nature communications

Article

https://doi.org/10.1038/s41467-024-53868-0

Patterns of partisan toxicity and engagement reveal the common structure of online political communication across countries

Received: 16 December 2023

Accepted: 23 October 2024

Published online: 14 November 2024

Check for updates

Max Falkenberg[1,2] ✉, Fabiana Zollo[3,4], Walter Quattrociocchi[5], Jürgen Pfeffer[6] & Andrea Baronchelli[2,7] ✉

Existing studies of political polarization are often limited to a single country and one form of polarization, hindering a comprehensive understanding of the phenomenon. Here we investigate patterns of polarization online across nine countries (Canada, France, Germany, Italy, Poland, Spain, Turkey, UK, USA), focusing on the structure of political interaction networks, the use of toxic language targeting out-groups, and how these factors relate to user engagement. First, we show that political interaction networks are structurally polarized on Twitter (currently X). Second, we reveal that out-group interactions, defined by the network, are more toxic than in-group interactions, indicative of affective polarization. Third, we show that out-group interactions receive lower engagement than in-group interactions. Finally, we identify a common ally-enemy structure in political interactions, show that political mentions are more toxic than apolitical mentions, and highlight that interactions between politically engaged accounts are limited and rarely reciprocated. These results hold across countries and represent a step towards a stronger cross-country understanding of polarization.

Political polarization has important democratic consequences. Some ideological polarization is critical for driving debate in public policy and improving the deliberation of ideas[1]. However, severe polarization can stifle debate, drive animosity between groups, and may result in democratic backsliding[2] or violence[3]. As a result, a vast literature has emerged aiming to better understand political polarization and its diverse sub-types.

Scholars place a particular emphasis on affective polarization, defined as the tendency to dislike ones partisan opponents, given that it may undermine the mechanisms which allow a democracy to function[4]. Researchers have shown that affective polarization has grown steadily in the USA[5], linking this growth to social identity theory[4] and partisan sorting[6,7]: Political parties are increasingly associated with specific social identities and demographics[8,9], increasing the perceived distance between partisans, which in turn drives animosity between political opponents[10].

However, recent work highlights why we should be careful not to assume that findings on polarization generalize between regions: Across 12 OECD countries, the USA has seen the largest increase in affective polarization over the last four decades[5], but across 53 countries it remains middle of the pack when measured in absolute terms[2]. Similarly, recent studies have shown how polarization interventions differ between countries, with conflicting outcomes regarding the effect of deactivating social media on political polarization in the USA and in Bosnia-Herzegovina[11,12]. Despite this, the majority of studies,

[1]Department of Network & Data Science, Central European University, Vienna, Austria. [2]Department of Mathematics, City St George's, University of London, London, UK. [3]Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Venice, Italy. [4]The New Institute Centre for Environmental Humanities, Venice, Italy. [5]Department of Computer Science, Sapienza University of Rome, Rome, Italy. [6]School of Social Science and Technology, Technical University of Munich, Munich, Germany. [7]The Alan Turing Institute, British Library, London, UK.
✉e-mail: max.falkenberg@protonmail.com; a.baronchelli.work@gmail.com

including in high impact factor journals[13–23], still only consider polarization in the United States.

Research on affective polarization has traditionally been carried out using survey data, often measuring an individual's self-reported attitudes towards their out-party on a "feeling thermometer"[4]. However, the rise of the internet has seen social media emerge as an alternate public for the study of polarization[24,25].

The role of social media in driving polarization is disputed[18,19,26–28]; many argue that polarization on social media simply mirrors the underlying polarization of our societies. However, the richness and availability of social media data has made it an invaluable forum for studying the mechanisms of polarization[15,29–31], depolarization[32,33], its evolution over time[16,34], and for testing potential interventions and countermeasures[18]. Twitter (currently X) is particularly important for polarization research given its outsized influence on politicians[35] and journalists[36,37].

The most commonly studied form of polarization on social media is interactional polarization[38]—sometimes referred to as structural[39] or social network polarization[40]—which looks at how the interaction patterns between ideological groups are segregated[13,41]. However, many social media studies do not use this language, referring to patterns of network homophily as "polarization" in a general sense[27].

Affective polarization has also been studied on social media since it allows for a direct measurement of partisan animosity through the analysis of inter-group messages (e.g., using toxicity analysis, or dictionaries of polarized language[42]). Most prominently, researchers have studied how moral-emotional language is used both across and within political groups[43], and how this language grabs our attention, drives increased engagement with like-minded content (both organically and due to the design of social media algorithms), and results in the reinforcement of political group identities (see the MAD[44] and SPIR models[45]). While these studies find that moral-emotional language increases social media engagement in general[43,46], animosity towards ones political out-group is consistently the strongest predictor of increased social media engagement[14].

In the current study we focus on the use of toxic language, a common approach across many studies of political communication online[47,48]. However, because toxic language is explicitly defined as "rude, disrespectful, or unreasonable [language] likely to make someone leave a discussion" [49], it is possible that engagement patterns with toxic language may differ from the broader category of moral-emotional language.

Other studies have also looked at how interactional polarization aligns with affective polarization, for instance in relation to US elections[50,51], the US far-right[52], UK politicians[53], in relation to Covid-19[54,55], and following violent events in Israel[38]. However, these examples remain limited to individual countries and contexts. For this reason, it is important to investigate how different forms of polarization are related across countries in order to identify common trends and potential outliers.

One of the reasons for the lack of cross-country studies on social media is the difficulty acquiring sufficiently large datasets which are not keyword specific[56]. Here we overcome this limitation by using a complete Twitter dataset which includes all public interactions across a 24 h period[57]. Coupled with a second dataset of known elected politicians on Twitter[58], we are able to study how affective polarization aligns with interactional polarization across nine countries (Canada, France, Germany, Italy, Poland, Spain, Turkey, UK, USA) covering seven languages.

In the remainder of this paper, we first give an overview of the datasets studied and visualize the network of politicians on Twitter. Then, we compute the spectrum of interactional polarization, showing that it broadly aligns with a left-right political dimension (with the exception of Germany, where the primary divide is establishment-populist). Grouping users on each side of the structural divide, we show that for all nine countries out-group interactions are more toxic, but receive lower engagement, than in-group interactions. We then show that highly toxic content receives lower engagement than low toxicity content and that only a minority of interactions from politically engaged accounts are with other politically engaged accounts. We identify that interactions between politically engaged accounts, and with apolitical accounts, share a common ally-enemy structure across political groups, and that interactions between partisans are rarely reciprocated. Comparing political and apolitical interactions, we show that political content is consistently more toxic than apolitical content. Finally, we contextualize our work and discuss its implications for the wider study of political polarization.

Results

To study polarization across countries on Twitter (currently X), we use a complete dataset of all public Twitter posts (including retweets) from a 24 h period in September 2022 (see "Methods"), totaling 375 million tweets (see "Data Availability"). In this paper we are interested in political interactions. To identify these, we use a second dataset of known elected politicians from 26 different countries[58] to label all content involving politically engaged Twitter users. This includes all posts authored by politicians, all interactions with those politicians (excluding likes), and all posts and interactions by the accounts who at any point have engaged with these politicians (these posts do not themselves have to mention an elected politician). Previous studies of political communication online have included interactions with news outlets when constructing political interaction networks. However, recent research suggests that most social media users have politically moderate news diets[59–62], and most news they interact with is apolitical in nature (e.g., relating to sports or food recipes[61]). Since these factors may suppress the identification of interactional polarization on Twitter, we do not include news outlet mentions in the construction of our political interaction networks.

Having identified political interactions on Twitter, we focus on the nine countries where there is sufficient engagement with politicians across the 24 h period (at least 5000 unique user pairs between politicians and their retweeters) to enable a robust comparison of affective and interactional polarization. This threshold is determined experimentally and is required since the computed latent ideology is overly dependent on the Twitter interactions of a small number of highly active accounts in countries where the total number of unique user pairs is small. The resulting distribution of ideology scores is often not representative of the diverse range of political views found in a country. Countries meeting the required user pair threshold are Canada, France, Germany, Italy, Poland, Spain, Turkey, the United Kingdom, and the United States. The filtered dataset of political Twitter interactions is broken down in detail in Supplementary Note 1 (SN1) and Fig. S1. In its totality, the filtered dataset includes the interactions of 140 thousand unique users with 1837 unique elected politicians across the 24 h period.

Visualizing politicians on Twitter

To start, we visualize the network of political Twitter interactions to gain an intuitive understanding of the structure of multi-national political communication.

Since our aim is to identify politicians who are ideologically aligned, our polarization analysis uses retweet networks, a common approach in many Twitter-based polarization studies[13,34,39]. We focus on retweets since they are generally evidence of a Twitter user endorsing the message of the original poster[63], as opposed to other Twitter interactions (mentions, quotes or replies) which may indicate a positive, negative or neutral relationship between the two users. Retweets also uniquely refer to a single user as opposed to replies and mentions where multiple users may be referenced. Using the retweet framework, if two politicians share a large number of common

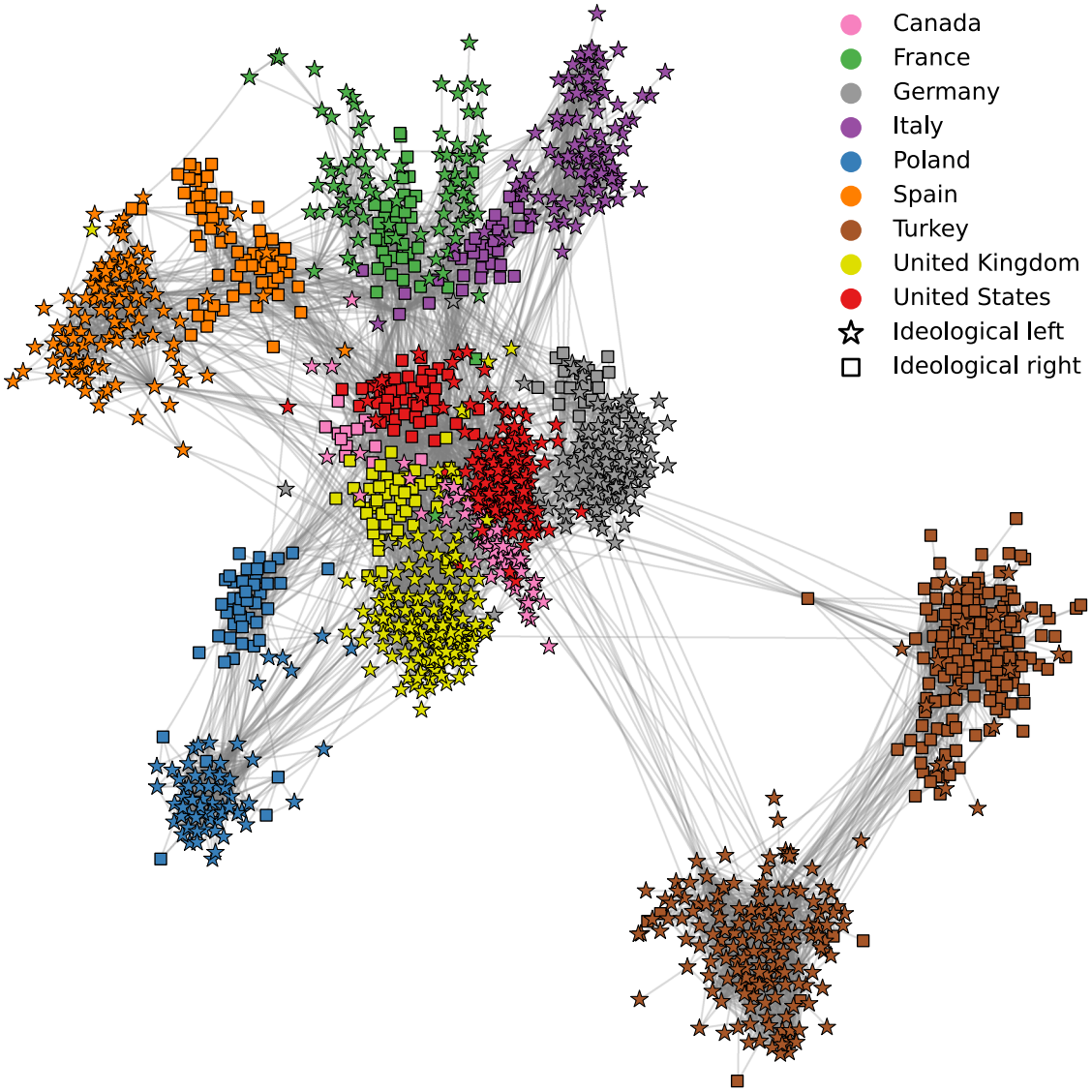

Fig. 1 | The network of political Twitter interactions is segregated across countries and polarized within then. The figure depicts a co-retweet network where nodes correspond to individual politicians and an edge is drawn between two nodes if those politicians share common retweeters (see "Methods"). Square nodes correspond to politicians classified as members of the ideological right, star nodes correspond to politicians classified as members of the ideological left, as determined using the latent ideology (see Fig. 2). The network visualization is produced using a force directed drawing algorithm (see "Methods"), in which repulsive forces are applied between pairs of nodes to push nodes apart, and attractive forces are applied to any pair of nodes which are connected by an edge. Nodes are colored according to the country of each elected politician. Canada: Pink, France: Green, Germany: Gray, Italy: Purple, Poland: Blue, Spain: Orange, Turkey: Brown, United Kingdom: Yellow, United States: Red.

retweeters they likely share similar ideological views or the same partisan identity (see for example[34]).

Figure 1 shows the co-retweet network of elected politicians from the nine countries studied. Each node in the network corresponds to a single elected politician, colored according to country. Two politicians are connected by an edge if they share at least two common retweeters. To avoid spurious connections, we exclude a small number of retweets from highly active accounts (possibly spam) who engage with many politicians. We note that the lower (2) and upper (10) bounds for the number of interactions are used to ensure visual clarity and only apply to the visualization in Fig. 1. Results in the remainder of the paper are not restricted by these bounds.

From the network visualization we make three qualitative observations: (1) The political Twitter discussion in each country is largely self-contained, separate to the political discussion from other countries. (2) Within each country, there is a clear separation between politicians classified as members of the political left and classified as members of the political right, forming well defined clusters. (3) Across the nine countries studied, Anglophone countries (US, UK, Canada) are at the center of the political retweet network. These countries exhibit some overlap of their political factions: The US left (right) is structurally closer to the Canadian left (right), than to the US right (left). In SN1 and Fig. S1 we provide evidence that these countries share a larger fraction of common users than country pairs which do not share the same language, but show that, in general, very few users interact with politicians from more than one country.

Interactional polarization across countries

To formalize our observation of polarized political Twitter networks, we now measure the spectrum of interactional polarization in each of the nine countries studied. For each country, we construct a bipartite network between the country's elected politicians active on Twitter (the "influencers"), and all remaining Twitter users who retweet those politicians (the "retweeters"). Connections between politicians, and

between retweeters, are not required to compute the spectrum of interactional polarization and are ignored in the current analysis.

From each bipartite network, we compute a one-dimensional spectrum of ideological scores using the latent ideology method, originally developed for follower networks in ref. 64, adapted to retweet networks in ref. 13, and applied using elected politicians as the set of influencers in ref. 34. A precise mathematical formulation for the latent ideology is provided in the "Methods", where we also discuss its extension to a second structural dimension. Intuitively, the method produces a one-dimensional ordering where Twitter users who retweet similar sets of politicians are close to each other in the ordering. Exact ideological scores produced are arbitrary and should not be compared across networks. Here, we rescale the derived ideological scores so that the two dominant peaks of the ideology distribution align with scores of −1 and +1 respectively (robustness checks using an alternate rescaling are shown in SN1 and Figs. S2 and S3).

The distribution of user ideology scores for each country is shown in Fig. 2. The histogram is shaded according to the modal political party of the politicians retweeted by users in the binned range of ideology scores. Users who do not retweet a unique political party, or whose modal retweeted party received little engagement, are not shaded.

Figure 2 shows that political Twitter interactions are polarized in each of the nine countries, with a bimodal (or multi-modal) distribution of ideology scores. In general, retweeters who align with a specific political party are found in only one of the two dominant peaks in the ideology distribution of each country, not both. For example, in Canada (Fig. 2a) retweeters who align with the left-leaning Liberal party are found in the left peak with ideology scores less than 0. Conversely, retweeters who align with the right-leaning Canadian Conservative party are found in the right peak with ideology scores greater than 0. Similarly, in the USA, most users who align with the Democrats are found in the left peak, whereas most users who align with the Republicans are found in the right peak. In Poland, we find the center-right Platforma Obywatelska party (PO; Civic Platform) in the left peak and the populist-right Prawo i Sprawiedliwość party (PiS; Law & Justice) in the right peak. Both parties are on the political right, but in a relative sense PO is further left than PiS.

The only case where the latent ideology does not align with the left-right dimension is Germany, where the primary structural divide is along the establishment-populist dimension. Users who retweet politicians from political parties in the governing coalition have ideological scores less than zero, whereas most users from the Left party (LP) and the far-right Alternative for Germany (AfD) have ideological scores greater than zero. This merger is discussed in SN1 and relates to the unified AfD/LP position criticizing the German government's stance on the Russia-Ukraine war.

Based on these observations, in the following we refer to the ideological left as all influencers (politicians) and retweeters with an ideology score less than 0, and the ideological right as all politicians and retweeters with an ideology score greater than 0. Note, however, that references to the left-right political spectrum are used loosely; individual countries have their own political nuances. The choice of a one-dimensional space aims to make the comparison across countries as transparent as possible, and extending the analysis to two dimensions shows that this approach successfully captures the essential structure of political interactions in most cases (see Fig. S4).

Returning to the ideology distributions shown in Fig. 2, there are some cases where the retweeters of a political party do not align with the rest of their party in the interaction network. The best example of this is in the US (Fig. 2i) where a number of users labeled as Republicans are shown on the left, and a number of users labeled as Democrats are shown on the right. This reflects users who retweet party outliers. In the case of the US, most users whose modal party is Republican but have an ideological score less than zero are retweeters of Liz Cheney (an elected Republican at the time of our data collection), whereas most users whose modal party is Democrat but have an ideological score greater than zero are retweeters of Tulsi Gabbard (an elected Democrat; defected to the Republicans in October 2022, after our data collection period). Both politicians are known outliers from the dominant position of their parties, and their structural alignment with opposition parties on Twitter has been noted previously in work studying the US far-right[52]. The alignment of these politicians with the political opposition is also demonstrated by investigating the second dimension of the latent ideology, as shown in Fig. S4, where we also discuss political outliers from other countries.

Out-group interactions are more toxic than in-group interactions

We have shown that in each of the nine countries studied the network of political Twitter interactions is structurally polarized, in most cases along a broadly left-right spectrum. We now ask whether this spectrum of interactional polarization aligns with affective polarization, referring to out-group hostility[23].

As we have seen, defining groups according to political party is limiting due to the presence of party-outliers (e.g., Liz Cheney and Tulsi Gabbard in the US). Hence, we use an interactional approach, defining an "in-group" interaction as any Twitter mention where both users (the mentioner and mentionee) are classified as members of the same ideological group, i.e., both from the left (scores < 0) or right (scores > 0). Conversely, we refer to a Twitter interaction as "out-group" when the two users are classified as members of opposed ideological groups (one left, one right). We acknowledge that this grouping may oversimplify the political reality of some countries, and may miss some of the ideological nuance across factions within a party, but we consider this approach an acceptable approximation in order to ensure a consistent methodology throughout; for alternative approaches to studying polarization in multi-party contexts see refs. 65–67.

To measure affective polarization, we calculate the toxicity of original posts (not retweets) which include a mention between ideologically labeled users. For English, French, Italian and Spanish language posts we compute toxicity scores twice, once using Google Perspective API[49], and a second time with Detoxify using BERT sentence classifiers as a robustness check (see Fig. S5). For German and Polish posts we compute toxicity scores with Google Perspective API only; these languages are not compatible with Detoxify. For Turkish we compute toxicity scores with Detoxify only; Turkish is not compatible with the Perspective API. In the SI, we show that our results are robust using either toxicity model, and are robust to an alternate rescaling of the ideological scores.

The derived toxicity scores from both models fall in the range [0, 1]. Posts with scores near 0 are the least likely to be considered toxic by a human labeler. Conversely, posts with scores near 1 are the most likely to be considered toxic by a human labeler. To ease comparison between countries, in the following we analyze posts according to their toxicity quantile rather than raw toxicity score.

Figure 3a shows boxplots of the bootstrapped difference in the median toxicity quantile (in the range [0, 1]) of out-group interactions less the median toxicity quantile of in-group interactions for each country. Here, if the median in-group toxicity quantile for a country is 0.4, and the median out-group toxicity quantile is 0.6, the value shown on Fig. 3a will be 0.6 − 0.4 = 0.2. The individual toxicity quantiles for in-group interactions, and for out-group interactions, are shown in Fig. S6. Comparing the toxicity distributions using a two-sided non-parametric Mann-Whitney U test (see "Methods"), out-group interactions are significantly more toxic than in-group interactions. Full statistical reporting is provided in Supplementary Note 2 (SN2). This result shows that in each of the nine countries studied, out-group interactions, defined based on the interaction network, are more toxic

https://doi.org/10.1038/s41467-024-53868-0

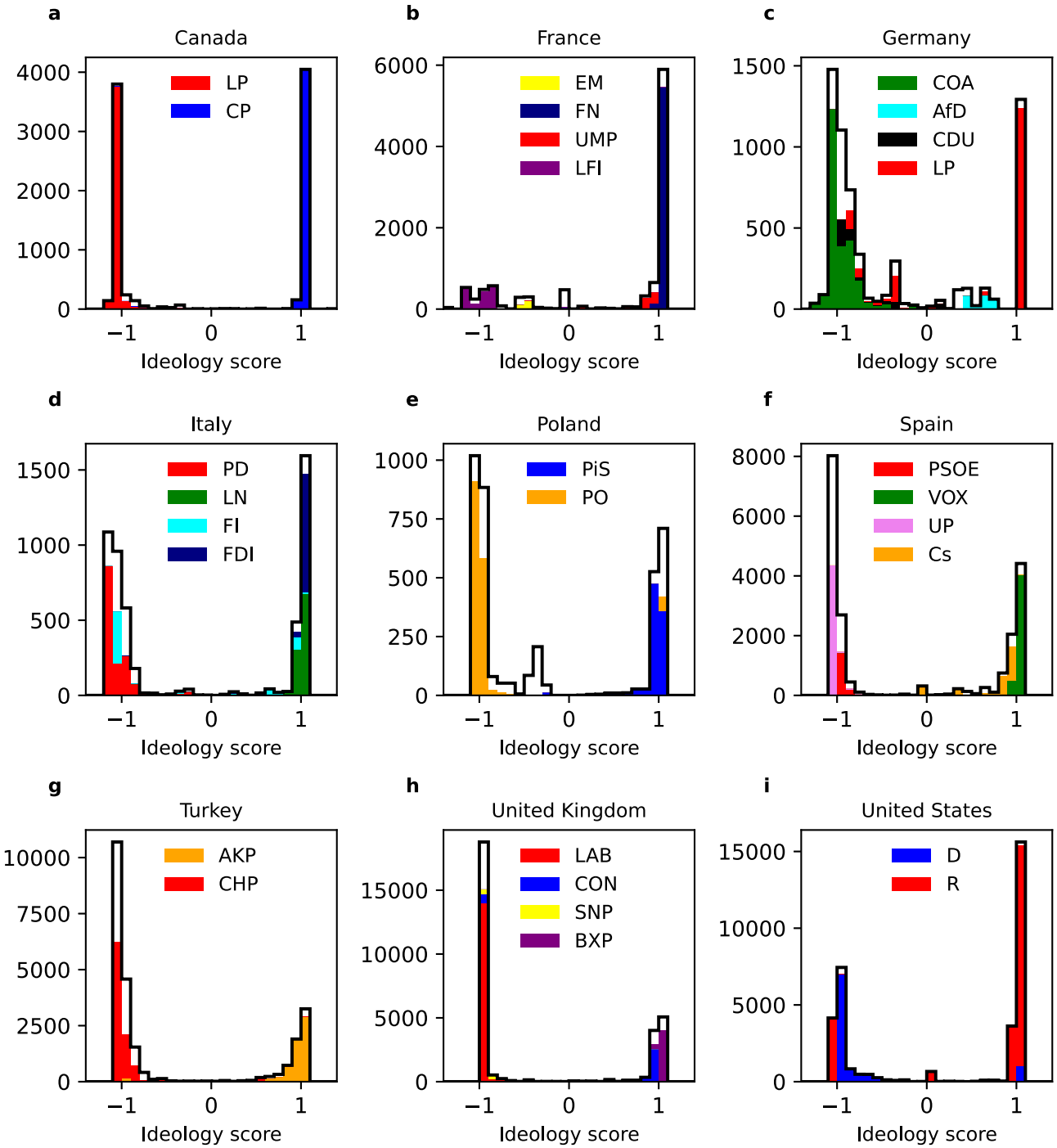

Fig. 2 | Political interaction networks are structurally polarized in each of the nine countries studied. We compute the latent ideology of Twitter users based on their retweet interactions with elected politicians from each country (see "Methods"). In each subfigure, the histogram outlined in bold shows the number of Twitter users with an ideology score in the binned range. Colored bars show the modal political party of users in the binned ideology range. Area in white corresponds to users without a unique modal political party or interacting with other political parties. **a** Canada: {LP: Liberal Party (Red), CP: Conservative Party (Blue)}. **b** France: {EM: En Marche (Yellow), FN: Rassemblement National (Navy), UMP: Les Républicains (Red) LFI: La France Insoumise (Purple)}. **c** Germany: {COA: Coalition (Sozialdemokratische Partei Deutschlands/Freie Demokratische Partei/Bündnis 90/Die Grünen) (Green), AfD: Alternative für Deutschland (Cyan), CDU: Christlich Demokratische Union Deutschlands/Christlich-Soziale Union in Bayern (Black), LP: Die Linke (Red)}. **d** Italy: {PD: Partito Democratico (Red), LN: Lega Nord (Green), FI: Forza Italia (Cyan), FDI: Fratelli d'Italia (Navy)}. **e** Poland: {PiS: Prawo i Sprawiedliwość (Blue), PO: Platforma Obywatelska (Orange)}. **f** Spain: {PSOE: Partido Socialista Obrero Español (Red), VOX: Vox (Green), UP: Podemos (Pink), Cs: Ciudadanos (Orange)}. **g** Turkey: {AKP: Adalet ve Kalkınma Partisi (Red), CHP: Cumhuriyet Halk Partisi (Orange)}. **h** United Kingdom: {LAB: Labour (Red), CON: Conservative (Blue), SNP: Scottish National Party (Yellow), BXP: Brexit Party/UK Independence Party/Reform (Purple)}. **i** United States: {D: Democrats (Blue), R: Republicans (Red)}.

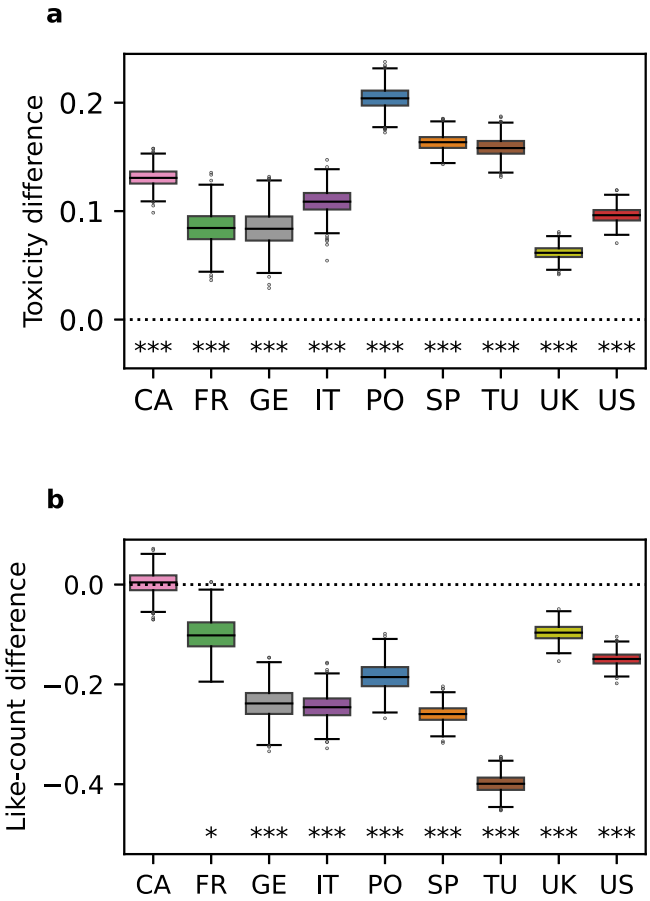

Fig. 3 | Out-group interactions are more toxic, but receive lower engagement, than in-group interactions. a Boxplots for the bootstrapped difference between the median out-group interaction toxicity quantile, less the median in-group interaction toxicity quantile for each of the nine countries. **b** Boxplots for the bootstrapped difference between the mean log likes received by an out-group post relative to the mean log likes received by an in-group post. Boxplots show the bootstrapped median, interquartile range (IQR), whiskers for 1.5 times the IQR from the hinge, and points for outliers (see "Methods"). Stars indicate statistical significance using a two-sided non-parametric Mann-Whitney U test on the full distribution (see "Methods"): $p < 0.05$: *, $p < 0.01$: **, $p < 0.001$: ***. Full statistical reporting including sample sizes, test statistics, exact p-values and boxplot element values are provided in Tables S1–S9 in Supplementary Note 2 (SN2). Country abbreviations and colors: CA Canada (Pink), FR France (Green), GE Germany (Gray), IT Italy (Purple), PO Poland (Blue), SP Spain (Orange), TU Turkey (Brown), UK United Kingdom (Yellow), US United States (Red).

than in-group interactions, demonstrating that affective and interactional polarization are aligned. Expanded results showing the distribution of raw toxicity scores for in-group and out-group interactions are shown in Fig. S7, and are consistent with the results shown in Fig. 3a using the median toxicity quantile.

Our results are robust if we only consider posts authored by accounts classified as members of the political left, or accounts classified as members of the political right (see Fig. S6b, c). For all nine countries, the political left are more toxic when interacting with the political right than when interacting with the political left (statistically significant in each case). Similarly, the political right are more toxic when interacting with the political left than when interacting with the political right in eight of the nine countries (statistically significant in each case). The UK is the only outlier where accounts classified as members of the political right are more toxic when interacting with the right than when interacting with accounts classified as members of the left; this is likely due to toxic interactions between the UK right (Conservative party) and far-right (e.g., Brexit party, Reform party), but may also be due to animosity between party factions which recent research suggests can be as extreme, if not worse than, out-party animosity[68]. In our view, these differences are noteworthy since they stress the importance of not over-generalizing results, and of repeating experiments across a large number of countries.

Finally, we assess whether posts including an out-group interaction authored by accounts classified as members of the political right and more toxic than those authored by accounts classified as members of the political left, or vice versa. This is important given that some research has suggested that out-group discrimination and animosity are predominantly a feature of the political right, whereas others argue that partisans exhibit these traits independent of political ideology[69]. Across the nine countries studied, we do not find a consistent trend suggesting that the political left are more toxic than the political right, or vice versa (see Fig. S6d). Statistically significant differences are only detected in Poland, where right-to-left interactions are more toxic than left-to-right, and in the UK, where left-to-right interactions are more toxic than right-to-left. Future work is needed to confirm and generalize these results given that out-group interactions are rare relative to in-group interactions in our dataset.

Out-group interactions receive lower engagement than in-group interactions

Having found that affective polarization aligns with interactional polarization, we now ask whether out-group interactions receive lower engagement than in-group interactions. Figure 3b shows boxplots of the bootstrapped difference in the mean log like-count ($\log_2$[likes + 1]; likes recorded 10 min after a post first appeared online, see "Methods") received on posts with an out-group mention and the mean log like-count received on posts with an in-group mention. Here, the mean is used rather than the median since the like-distribution is fat-tailed, and between 66% (Poland) and 78% (United States) of all posts receive 0 likes in the first 10 min after posting. The logarithm of the like-count is used to avoid a small number of posts with very large engagement dominating the mean; the +1 avoids errors due to posts with zero likes. The panel shows that out-group posts receive lower engagement than in-group posts (result not statistically significant in Canada). This result is robust if we only consider posts authored by the accounts classified as members of the political left or the political right (see Fig. S8). Across all nine countries, posts authored by accounts classified as members of the political left mentioning another user from the political left receive higher engagement than posts mentioning a user classified as being from the political right (statistically significant in each case). For posts authored by accounts classified as members of the political right, mentions of another user from the political right receive higher engagement than mentions of a user classified as being from the political left in seven of the nine countries (no significant difference in Canada and France; see SN2).

These results suggest that out-group interactions receive lower engagement than in-group interactions. Consequently, given the realities of an attention economy, this is arguably an incentive for Twitter users to prioritize in-group interactions over out-group interactions.

One possible reason for lower out-group engagement may be the confounding factor of post toxicity. Figure 4 shows boxplots for the bootstrapped mean engagement (log-likes) received by the most toxic posts in each country, less the mean engagement received by lower toxicity posts. The figure shows that for all nine countries posts with high toxicity receive lower engagement than low toxicity posts (result for Germany not statistically significant; see SN2). This may be an authentic reflection of Twitter-users' behavior, showing that users are less likely to interact with toxic posts. However, it is also possible that this difference is the result of content moderation policies reducing the visibility of, or removing, offensive posts.

The majority of Twitter interactions are with apolitical accounts

Our focus thus far has been on political Twitter interactions. However, there is substantial evidence to suggest that social media is primarily used for non-political purposes; most social media users are not politically engaged[59]. Here, we study the differences in how politically engaged Twitter users interact with each other, as opposed to with non-politically engaged users.

To achieve this, we compute the "quote-ratio", a metric of cohort-level endorsement defined as the ratio between the number of times a fixed cohort of ideologically aligned users mention a given account in an original tweet (quote tweet, original tweet, or reply), normalized by the number of times the account is mentioned by the same cohort in any tweet (including retweets). This metric was developed in ref. 52 based on the premise that retweets are generally indicative of an endorsement on Twitter[63], whereas non-retweet mentions can be used in a positive, negative, or neutral manner. Hence, if a cohort of ideologically aligned users frequently mention, but never retweet, an account, then the members of the cohort likely disagree with the views of the mentioned account. In contrast, given that retweets are more common than other interaction types, accounts which are disproportionately retweeted are generally seen as endorsed by members of the cohort. The efficacy of this metric is demonstrated in ref. 52 where the US far-right are shown to have a low quote-ratio when mentioning Republican politicians and right leaning media sources, and a high quote-ratio when mentioning Democrat politicians and left leaning media sources.

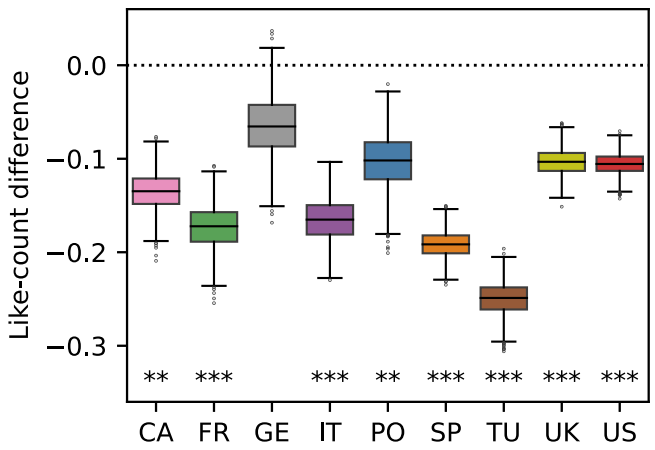

Fig. 4 | High toxicity interactions receive lower engagement than low toxicity interactions. Boxplots show the bootstrapped mean log like-count received by high toxicity posts, minus the mean log like-count received by low toxicity posts. High toxicity interactions are defined as the top 10% most toxic posts in each country. Boxplots show the bootstrapped median, interquartile range (IQR), whiskers for 1.5 times the IQR from the hinge, and points for outliers (see "Methods"). Stars indicate statistical significance using a two-sided non-parametric Mann-Whitney U test on the full distribution (see "Methods"): $p < 0.05$: *, $p < 0.01$: **, $p < 0.001$: ***. Full statistical reporting including sample sizes, test statistics, exact p-values and boxplot element values are provided in Tables S1–S9 in SN2. Country abbreviations and colors: CA Canada (Pink), FR France (Green), GE Germany (Gray), IT Italy (Purple), PO Poland (Blue), SP Spain (Orange), TU Turkey (Brown), UK United Kingdom (Yellow), US United States (Red).

Figure 5 shows the quote-ratio computed using users classified as members of the ideological left as the mentioning cohort (vertical axis) and users classified as members of the ideological right as the mentioning cohort (horizontal axis), broken down according to the group of users mentioned. Figure 5a shows the quote-ratio for mentioned users who do not engage with elected politicians on Twitter and are therefore not assigned an ideological score. We refer to this cohort of users as "apolitical" accounts, but we stress that some of these interactions may still be political in nature (Twitter posts can be political without explicitly mentioning elected politicians). However, this cohort also includes a large number of interactions with content which is not political in nature such that, on average, this cohort is expected to be less explicitly political than the cohorts identified using the latent ideology in Fig. 2. Figure 5b, c show the quote-ratio for mentioned accounts classified as from the ideological left and right respectively. Each panel is averaged across the nine countries studied, with each country equally weighted (individual countries are shown in Fig. S9).

The figure shows that, when interacting with accounts who are not politically engaged (panel a), most accounts are disproportionately retweeted (i.e., endorsed) by both the accounts classified as members of the ideological left and those classified as members of the right. These apolitical interactions represent the majority of interactions in each country (between 89% and 96%; see SN1). In contrast, accounts from the left (panel b) have a low quote-ratio when mentioned by other accounts from the left (i.e., the left endorse the left), but a large quote-ratio when mentioned by accounts from the right (i.e., the right mention, but do not endorse, the left). The reverse is also true, with mentioned accounts classified as from the right (panel c) having a low quote-ratio when mentioned by other accounts from the right, but a high quote-ratio when mentioned by accounts classified as from the left.

This pattern is largely robust at the individual country level, see Fig. S9, revealing the common ally-enemy structure of political

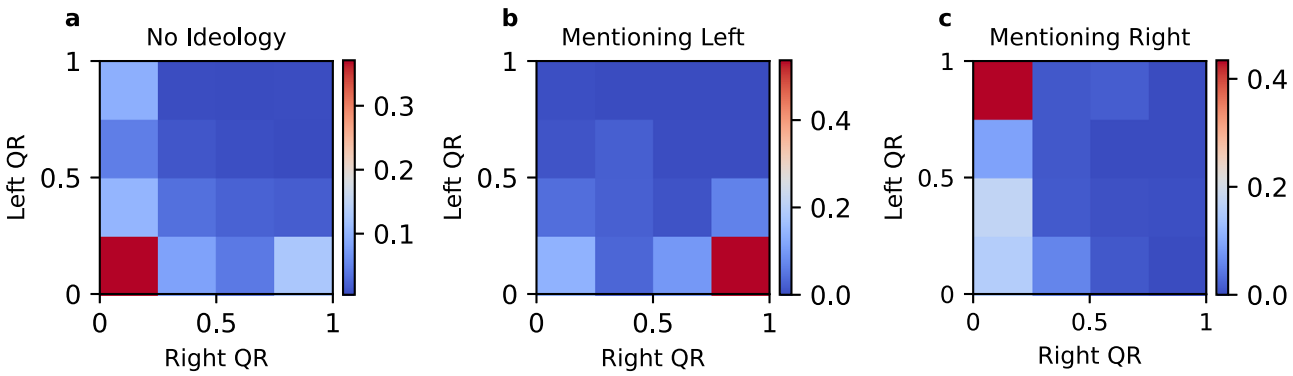

Fig. 5 | Accounts classified as members of the political left and right differentiate between allies and enemies in their interaction patterns, but treat apolitical accounts equally. Binned quote-ratios (QR) for users mentioned who (**a**) are not classified as having a political ideology, (**b**) are classified as having a left-leaning ideology, and (**c**) are classified as having a right-leaning ideology. The quote-ratio is computed twice for each mentioned group; once with users who are classified as members of the ideological left as the mentioners (y-axis: Left QR) and once with users who are classified as members of the ideological right as the mentioners (x-axis: Right QR). Subpanels show the mean computed across all nine countries equally weighted; individual countries in Fig. S9.

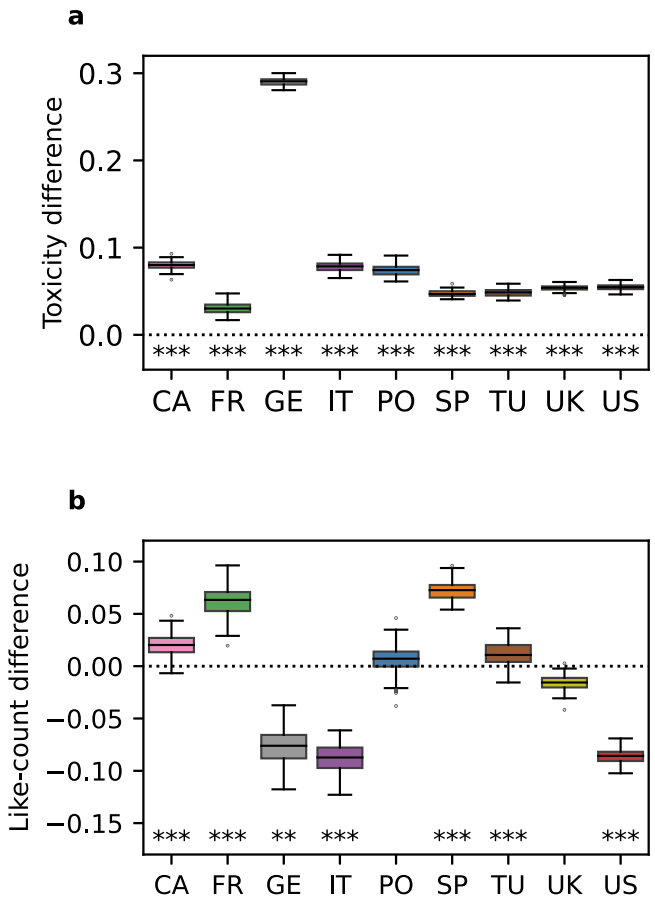

Fig. 6 | Political interactions are more toxic than apolitical interactions, but differences in engagement vary by country. a Boxplots for the bootstrapped difference between the median interaction toxicity quantile between politically engaged accounts, less the median interaction toxicity quantile mentioning an apolitical account for each of the nine countries. **b** Boxplots for the bootstrapped difference between the mean log likes received by a political interaction, less the mean log likes received by a post mentioning an apolitical account. Boxplots show the bootstrapped median, interquartile range (IQR), whiskers for 1.5 times the IQR from the hinge, and points for outliers (see "Methods"). Stars indicate statistical significance using a two-sided non-parametric Mann-Whitney U test on full distribution (see "Methods"): $p < 0.05$: *, $p < 0.01$: **, $p < 0.001$: ***. Full statistical reporting including sample sizes, test statistics, exact p-values and boxplot element values are provided in Tables S1–S9 in Supplementary Note 2 (SN2). Country abbreviations and colors: CA Canada (Pink), FR France (Green), GE Germany (Gray), IT Italy (Purple), PO Poland (Blue), SP Spain (Orange), TU Turkey (Brown), UK United Kingdom (Yellow), US United States (Red).

interactions across countries. Mentions of the political right in France and Germany are outliers. In the case of Germany, this is likely because the main structural divide observed in Fig. 2c is between establishment and populist parties, not right and left. In the case of France, this is likely due to relatively low activity from French left leaning parties (see Fig. 2b).

Political interactions are more toxic than apolitical interactions

We now assess whether interactions between politically engaged accounts (either in- or out-group) are more or less toxic than interactions mentioning one of the accounts from the apolitical cohort.

Figure 6a shows that mentions between accounts which are both politically engaged are more toxic than posts mentioning an apolitical account in all nine countries. Regarding engagement, Fig. 6b shows that political mentions receive higher engagement than apolitical mentions in Canada, France, Spain and Turkey, but receive lower engagement in Germany, Italy, and the United States. No statistically significant difference in engagement is detected in Poland or the United Kingdom (see SN2).

These results highlight that not only are political out-group interactions more toxic than political in-group interactions, as shown in Fig. 3, but political interactions are, in general, more toxic than apolitical interactions. In Fig. S10, we repeat the analysis shown in Fig. 4 for apolitical mentions, showing that high toxicity posts receive lower engagement than low toxicity posts.

Out-group interactions are rarely dyadic

Our analysis has not considered whether interactions are unidirectional or dyadic (i.e., if user A mentions user B, does user B then mention user A?). In Fig. S11, we show that across all countries dyadic interactions are rare. On average, for in-group user pairs, 8.2% of interactions are dyadic, whereas for out-group interactions 2.2% are dyadic. For every country, in-group interactions are more likely to be dyadic than out-group interactions. This demonstrates that, across countries, out-group conversation with political opponents, as opposed to unidirectional broadcasting, is rare.

A natural question is whether there is a difference in the toxicity of unidirectional interactions as opposed to dyadic interactions. Unfortunately, we do not observe enough dyadic interactions to allow for a robust analysis of dyadic toxicity in the current study. This question should be investigated in future work.

Users classified as members of the political right reference lower reliability news outlets than users classified as members of the political left

Many tweets by politically engaged accounts reference news media outlets[70]. In Fig. S12, we show that for the six countries (Canada, France, Germany, Italy, UK, USA) where we have news media reliability ratings provided by NewsGuard (see "Methods"), accounts classified as members of the political right reference lower reliability news outlets than accounts classified as members of the political left (statistical reporting in SN2 and Table S18). These results extend previous research investigating differences in the reliability of media sources shared by partisans[71,72].

Discussion

We have shown that there are common patterns of partisan animosity online, with affective and interactional polarization aligning across nine countries (Canada, France, Germany, Italy, Poland, Spain, Turkey, UK, USA) and seven languages (English, French, German, Italian, Polish, Spanish, Turkish) on Twitter (currently X). When dividing a country's political interaction network into its two primary structural groups, out-group interactions are more toxic than in-group interactions. These results are robust for all nine countries and for both accounts classified as members of the political left and members of the political right.

We have addressed the multi-faceted nature of polarization by drawing on the strengths of social media research, while also employing insights from political science. Specifically, our analysis classifies the ideology of Twitter users through their partisan association, identified via their endorsement of elected politicians. Our results show how the supporters of a given political party typically cluster in a single group, structurally separated from their political opponents. However, our results also reveal how partisan nonconformists are, essentially, treated as members of the political opposition. This behavior is observed across a number of countries (see SN1), raising the worrying prospect that, online, there is no political middle ground.

Our social media lens shows how out-group interactions generally draw lower engagement than in-group interactions. The observation that higher toxicity mentions receive lower engagement is perhaps surprising given that previous studies have shown how moral-

emotional language increases engagement on Twitter[43], especially if this language targets the political out-group[14]. However, the explicitly harmful nature of toxic language means that it is not clear whether lower engagement with toxic content is an authentic user-driven result, or an artifact of content down-ranking by the Twitter recommendation system (for structural reasons, because the content is toxic, or otherwise). Without losing sight of this limitation, our results do suggest an incentive for politically engaged users to prioritize engagement from their own political group as opposed to from a politically diverse user base. Psychological models suggest that this could induce a reinforcing cycle which may worsen interactional polarization over time[44,45].

In the current study, our primary focus has been on political Twitter interactions. However, politically engaged Twitter users interact with apolitical users as well as with other partisans. Focusing on this divide, we have shown how partisans differ in their interactions with other partisans, as opposed to with individuals who are not politically engaged. We find that there is a common ally-enemy structure in how members classified as part of the political left and right interact with each other and that interactions classified as being with apolitical accounts are structurally similar. These represent the majority of the interactions in our dataset, extending previous work for the USA which showed that most Twitter users are not politically engaged[59].

Comparing mentions of these apolitical accounts to mentions of accounts with an assigned ideological score, we find that political mentions are consistently more toxic than apolitical mentions. But, it is important to caveat these results by acknowledging that our definition of political mentions—those involving users who, due to their engagement with elected politicians, have an assigned ideological score—will not capture all political content on Twitter. Some implicitly political content may be included in the cohort of mentions we refer to as apolitical. However, this cohort will also include many posts which are not political in nature.

Finally, our study shows that despite the prevalence of, often toxic, out-group communication, these interactions are rarely reciprocated. In the social media literature, such structures are often referred to as echo chambers[30,73] given that partisan homophily means that individuals are predominantly exposed to content from like-minded individuals and do not appear willing to engage in active conversation with political opponents. Such structures may be further exacerbated by the algorithm mediated filtering of content, often referred to as filter bubbles[74], which can increase the visibility of politically-aligned content in an individual's feed and may suppress politically discordant content. This is especially likely if these out-group interactions are more toxic than in-group interactions and are, therefore, more likely to be down-ranked (or removed) by a platform's moderation tools. However, recent surveys have found little evidence for the presence of these, so-called, filter bubbles on social media[74,75].

Whether the presence of politically homogeneous communities online is a positive or a negative is unclear; there are arguments to suggest that cross-party communication may reduce affective polarization[76]. However, it also comes at the risk of increased exposure to toxic content and hate speech, and the negative consequences that an individual experiences as a result.

There are limitations to our study which present opportunities for future work. First, the countries we study are largely Western and developed. While the importance of these countries should not be underestimated, there remain open questions as to whether our results generalize to other regions, particularly to countries in the global south. Future work should expand our analysis to a broader set of countries, and may consider whether the magnitude of the observed difference between out-group and in-group toxicity correlates to important societal metrics (e.g., GDP, inequality), as has been carried out in previous cross-national studies on differences between in- and out-group behavior[77,78].

Second, our analysis primarily considers a one-dimensional representation of interactional polarization, with a binary divide between the accounts classified as being from the political left and right—a common approach in many polarization studies on social media[21,34,43,48]—across which affective polarization is measured. Particularly in some multi-party states, this one-dimensional representation of polarization may lose some of the important nuances of a country's political interaction network (e.g., France and Italy; see Fig. S4). However, we have shown that a one-dimensional picture of interactional polarization does capture the primary political divide in most countries. Therefore, to ensure a consistent methodology we retain this one-dimensional representation across the countries studied here but emphasize that future work should consider in greater detail how to compare the structures of political polarization across countries with different political systems and a variable number of political parties. Importantly, this work should further investigate whether there are specific ideological asymmetries (see ref. 21) in certain countries and whether the behaviors observed on the political left differ from the behaviors on the political right.

Third, our study focuses exclusively on Twitter (now X). Future work should consider a similar analysis on other platforms. However, we stress that understanding polarization on Twitter remains critically important: It is one of the most influential social media sites for politicians and journalists[35,37], and results for Twitter will likely have some relevance for Twitter's emerging competitors (e.g., Threads, Bluesky) which use similar interaction mechanisms. In the current study, we have focused primarily on politically engaged accounts which are defined based on their interactions with known elected politicians. However, we acknowledge that some political content, authored by users who do not interact with elected politicians, will not be captured by this definition. Future work should consider alternate methods for identifying political content on social media, for instance using topic models.

Fourth, our study uses data gathered during a single 24 h observation window. While this ensures coherent results from the slow changes in political discourse, it also raises the possibility that results may differ across larger time windows. Specific observations in the current study which may be explained by the short observation window include (1) that the primary structural divide in Germany is between establishment and populist parties, not between the political left and right, (2) the observation of higher in-group toxicity than out-group toxicity on the political right in the UK, and (3) the diversity of political outliers observed in Italy (which was in an election period at the time our data was collected). Observations over longer time periods could clarify whether these results are robust over time, or specific to the 24 h observation window. There is evidence that the structure of political interactions evolve to reflect the changing political landscape of a country (for example in Pakistan[79]). Similar long-term tracking should be tested with cross-country datasets in the future, although we note that the feasibility of such studies is now in question given recent restrictions to academic social media data access[56].

Finally, our study is observational in nature. Our results do not point towards a causal relationship between observed interactional polarization and observed affective polarization. Despite this, our research emphasizes that studying polarization in a siloed manner may be counterproductive since the different forms of polarization may have interdependent mechanisms.

In summary, our findings contribute towards a more unified understanding of how different forms of polarization are related and the extent to which results generalize across countries. This provides context for future work looking at how to reduce partisan animosity which, if left untackled, may have damaging democratic consequences.

Methods

Ethical approval

As no new data was acquired for this study, and because all Twitter data used is from publicly available sources, no ethical approval was sought for the current study.

To ensure the appropriate processing of data and to comply with data protection regulations including GDPR, a Data Protection Impact Assessment threshold test (DPIA; reference number Reference Number DPIA0001277) was completed at the corresponding authors' institution, City University of London (now City St. George's, University of London). The DPIA confirms that the data processing carried out complies with all necessary regulations, and that the matching of datasets is permitted given only elected politicians are matched, and only publicly available data is used.

Data

The Twitter (currently X) data analyzed was acquired by Pfeffer et al.[57]. The dataset includes all public Twitter posts across a 24 h period starting on September 21, 2022 (see also "Data Availability"). Posts were downloaded almost exactly 10 min after they appeared online. Politicians are identified using the dataset in ref. 80 which lists politicians across 26 countries known to be active on Twitter, and their respective political parties. Media reliability scores are calculated using data from NewsGuard.

Network visualization

The network visualization in Fig. 1 is a co-occurence network of elected politicians in the nine countries studied. Each node corresponds to an elected politician who was active during the 24 h period. For visual clarity, we only show politicians who were retweeted by at least two unique users in the 24 h observation window. Two politicians (nodes) are connected by an edge if they were both retweeted by the same users. When constructing the network, we remove edges which are due to highly active accounts, defined as accounts who have retweeted over ten different politicians. This limits the number of spurious edges in the network, which may be due to automated accounts spamming retweets. We stress that these lower and upper bounds on the number of retweets are only applied for the network visualization, and not to the analyses in the remainder of the paper. Finally, for visualization purposes, we remove any politicians who are not part of the giant connected component. The resulting network is drawn manually with a layout derived using ForceAtlas2[81]. This layout uses repulsive forces between all node pairs, pushing nodes apart, and attractive forces between any nodes connected by an edge, pulling connected nodes together. Nodes are colored according to country. The shape of nodes is determined by their ideological score as computed using the latent ideology, see Fig. 2.

Latent ideology

Ideological scores for each country's political Twitter network are derived using the latent ideology method developed in ref. 64. We use the same adaptation applied in ref. 34 for use with a bipartite representation of Twitter retweet networks between a set of m politicians (influencers) and their n retweeters (users). Influencer-influencer connections and user-user connections are ignored when constructing the bipartite network.

We start with an $n \times (m+1)$ matrix $\mathbf{A}$ where each element a_{ij} is the number of times user i has retweeted politician j. For each country, we include all known elected politicians who were active in the 24 h period and were retweeted at least once. In column $m+1$ we include a "dummy politician" who is retweeted by every user (i.e., a columns of 1s) to ensure that the bipartite network is a single connected component. This dummy politician is removed from the analysis once ideological scores have been derived.

Once the matrix $\mathbf{A}$ has been constructed, we compute the matrix normalized according to the number of retweets as $\mathbf{P} = \mathbf{A}(\sum_{ij} a_{ij})^{-1}$. We then define the column vector $\mathbf{r}$ as the sum over the n rows given by $\mathbf{r} = \mathbf{P1}$, and define the row vector $\mathbf{c}$ as the sum over the $m+1$ columns given by $\mathbf{c} = \mathbf{1}^T\mathbf{P}$. Using these row and column vectors we also define the diagonal matrices $\mathbf{D}_r = \mathrm{diag}(\mathbf{r})$ and $\mathbf{D}_c = \mathrm{diag}(\mathbf{c})$. We can, therefore, compute the matrix of standardized residuals of the adjacency matrix as $\mathbf{S} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{rc})\mathbf{D}_c^{-1/2}$, where $\mathbf{rc}$ denotes the outer product of the column vector $\mathbf{r}$ with the row vector $\mathbf{c}$ resulting in an $n \times (m+1)$ matrix. This residual matrix accounts for differences in activity of retweeters and differences in the popularity of individual politicians (i.e., how often each politician is retweeted). Next, single value decomposition is applied to the matrix $\mathbf{S}$ as $\mathbf{S} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^T$ with $\mathbf{U}\mathbf{U}^T = \mathbf{V}\mathbf{V}^T = \mathbf{I}$ and $\mathbf{D}_\alpha$ being the singular values diagonal matrix. The ideological scores of the n users who have retweeted at least one of the m elected politicians is given by the standard row coordinates $\mathbf{X} = \mathbf{D}_r^{-1/2}\mathbf{U}$. In line with previous studies, our primary analysis only considers the first dimension that corresponds to the largest singular value. In Fig. S4, we investigate the extent to which this first dimension of the latent ideology adequately captures the core underlying structure of the political interaction network of each country by investigating the second dimension of the latent ideology, computed using the second largest eigenvalue. Our analysis shows that in most countries the first dimension of the latent ideology does an excellent job of capturing the primary interaction structure on Twitter, but that some additional structure is revealed in some cases (e.g., France and Italy).

We rescale the ideological scores of retweeters such that the two largest peaks in the ideology score distribution align with scores of -1 and $+1$ respectively. This rescaling is only possible if the distribution of ideology scores is multi-modal, which is the case for each of the country-specific retweet networks in the current paper, but is not necesarily true for all social media interaction networks. In Figs. S2 and S3, we show that our results are robust using an alternate rescaling of the ideological scores.

Toxicity analysis

Toxicity analysis is a common method in digital media research for identifying content which is "rude, disrespectful, or unreasonable … [and is] likely to make someone leave a discussion" [49].

There are a range of tools available for the automated detection of toxic content on social media. Here, we primarily use the Perspective API[49], developed by the Jigsaw team at Google, which provides toxicity scores between 0 and 1 corresponding to the probability that the classified content would be labeled as toxic by a human labeler. The Perspective API provides toxicity scores for English, German, French, Italian, Spanish, and Polish. The Perspective API cannot classify comments in Turkish. For Turkish language posts (and for English, French, Italian and Spanish language posts as a robustness check) we compute toxicity scores using Detoxify[82], developed by Unitary. In both cases, posts classified using these models do not require pre-cleaning.

For the analysis of affective polarization in Figs. 3 and 4, we compute the toxicity of all original tweets authored by a Twitter user with an assigned ideological score which mentions a single other user who also has an assigned ideological score for the same country. We only classify posts labeled as being in one of the seven languages covered by our toxicity analysis models. This does not include posts which only include URLs, or are authored in another language. We exclude tweets where multiple users are mentioned. Tweets analyzed include original posts, replies, and quote tweets. Importantly, we do not analyze the toxicity of retweets since the text of the retweet is attributable to the original author and not the retweeter.

For the analysis of the differences between the toxicity of political mentions compared to apolitical mentions (see Fig. 6), we define political mentions as the set of all mentions between accounts with an assigned ideological score (in-group and out-group mentions

merged). For apolitical mentions, we consider all posts authored by any account (both accounts with and without an ideological score) mentioning any account which does not have an ideological score, but which was mentioned by a political account (i.e., one with an assigned ideological score). These correspond to users included in the cohort shown in Fig. 5a. To ensure robust comparisons at the country level, we only consider mentions authored in the primary language of a country. For each country, mentioned accounts are only included if they receive more mentions from that country's cohort of users, than from other country's users. This ensures that apolitical accounts are not included in multiple different country cohorts.

Statistical analysis

Statistical analysis is carried out using a two-sided non-parametric Mann-Whitney U test. The test returns a p-value corresponding to the probability that the two distributions are drawn from the same parent distribution. Under the null hypothesis that the two samples are drawn from the same distribution, we use the standard convention that the null hypothesis can be rejected at three different significance levels: $p < 0.05$: *, $p < 0.01$: **, $p < 0.001$: ***. Data tested using Mann-Whitney U test meets necessary independence assumptions.

Point estimates for the median (mean) of the observables shown in Figs. 3, 4, and 6 are computed using a bootstrapping procedure where the point estimate is sampled 1000 times using a 50% sample of the full distribution with replacement. The distribution of these point estimates is then shown as a boxplot for each country.

Full statistical reporting including sample sizes, test statistics, exact p-values and boxplot element values are provided in Tables S1–S9 in Supplementary Note 2 (SN2).

News media classification

Media reliability data was provided by NewsGuard[83]. The data is proprietary and requires a license for use. The dataset by Newsguard includes a range of news media outlets in the US, UK, Canada, Germany, France and Italy from across the political spectrum and classifies the reliability of each outlet according to a set of journalistic criteria. Each outlet receives a score from 0 to 100 for each of the criteria, assigned by a team of independent journalists. Outlets with a score of 100 are considered the most reliable, whereas outlets with a score of 0 are considered the least reliable; lower scores reflect lower reliability. For the news media outlets classified, Newsguard list their online domains. From the tweets analyzed, we extract URLs and search for domains which correspond to a classified news domain. For each post which includes such a domain we assign the corresponding reliability score as provided by Newsguard. Analysis using media reliability scores provided by Newsguard have been shown to be similar to the results obtained using other media reliability datasets[84].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

No new data was acquired for this study. Three existing datasets were used: (1) The Twitter 24 h dataset collected by Pfeffer et al.[57]. (2) The dataset of politicians active on Twitter collected by Van Vliet et al.[58]. (3) The news media reliability dataset provided by NewsGuard[83]. The Twitter 24 h dataset and the Twitter politicians dataset are publicly available in accordance with Twitter's (currently X's) terms of service in the form of tweet IDs at ref. 85, and in the form of Twitter user IDs at ref. 80 The combined dataset can be produced by acquiring the two individual datasets separately and then labeling user IDs in the Twitter 24 h dataset using the politician user IDs listed in the politicians dataset. The authors of ref. 57 acknowledge that following restrictions to the Twitter API for academics, downloading these tweets using the tweet IDs provided may be difficult. Consequently, the authors encourage anyone interested in the dataset to contact them for collaboration. The NewsGuard media reliability dataset is a proprietary dataset and is not publicly available. Access to the dataset requires a license which can be purchased from NewsGuard[83].

Code availability

All analysis carried out in Python 3.7 using publicly available packages. Statistical analysis carried out using the `mannwhitneyu` function from Scipy 1.14.1. Latent ideology calculation based on previous work in ref. 34. Toxicity analysis carried out using Google Perspective API[49] and Detoxify[82] which are freely accessible online.

References

1. Iandoli, L., Primario, S. & Zollo, G. The impact of group polarization on the quality of online debate in social media: a systematic literature review. *Technol. Forecast. Soc. Change* **170**, 120924 (2021).
2. Orhan, Y. E. The relationship between affective polarization and democratic backsliding: comparative evidence. *Democratization* **29**, 714–735 (2022).
3. McCoy, J., Rahman, T. & Somer, M. Polarization and the global crisis of democracy: common patterns, dynamics, and pernicious consequences for democratic polities. *Am. Behav. Sci.* **62**, 16–42 (2018).
4. Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N. & Westwood, S. J. The origins and consequences of affective polarization in the united states. *Annu. Rev. Polit. Sci.* **22**, 129–146 (2019).
5. Boxell, L., Gentzkow, M. & Shapiro, J. M. Cross-country trends in affective polarization. *Rev. Econ. Stat.* **106**, 557–565 (2024).
6. Roccas, S. & Brewer, M. B. Social identity complexity. *Personal. Soc. Psychol. Rev.* **6**, 88–106 (2002).
7. Ojer, J., Cárcamo, D., Pastor-Satorras, R. & Starnini, M. Charting multidimensional ideological polarization across demographic groups in the United States. *arXiv preprint*, 2311.06096 (2023).
8. Mason, L. "I disrespectfully agree": the differential effects of partisan sorting on social and issue polarization. *Am. J. Polit. Sci.* **59**, 128–145 (2015).
9. Mason, L. *Uncivil Agreement: How Politics Became our Identity* (University of Chicago Press, 2018).
10. Levendusky, M. S. & Malhotra, N. (mis) perceptions of partisan polarization in the American public. *Public Opin. Q.* **80**, 378–391 (2016).
11. Allcott, H., Braghieri, L., Eichmeyer, S. & Gentzkow, M. The welfare effects of social media. *Am. Econ. Rev.* **110**, 629–676 (2020).
12. Asimovic, N., Nagler, J., Bonneau, R. & Tucker, J. A. Testing the effects of facebook usage in an ethnically polarized setting. *Proc. Natl Acad. Sci. USA* **118**, e2022819118 (2021).
13. Flamino, J. et al. Political polarization of news media and influencers on twitter in the 2016 and 2020 us presidential elections. *Nat. Hum. Behav.* **7**, 904–916 (2023).
14. Rathje, S., Van Bavel, J. J. & Van Der Linden, S. Out-group animosity drives engagement on social media. *Proc. Natl Acad. Sci. USA* **118**, e2024292118 (2021).
15. Bail, C. A. et al. Exposure to opposing views on social media can increase political polarization. *Proc. Natl Acad. Sci. USA* **115**, 9216–9221 (2018).
16. Waller, I. & Anderson, A. Quantifying social organization and political polarization in online platforms. *Nature* **600**, 264–268 (2021).
17. Robertson, R. E. et al. Users choose to engage with more partisan news than they are exposed to on Google search. *Nature* **618**, 342–348 (2023).
18. Guess, A. M. et al. Reshares on social media amplify political news but do not detectably affect beliefs or opinions. *Science* **381**, 404–408 (2023).

19. Guess, A. M. et al. How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science* **381**, 398–404 (2023).
20. Nyhan, B. et al. Like-minded sources on Facebook are prevalent but not polarizing. *Nature* **620**, 137–144 (2023).
21. González-Bailón, S. et al. Asymmetric ideological segregation in exposure to political news on Facebook. *Science* **381**, 392–398 (2023).
22. Voelkel, J. G. et al. Interventions reducing affective polarization do not necessarily improve anti-democratic attitudes. *Nat. Hum. Behav.* **7**, 55–64 (2023).
23. Finkel, E. J. et al. Political sectarianism in America. *Science* **370**, 533–536 (2020).
24. Bail, C. *Breaking the Social Media Prism: How to Make our Platforms Less Polarizing* (Princeton University Press, 2022).
25. Lee, J. K., Choi, J., Kim, C. & Kim, Y. Social media, network heterogeneity, and opinion polarization. *J. Commun.* **64**, 702–722 (2014).
26. Nordbrandt, M. Affective polarization in the digital age: testing the direction of the relationship between social media and users' feelings for out-group parties. *New Media Soc.* **25**, 3392–3411 (2021).
27. Kubin, E. & von Sikorski, C. The role of (social) media in political polarization: a systematic review. *Ann. Int. Commun. Assoc.* **45**, 188–206 (2021).
28. Lorenz-Spreen, P., Oswald, L., Lewandowsky, S. & Hertwig, R. A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nat. Hum. Behav.* **7**, 74–101 (2023).
29. Flores, A. et al. Politicians polarize and experts depolarize public support for covid-19 management policies across countries. *Proc. Natl Acad. Sci. USA* **119**, e2117543119 (2022).
30. Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W. & Starnini, M. The echo chamber effect on social media. *Proc. Natl Acad. Sci. USA* **118**, e2023301118 (2021).
31. Törnberg, P. How digital media drive affective polarization through partisan sorting. *Proc. Natl Acad. Sci. USA* **119**, e2207159119 (2022).
32. Chen, T. H. Y., Salloum, A., Gronow, A., Ylä-Anttila, T. & Kivelä, M. Polarization of climate politics results from partisan sorting: evidence from finnish twittersphere. *Glob. Environ. Change* **71**, 102348 (2021).
33. Xia, Y. et al. How the Russian invasion of Ukraine depolarized the Finnish NATO discussion. *EPJ Data Sci.* **13**, 1–12 (2024).
34. Falkenberg, M. et al. Growing polarization around climate change on social media. *Nat. Clim. Chang.* **12**, 1114–1121 (2022).
35. Stieglitz, S. & Dang-Xuan, L. Social media and political communication: a social media analytics framework. *Soc. Netw. Anal. Min.* **3**, 1277–1291 (2013).
36. Hu, M. et al. Breaking news on Twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 2751–2754 (Association for Computing Machinery, 2012).
37. Molyneux, L. & McGregor, S. C. Legitimating a platform: evidence of journalists' role in transferring authority to Twitter. *Inf. Commun. Soc.* **25**, 1577–1595 (2022).
38. Yarchi, M., Baden, C. & Kligler-Vilenchik, N. Political polarization on the digital sphere: a cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Polit. Commun.* **38**, 98–139 (2021).
39. Salloum, A., Chen, T. H. Y. & Kivelä, M. Separating polarization from noise: comparison and normalization of structural polarization measures. *Proc. ACM Hum. Comput. Interact.* **6**, 1–33 (2022).
40. Tokita, C. K., Guess, A. M. & Tarnita, C. E. Polarized information ecosystems can reorganize social networks via information cascades. *Proc. Natl Acad. Sci. USA* **118**, e2102147118 (2021).
41. Bovet, A. & Makse, H. A. Influence of fake news in twitter during the 2016 us presidential election. *Nat. Commun.* **10**, 1–14 (2019).
42. Simchon, A., Brady, W. J. & Van Bavel, J. J. Troll and divide: the language of online polarization. *PNAS Nexus* **1**, pgac019 (2022).
43. Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A. & Van Bavel, J. J. Emotion shapes the diffusion of moralized content in social networks. *Proc. Natl Acad. Sci. USA* **114**, 7313–7318 (2017).
44. Brady, W. J. & Van Bavel, J. J. The mad model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspect. Psychol. Sci.* **15**, 978–1010 (2020).
45. Harris, E., Rathje, S., Robertson, C. E. & Van Bavel, J. J. The SPIR Framework of Social Media and Polarization: Exploring the Role of Selection, Platform Design, Incentives, and Real-World Context. *Int. J. Commun.* **17**, 5316–5335 (2023).
46. Brady, W. J., Jackson, J. C., Lindström, B. & Crockett, M. Algorithm-mediated social learning in online social networks. *Trends Cogn. Sci.* **27**, 947–960 (2023).
47. Avalle, M. et al. Persistent interaction patterns across social media platforms and over time. *Nature* **628**, 582–589 (2024).
48. Mamakos, M. & Finkel, E. J. The social media discourse of engaged partisans is toxic even when politics are irrelevant. *PNAS Nexus* **2**, pgad325 (2023).
49. Google perspective api: Attributes and languages, accessed 16 March 2023. https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US (2023).
50. Grimminger, L. & Klinger, R. Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In *Proc. Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 171–180 (Association for Computational Linguistics, 2021).
51. Saveski, M., Roy, B. & Roy, D. The structure of toxic conversations on twitter. In *Proceedings of the Web Conference 2021*, 1086–1097 (Association for Computing Machinery, 2021).
52. Mekacher, A., Falkenberg, M. & Baronchelli, A. The systemic impact of deplatforming on social media. *PNAS Nexus* **2**, pgad346 (2023).
53. Agarwal, P. et al. Hate speech in political discourse: a case study of UK MPs on Twitter. In *Proc. 32nd ACM Conference on Hypertext and Social Media*, 5–16 (Association for Computing Machinery, 2021).
54. Cinelli, M. et al. Dynamics of online hate and misinformation. *Sci. Rep.* **11**, 22083 (2021).
55. Miyazaki, T., Uchiba, T., Tanaka, K. & Sasahara, K. Aggressive behaviour of anti-vaxxers and their toxic replies in English and Japanese. *Hum. Soc. Sci. Commun.* **9**, 1–8 (2022).
56. Roozenbeek, J. & Zollo, F. Democratize social-media research-with access and funding. *Nature* **612**, 404 (2022).
57. Pfeffer, J. et al. Just another day on Twitter: a complete 24 hours of Twitter data. In *Proc. 17th International AAAI Conference on Web and Social Media*, (Association for the Advancement of Artificial Intelligence, 2023).
58. Van Vliet, L., Törnberg, P. & Uitermark, J. The Twitter parliamentarian database: analyzing Twitter politics across 26 countries. *PLoS ONE* **15**, e0237073 (2020).
59. Wojcieszak, M., Casas, A., Yu, X., Nagler, J. & Tucker, J. A. Most users do not follow political elites on Twitter; those who do show overwhelming preferences for ideological congruity. *Sci. Adv.* **8**, eabn9418 (2022).
60. Eady, G., Nagler, J., Guess, A., Zilinsky, J. & Tucker, J. A. How many people live in political bubbles on social media? evidence from linked survey and Twitter data. *Sage Open* **9**, 2158244019832705 (2019).
61. Wojcieszak, M. et al. No polarization from partisan news: Over-time evidence from trace data. *Int. J. Press/Polit.* **28**, 601–626 (2023).
62. Guess, A. M. (Almost) Everything in moderation: new evidence on Americans' online media diets. *Am. J. Polit. Sci.* **65**, 1007–1022 (2021).
63. Metaxas, P. et al. What do retweets indicate? results from user survey and meta-review of research. In *Proc. International AAAI*

Conference on Web and Social Media, **9**, 658–661 (Association for the Advancement of Artificial Intelligence, 2015).
64. Barberá, P. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Polit. Anal.* **23**, 76–91 (2015).
65. Martin-Gutierrez, S., Losada, J. C. & Benito, R. M. Multipolar social systems: measuring polarization beyond dichotomous contexts. *Chaos Solitons Fractals* **169**, 113244 (2023).
66. Baumann, F., Lorenz-Spreen, P., Sokolov, I. M. & Starnini, M. Emergence of polarized ideological opinions in multidimensional topic spaces. *Phys. Rev. X* **11**, 011012 (2021).
67. Peralta, A. F., Ramaciotti, P., Kertész, J. & Iñiguez, G. Multidimensional political polarization in online social networks. *Phys. Rev. Res.* **6**, 013170 (2024).
68. Young, D. J. & de Wit, L. H. Affective polarization within parties. *Political Psychology*, 1–21 (2024).
69. Balliet, D., Tybur, J. M., Wu, J., Antonellis, C. & Van Lange, P. A. Political ideology, trust, and cooperation: in-group favoritism among republicans and democrats during a US national election. *J. Confl. Resolut.* **62**, 797–818 (2018).
70. Wojcieszak, M., Bimber, B., Feldman, L. & Stroud, N. J. Partisan news and political participation: exploring mediated relationships. *Political Commun.* **33**, 241–260 (2016).
71. Mosleh, M. & Rand, D. G. Measuring exposure to misinformation from political elites on twitter. *Nat. Commun.* **13**, 7144 (2022).
72. Lasser, J. et al. Social media sharing of low-quality news sources by political elites. *PNAS Nexus* **1**, pgac186 (2022).
73. Nguyen, C. T. Echo chambers and epistemic bubbles. *Episteme* **17**, 141–161 (2020).
74. Bruns, A. *Are Filter Bubbles Real?* (John Wiley & Sons, 2019).
75. Ross Arguedas, A., Robertson, C., Fletcher, R. & Nielsen, R. Echo Chambers, Filter Bubbles, And Polarisation: A Literature Review. Technical Report. (Oxford University, 2022).
76. Levendusky, M. S. & Stecula, D. A. *We Need to Talk: How Cross-Party Dialogue Reduces Affective Polarization* (Cambridge University Press, 2021).
77. Romano, A., Sutter, M., Liu, J. H., Yamagishi, T. & Balliet, D. National parochialism is ubiquitous across 42 nations around the world. *Nat. Commun.* **12**, 4456 (2021).
78. Dorrough, A. R. & Glöckner, A. Multinational investigation of cross-societal cooperation. *Proc. Natl Acad. Sci. USA* **113**, 10836–10841 (2016).
79. Baqir, A., Galeazzi, A., Drocco, A. & Zollo, F. Social media polarization reflects shifting political alliances in Pakistan. *arXiv preprint*, 2309.08075 (2023).
80. van Vliet, L. The Twitter Parliamentarian Database. https://figshare.com/articles/dataset/The_Twitter_Parliamentarian_Database/10120685 (2020).
81. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS ONE* **9**, e98679 (2014).
82. Hanu, L. & Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify (2020).
83. Newsguard Technologies Inc. Newsguard rating process criteria, accessed 02 May 2023. https://www.newsguardtech.com/ratings/rating-process-criteria/.
84. Lin, H. et al. High level of correspondence across different news domain quality rating sets. *PNAS Nexus* **2**, pgad286 (2023).
85. Pfeffer, J. Just another day on Twitter: a complete 24 hours of Twitter data. GESIS—Leibniz-Institute for the Social Sciences. Data File Version 1.0.0, https://doi.org/10.7802/2516 (2023).

Acknowledgements

M.F., F.Z., W.Q., and A.B. acknowledge the 100683EPID Project "Global Health Security Academic Research Coalition" SCH-00001-3391. M.F. thanks Maddalena Torricelli for help contextualizing the distribution of Italian politicians.

Author contributions

M.F.: Conceptualization, methodology, analysis, writing—original draft, figures. F.Z., W.Q., J.P.: Writing—editing, data acquisition. A.B.: Conceptualization, writing—editing, supervision.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-53868-0.

Correspondence and requests for materials should be addressed to Max Falkenberg or Andrea Baronchelli.

Peer review information *Nature Communications* thanks Almog Simchon and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.