main effects. In the main effects only model, we moreover find that there is no significant difference between the probability of an in-party interaction conditioned on whether or not the interaction is in our original sample (p=.089).
- We find that *Republicans are more likely to use all three forms of interaction across party lines.* For this result to be invalidated by a samplign bias, it would have to be the case that Democrats use replies and/or quotes without our keywords more frequently than Republicans in cross-party interactions. As noted in the main text, we find that there *does* exist a difference here between the replies/quotes in our original sample versus those in the validation set. Specifically, we find that Republicans are significantly ($p < .001$) less likely to use quotes for cross-party interactions in the sample missed by our collection than they are in the sample in our original dataset. We temper conclusions relevant to this claim accordingly in the main text.
- We find that *cross-party interactions are largely consumed through a filtered lens, especially retweets of quotes.* This finding could be invalidated if cross-party replies not in our dataset were significantly more prevalent than retweets of cross-party quotes. We can use metadata from the Twitter v2 API to assess this possibility, as we can obtain the number of times each interaction that was not in our original sample was itself retweeted, quoted, or replied to. To this end, we find that on average per conversation, our sample misses 928 (95% bootstrapped CI [736,1144]) retweets of cross-party quotes, compared to 246 [229,263] direct L1 replies, 58 [41,78] replies to replies, and 110 [97,123] cross-party quotes. As we note in the main text, we cannot directly use our validation data to assess the possibility of lengthy reply chains that involve significantly many more cross-party replies, but prior work suggests this is unlikely (Shugars and Beauchamp 2019), as does the fact that on average, less than a quarter (23%, or 57/246) of top-level replies in our validation data get any replies at all. Overall, then, metadata from our validation sample suggests that, if anything, cross-party interactions are even more heavily concentrated within retweets of cross-party quotes.
- Finally, with respect to the network core analysis, we find that *core left-leaning users are largely either established media accounts or scientists that appear uninterested in cross-party dialogue, whereas the core of the right-leaning network relies heavily on bringing content from the left into their networks.* For this result to be invalidated by biases in data collection, it would have to be the case that there existed core network members who interacted almost entirely without use of our keywords. If this were to be the case, though, then we would likely be uninterested in these interactions, given our interest in COVID (and the relevant keywords). We therefore believe that results for RQ3 are not directly impacted by this sampling bias.

In sum, we find limited evidence in our validation sample that would lead us to believe that our main claims are invalid. Of course, content analyses presented in the paper will be impacted by these sampling biases; we take care in the main text of the paper to note this and so do not attempt to address this point here.