PNAS Nexus, 2021, pp. 1–12

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

Unpacking Polarization: Antagonism and Alignment in Signed Networks of Online Interaction

Emma Fraxanet,[a,*] Max Pellert,[b] Simon Schweighofer,[c] Vicenç Gómez[a] and David Garcia[d,e]

[a]Department of Information and Communication Technologies, Universitat Pompeu Fabra, Tànger 122-140, 08018, Barcelona, Spain, [b]Chair for Data Science in the Economic and Social Sciences, University of Mannheim, L15 1–6, 68161, Mannheim, Germany, [c]Department of Media & Communication, Xi'an Jiaotong-Liverpool University, Suzhou Industrial Park, 215123, Suzhou, P.R.China, [d]Complexity Science Hub, Josefstäfter Strasse 39, 1080, Vienna, Austria and [d]Department of Politics and Public Administration, University of Konstanz, Universitätstrasse 10, 78464, Konstanz, Germany

[*]To whom correspondence should be addressed: email:emmafraxanet@gmail.com

FOR PUBLISHER ONLY Received on Date Month Year; accepted on Date Month Year

Abstract

Political conflict is an essential element of democratic systems, but can also threaten their existence if it becomes too intense. This happens particularly when most political issues become aligned along the same major fault line, splitting society into two antagonistic camps. In the 20th century, major fault lines were formed by structural conflicts, like owners vs workers, center vs periphery, etc. But these classical cleavages have since lost their explanatory power. Instead of theorizing new cleavages, we present the FAULTANA (FAULT-line Alignment Network Analysis) pipeline, a computational method to uncover major fault lines in data of signed online interactions. Our method makes it possible to quantify the degree of antagonism prevalent in different online debates, as well as how aligned each debate is to the major fault line. This makes it possible to identify the wedge issues driving polarization, characterized by both intense antagonism and alignment. We apply our approach to large-scale data sets of Birdwatch, a US-based Twitter fact-checking community and the discussion forums of DerStandard, an Austrian online newspaper. We find that both online communities are divided into two large groups and that their separation follows political identities and topics. In addition, for DerStandard, we pinpoint issues that reinforce societal fault lines and thus drive polarization. We also identify issues that trigger online conflict without strictly aligning with those dividing lines (e.g. COVID-19). Our methods allow us to construct a time-resolved picture of affective polarization that shows the separate contributions of cohesiveness and divisiveness to the dynamics of alignment during contentious elections and events.

Key words: Polarization, Signed Networks, Political Cleavages, Network Analysis, Social Media

Significance statement
Online media can fuel and amplify societal division, particularly in relation to (perceived) political identities. While a number of extremism and polarization mechanisms have been uncovered, effective mitigation interventions are yet to be determined. Strategies that naively create common spaces between factions can easily generate backfire effects that, counter-intuitively, worsen the situation. In our work, we gather existing methods previously mostly used in small-scale elite political systems and generate a pipeline applied to large-scale online environments, in an effort to leverage platforms' features such as fine-grained temporal resolution and diversity of topics to study *popular* polarization.

Introduction

It is nowadays difficult to watch a news broadcast, listen to a campaign speech, or read a political commentary without coming across the term *polarization*. It seems that, when political commentators need a catchy, one-word description of the current state of political affairs, they habitually default to *polarized*. But this inflationary usage of the concept of political polarization lumps together very different forms of political conflict. In a world where even apparently apolitical questions of lifestyle and taste have become associated with ideological positions [1], it may seem like every political conflict is being fought along the lines of left versus right, neatly splitting the political spectrum into two opposed factions. But neither in theory nor in practice is this the only way in which political antagonism can manifest in democratic societies.

The conflation of concepts when talking about polarization also explains the seemingly ambivalent role of political antagonism in democratic societies: On the one hand, polarization is usually conceptualized as detrimental to political stability and efficient governance. On the other hand, conflict and competition are recognized as essential parts of a functioning political system. This apparent contradiction is easily resolved by stipulating that political antagonism is not automatically detrimental to the stability of the system, as long as it is not exclusively located along the same dividing line, or *cleavage*. If political antagonism is located along multiple *cross-cutting cleavages* [2, 3], it can actually increase systemic cohesion by putting political actors into ever-changing configurations of alliances. In such a system, the opponents of yesterday may become the allies of tomorrow (and vice versa), which creates an incentive to maintain a minimum of civility [4]. In contrast, if conflicts are predominantly organized along a single cleavage, political actors will always find themselves alongside, and across from, the same group of people. It is easy to see why in such a system civility tends to be replaced by partisan hostility and political sectarianism [5].

The analysis of cleavage structure has been a central concern for political scientists (especially in Europe) since the seminal work of Lipset and Rokkan in 1967 [6]. They theorized that party systems in Western democracies are the results of four basic societal conflicts: center vs. periphery, state vs. church, owner vs. worker, and land vs. industry, which are present to differing degrees in different societies. The four cleavages initially introduced by Lipset and Rokkan in the 1960s have since lost a large degree of their explanatory power [7]. New cleavages have been proposed by various authors, determined, for example, by conflicts around globalization [8], migration [9], or European integration [10]. However, it has been criticized that, similar to 'polarization', the term 'cleavage' has been overexpanded, and thus lost most of its meaning, serving now merely as a redescription of differences in political attitudes among the electorate [11, 12].

In this study, we aim to harness the extensive and longitudinal features of online environments to discover cleavages based on high-resolution and contextualized data as a supplementary approach to theorizing specific cleavages *ab initio*. Moreover, once known the configuration of these dividing lines, we identify and analyze two distinct factors of *popular* political polarization: First, the degree of *Antagonism* in a community, a metric reflecting the prevalence of negativity in the interactions that are triggered by a controversial issue. And second, the degree of *Alignment* of a community around an issue, reflecting how much the issue 'fits', and thereby reinforces, these main dividing lines in a community. Our intuition is that political polarization can then be defined as the product of Antagonism and Alignment, both of which have to be present for a system to fission into radically opposed factions.

Our detection of cleavages, as well as our quantification of Antagonism and Alignment is based on the identification of optimal divisions separating factions in networks of signed relations. In such signed networks, each node represents an individual and their relations are represented by positive or negative edges.

In social media, positive interactions are captured by liking, praising, forming friendships, or establishing trust, while negative interactions are captured by disliking, toxic behavior, hostility, or distrust. By considering explicitly negative interactions within social media, we gain a deeper understanding of community structures and relations than by only analyzing positive interactions. For example, relying only on positive interaction data creates biases that lead to an overestimation of online fragmentation and distorted pictures of the polarization of a community [13, 14]. This is particularly important when assessing the degree of political polarization in social media use, which might have been overstated due to missing information on negative interactions [15].

There is a considerable set of literature considering the detection of communities in signed networks. While most methods utilize the concept of balance theory [16, 17], which postulates that positive interactions happen with a higher likelihood between individuals belonging to the same group (e.g. political faction), whereas negative interactions happen predominantly between opposed factions, we find different approaches to it. On one hand, there are methods aimed at detecting dense communities informed by balance [18, 19, 20, 21]. These could be strictly categorized as community detection, as the groups found are cohesive and can easily be numerous. On the other hand, we have what we could call faction detection approaches, in which the aim is to solve a MINCUT problem where you want to cut in a way that you minimize the number of edges that violate the partition model (known as frustration) [22, 23, 24, 25, 26]. These methods tend to offer solutions of binary or few factions.

Building on the latter [27], we designed FAULTANA, our proposed pipeline for detecting the fault lines of an online community and assessing Alignment and Antagonism in its interactions. Our framework can track changes in Alignment over time and compare how group structure manifests across issues in society. Furthermore, we analyze the two independent mechanisms that contribute to Alignment, namely *Cohesiveness* and *Divisiveness* [23], which account for in-group agreement versus out-group disagreement. At present, out-group disaffection is the most relevant variable in the steep increase of political sectarianism, especially in the US [5]. Hence, a proper consideration of negative interactions and relations is crucial to the analysis of polarization within online systems.

We apply this framework to two unique datasets that contain explicit signed interactions between users extracted from two different online platforms: Birdwatch, the US pilot stage of a crowd-sourced fact-checking Twitter system; and DerStandard, an Austrian online newspaper with discussions on news pieces.

The signed network data of Birdwatch and DerStandard offer a unique opportunity to directly measure positive and negative relationships in large-scale systems, as previous research struggled to infer negative relationship information from unsigned data [28, 29, 30]. This difficulty is particularly pronounced in online social systems, where distinguishing between users not interacting due to animosity versus chance becomes infeasible [13]. Even the inference of positive interactions from endorsing actions, such as retweets, has been called into question [31]. However, there still are examples of platforms with signed interaction features that have particular functions away from general discussion (e.g. Epinions [32], Slashdot [33] or Wikipedia RfA [34]), as well as datasets obtained through the inference of implicit signed interactions from text data such as Reddit [35, 36], or co-edit networks of Wikipedia [37, 38].

Similarly to our approach but aimed at the context of political elites, there has been plenty of work on faction detection in systems of party politics and elite polarization, such as an analysis on the House of Commons regarding Brexit [39], government formation in parliamentary democracies [40], the US Congress [41, 27] or the US House of Representatives [22]. A key factor for the development of this type of work is the existence of signed graphs of political elite interactions, which are usually inferred from co-voting or co-sponsorship data. These systems are small-scale, have a high density of interactions and usually allow for manual labeling of node attributes (e.g. party affiliation). Other work has covered the study of international relations, which have similar features [42, 43, 44, 45].

Alternatively, our two datasets provide information on extensive general popular discussions with a strong presence of political content and explicitly signed interactions in the form of positive and negative user-author ratings based on spontaneous behavior. Both datasets also have fine-grained temporal information and contextualization features encoded in news tags in the case of DerStandard and text in both datasets.

Our DerStandard dataset is novel and comprises eight years of signed interaction information between regular users of efficiently moderated news discussions. This dataset fills a valuable gap that has been present in the study of signed networks.

On the other hand, Birdwatch has been studied in previous research [46, 47, 48, 49, 50, 51, 52, 53, 54, 55] and has been found to be a strongly polarized platform: from general user behavior in the platform [46] to the interplay of political biases of the users and their assessments [47], as well as the study of its network structure [48, 49]. This makes it an interesting case study for our pipeline since it allows us to unfold the particularities of this polarization. Moreover, there is still ongoing doubt about the effectiveness of crowdsourced approaches to fact-checking, which is reflected in other research that has contrasted crowdsourced content in Birdwatch with expert fact-checking, noting variations in content selection, resource use, and efficiency [50]; explored believability and harmfulness of Birdwatch posts [51] or their diffusion in retweet cascades [52].

Data description and preparation

We use these key features of our two data sources: (a) DerStandard: positive (+) and negative (-) ratings on postings in the forum below articles on the online newspaper page. (b) Birdwatch: agreement and disagreement between raters and their notes, which we treat as positive and negative interactions. In both cases we also have temporal information (timestamp of postings or note).

We differentiate between: (i) *Interactions*: directed pairwise interactions based on the reaction of a user (rater) to the content posted by another user (author), with the timestamp corresponding to the posting of that piece of content, and (ii) *Edges*: undirected and signed relations between users of the platform, based on aggregated interactions exchanged between them through their postings or notes.

Network Creation: From interactions to edges

Both datasets contain pairwise interactions between users. Considering a dataset of n users, we model each relation between user i and user j from such interactions as a random variable that follows a Bernoulli distribution with parameter p_{ij}. We follow a Bayesian model using a beta prior for estimating p_{ij} with parameters α_0, β_0. After observing all the interactions between i and j in the dataset, the posterior probability also follows a beta distribution, in this case parametrized by $\alpha_0 + \text{pos}, \beta_0 + \text{neg}$, where pos and neg correspond to the number of positive and negative interactions respectively.

From these posterior probabilities, we build an undirected signed network $G = (V, E, \sigma)$, where V is a set of n nodes, E is a set of m edges, and σ_{ij} is the edge sign. Edges are only defined for pairs of users who have a certain bias towards 0 or 1, i.e., $\mathbb{E}[p_{ij}] > 0.6$ or $\mathbb{E}[p_{ij}] < 0.4$, and very low uncertainty, i.e., $\text{Var}[p_{ij}] < 10^{-4}$, where $\mathbb{E}[p_{ij}] = \frac{\alpha}{\alpha+\beta}$ and $\text{Var}(p_{ij}) = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$. For defined edges, we set their sign according to $\sigma_{ij} = \text{sign}\left(\mathbb{E}[p_{ij}] - \frac{1}{2}\right)$, i.e, two users have a positive (negative) edge if their expected posterior is above 0.6 (below 0.4) with high certainty.

Birdwatch

Launched in January 2021, the platform aimed at fighting Twitter misinformation via crowd-sourced fact-checking by selected volunteer users, or *birdwatchers*. These users assessed tweet trustworthiness with evaluative notes, including sources and arguments. The platform served as a small scale trial for the current known *Community Notes*. Previous work analysis has shown high political alignment and polarization among users [46, 47, 48, 49] and a tendency to scrutinize content from counter-partisans while following a partisan cheerleading behavior in ratings [47].

Twitter regularly published updated and publicly available datasets containing metadata of notes (text, tweet ID, note timestamp, classification, note ratings) and anonymized users data. We retrieved all data covering the time span between the start of Birdwatch in January 2021 and August 2022. Moreover, we re-hydrated the the content and metadata of the targeted tweets with the academic access to the Twitter API and computed an ideology score for the corresponding tweet author with Bayesian Ideal Point Estimation [15] implemented by the package *tweetscores*.

During the time span covered by our data, the platform changed their rating procedure from a simple `agree` versus `disagree` (Jan 2021 - Jun 2021) to `helpful`, `somewhat helpful` and `not helpful` for the remaining months in our data set (Jul 2021 - Dec 2022). Moreover, the platform launched a new algorithm to compute note statuses in February 2022, which searched for agreement across different viewpoints [54]. Since these are substantial platform changes, we split the dataset into two parts accordingly: BW1, and BW2, and center our study mostly on BW1, leaving BW2 as comparison only since it includes a series of platform changes. BW1 includes $\sim 32k$ pairwise interactions between $2,676$ users, while BW2 is a larger dataset comprising $\sim 235k$ pairwise interactions amongst $10,662$ users.

On this platform, positive and negative interactions are present in similar proportions (see SI Appendix, Table S1). Both interactions in Birdwatch, agreement and disagreement, can be considered to be equally meaningful because they require an argumentation. Consequently, we use a uniform prior for the beta distribution that characterizes the user relations on Birdwatch.

DerStandard

The web page of the Austrian newspaper has a long tradition (dating back to the 1990's) of offering users discussion forums. Compared to other platforms with similar features, DerStandard uniquely provides information on which users rated a posting in addition to the sign of the rating (see SI Appendix, Fig. S1 for an example of the interface). A recent study shows that users that are active on DerStandard tend to be more often male, younger, more highly educated, and more often from Vienna or Upper Austria than respondents of a representative survey in Austria [56].

With permission from DerStandard, we automatically retrieved all publicly available postings and user ratings in the discussion forums below each news piece on DerStandard between Jan 2014 and Dec 2021. In addition to postings and ratings, we also retrieved tags that classify news pieces into topics according to the platform (e.g. sports, refugees in Austria, Op-Ed columns, etc).

To analyze a stable user sample from DerStandard and avoid results originated from a large influx or outflux of users, we consider only users that rated at least once yearly in our observation period (begin of 2014 - end of 2021), thus removing accounts that have spurious activity levels. This allows us to identify roughly $14.8k$ users that we track over 8 years, comprising a total of $\sim 76M$ pairwise interactions. Our observation period spans major events, including the European refugee crisis (2015-16), the Austria government coalition dissolution due to corruption scandals (2019), and the COVID-19 pandemic (2020-21).

On DerStandard, negative interactions are underrepresented (SI Appendix, Table S1) and can carry a stronger signal than positive interactions. To account for that, we use a prior distribution that slightly favors negative interactions, especially when the volume of interactions is low, i.e., a beta distribution with $\alpha = 1$ and $\beta = 2$. The resulting network contains a similar number of negative and positive edges.

Partitioning methods for signed networks

Our approach is based on finding the main division lines in a community and the posterior analysis of the reinforcement or challenge of those divisions. To find a robust partition, we build on previous work.

Main optimization problem

Following [57] notation, given a signed graph $G = (V, E, \sigma)$, and a partition $P = \{X, V \setminus X\}$, the frustration count will be the sum of the frustration state of all edges, $f_G(P) = \sum_{(i,j) \in E} f_{ij}$, where f_{ij} equals 1 for frustrated edges and 0 otherwise. Frustrated edges correspond to the edges that violate the assumptions of the optimal partition model, i.e. negative edges between members of the same group or positive edges between members of different groups. The problem thus is stated as finding the optimal partition P^* with the minimum number of frustrated edges $L^*_G = \min_P f_G(P)$.

Computational methods

The computation of L^*_G is known to be NP-hard [58]. For small scale networks, however, exact computation of the frustration index is feasible using the binary linear programming formulation [58]. Several approximate methods have been proposed that are applicable to large scale networks. For example, Doreian and Mvar apply blockmodeling [59], in which they optimize the criterion function $P(X) = E_{f,p} + E_{f,n}$ via a relocation algorithm, with $E_{f,p}$ defined as the frustrated positive edges and $E_{f,n}$ the frustrated negative edges. In practice, this method, combined with simulated annealing as in the *Signnet* implementation [60], provides approximate values of L^*_G that correspond to robust partitions. Moreover, given that it involves a stochastic algorithm, we execute it 200 times and select the partition yielding the minimum L^*_G value. Further details regarding this approach can be found in the SI Appendix, section 2A. Any approximated value for L^*_G will necessarily be equal or higher than its exact value, given that there is no sub-optimal partition that can provide a smaller number of frustrated edges, thus the best approximated value will be an upper bound.

Generalization to more than two groups

All the previous definitions and methods are generalizable to $k > 2$ partitions [61, 22]. In that case, each value of k provides an optimal solution $L^*_G(k)$, and a reasonable selection is to keep the value of k which yields the minimum L^*_G. In [59], it is shown that L^*_G follows a concave curve with a unique minimum value of k, which we refer to as k^*. For all the case studies presented in this paper we find $k^* = 2$, ergo two differentiated factions. See SI Appendix, section 2B for details in the multi-partition selection for our data.

Previously defined useful metrics

Given a partition, a metric that can be useful in determining the incidence of the general division on the community, or the degree up to which the network can be easily separated into groups, is the "Normalized Frustration Index" [57]: $1 - \frac{L^*_G}{m/2}$. The normalizing factor $m/2$, where m is the volume of edges in the network, accounts for different network sizes and is a soft upper bound on the number of frustrated edges. Note that the index decreases the more frustration there is (i.e. the more blended the groups are). Therefore, when working with an approximated L^*_G, the index value is a lower bound.

Interestingly, we can understand this level of grouping as a structural measure related to polarization under the assumption that there is a small number of groups which are of similar sizes. Otherwise, it would either be a complete fragmentation into small groups or a case of a majority versus minorities. We discuss further details and show an example of a case where this assumption is not fulfilled in the SI Appendix, Section S4A, Figures S10 and S11. For our data, we verify this assumption in the Results section.

It is important to clarify that even though we draw from previous research that aimed at defining partial balance

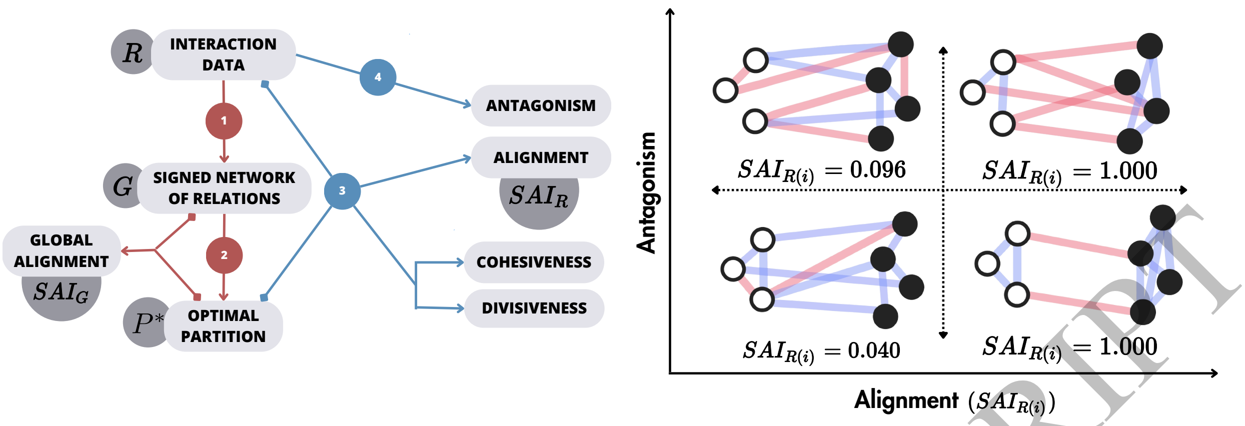

Fig. 1. (Left) Schema of the FAULTANA pipeline: our analysis framework for Antagonism, Alignment, Cohesiveness, and Divisiveness. Grey boxes indicate data structures and variables implicated in the pipeline. Step 1 creates the relation network based on an aggregation of interactions through time. Step 2 applies the optimization algorithm, either exact or approximated, to obtain the optimal number of groups and optimal partition. From these two steps we can retrieve a global Alignment metric SAI_G. Then, by selecting subsets of the interaction data and with the optimal partition information, steps 3 and 4 compute the four metrics of interest: Antagonism, Alignment, Cohesiveness (normalized) and Divisiveness (normalized).
(Right) Illustration of Antagonism and Alignment in signed network examples. The four networks have been constructed with the same edge density, number of nodes, and partitioning of nodes. Negative edges are red and positive edges are blue. The two upper networks have a higher proportion of negative edges, and thus higher Antagonism than the ones on the lower quadrants. Computed $SAI_{R(i)}$ values are provided to illustrate that the two right quadrants exhibit a higher level of Alignment, which is due to the lower amount of frustrated edges. Only the right upper quadrant corresponds to a strict definition of polarization in terms of both Antagonism and Alignment.

metrics, our aim diverges from assessing balance in these networks. In that case, the use of frustration could lead to incomplete descriptions of the interplay between polarization and balance in certain scenarios [45] (See SI Appendix, Section S4A). Instead, our focus lies on detecting a division between two or more factions based on frustration and studying the alignment of users to these factions.

Similar to earlier work [23], we can also analyze the two mechanisms that are involved in the coherence of users' links to the partition: alignment with one's own group (Cohesiveness) and alignment against the opposing group (Divisiveness). Cohesiveness (Divisiveness) is defined by the proportion of internal (external) edges that are positive (negative). Given our optimal partition P^*, internal edges are defined by $E_p^i = \{(i,j) \in E | i,j \in X \text{ or } i,j \notin X\}$ and external edges are defined by $E_p^e = \{(i,j) \in E | i \in X, j \notin X \text{ or } j \in X, i \notin X\}$.

The FAULTANA Pipeline

After constructing the signed relation network of each platform, we obtain their optimal partitions using the methods described above. Once the belonging of users to each group is fixed, we can assess the status of the platform globally (network of relations) or describe the status of directed sub-sets of the data (network of interactions). These allow us to find four metrics of interest: *Alignment*, *Antagonism*, *Cohesiveness* and *Divisiveness*. We designate this set of steps as the FAULTANA pipeline, which stands for FAULT-line Alignment Network Analysis (See Figure 1).

Re-normalization and global metrics

In order to be able to compare across subsets of our data, we have to re-think some of the characteristics of the previously designated metrics like the "Normalized Frustration Index", Cohesiveness or Divisiveness.

In the case of the "Normalized Frustration Index", we re-normalize it by comparing the empirical estimate of L^* versus its mean value in repeated measurements of a null model. The null model based on graph G randomly re-distributes sign attributes while keeping the overall structure of the network and the partition fixed ($\widetilde{G}$). The value of L in the null model simulations, $L_{\widetilde{G}}$, is consistently higher than the frustrated edges in our datasets, proving to be a tighter bound than only considering the number of edges with the term $m/2$. Thus, we define the *Global Signed Alignment Index*, which we also simply call *Global Alignment*, as:

$$SAI_G = \langle 1 - \frac{L^*_G}{L_{\widetilde{G}}} \rangle \tag{1}$$

We obtain 95% confidence intervals for SAI_G from the distribution resulting of repeated instances of the null model.

On the other hand, the measures of Cohesiveness and Divisiveness defined above cannot be compared between systems with different ratios of negative versus positive interactions. For example, a system A with a higher ratio of negative interactions than system B will have by construction a higher Divisiveness even if it is not more strongly divided along the fault line than system B. This can be observed in simulations of our null model, which show that the expected value of Divisiveness and Cohesiveness is perfectly correlated with the fraction of negative interactions (Fig.S4). To solve this, we design new metrics of Normalized Divisiveness and Normalized Cohesiveness by subtracting from the original metrics (i.e. not normalized) the mean values obtained in the null model simulations. For brevity, we will use the original metric names to refer to their normalized versions throughout the remainder of the manuscript. We assess the uncertainty of our measurements of Divisiveness and Cohesiveness through bootstrapping. For each measurement, we create 10,000 bootstrap samples of the network with replacement and of the same size as the original. On each bootstrap sample, we calculate Divisiveness and Cohesiveness and we use the resulting values to calculate the bootstrapping confidence interval around our original measurement.

Formalization of Alignment

Our normalization approach allows us to obtain a meaningful SAI for sub-sets of the interaction data. To do so, we maintain the optimal partition obtained from the network of relations (i.e. we fix the belonging of each user to a group that is defined by the long-term relation between users), and we proceed to assess how aligned the interactions within that subset of the data are to this partition. We refer to this metric as *Alignment*, or SAI_R. Since it follows the same laws of frustration (e.g. negative interactions within a group are frustrated interactions, and so on), we just have to re-define the SAI_G in the following way:

$$SAI_R = \langle 1 - \frac{L_R}{L_{\widetilde{R}}} \rangle \tag{2}$$

where R accounts for the network of directed interactions within a set or subset of the data, denoted by $R(t)$ in case of a temporal subset, or (i) for a selection based on issue or topic. L_R is then the number of frustrated interactions in that network given the existing assignment of nodes to groups. As in the case of SAI_G, $\widetilde{R}$ denotes an instance of the null model applied on R, by reshuffling the sign configuration while keeping the network structures and groups.

Formalization of Antagonism

Additionally, we formally describe *Antagonism* as the proportion of negative interactions in R, which is a simple indication of the prevalence of conflict or general disagreement. This measure is then not related to the network structure, like Alignment, but it indicates a property of the user-content interaction in terms of the overall presence of disagreement in comparison to agreement.

Formalization of Cohesiveness and Divisiveness (local)

Besides computing Cohesiveness and Divisiveness for the network of relations, providing a general overview of each platform, we can also analyse these metrics for subsets of interactions associated with topics or time periods, which can show how Cohesiveness and Divisiveness contribute in Alignment changes. Furthermore, given the directionality of ratings in the network of interactions, we can examine group asymmetries by calculating the separate contribution from each group to these metrics (e.g. how much of the division between two groups is driven by one of them).

Conceptualization of Alignment and Antagonism

The metric of *Alignment* captures how interactions follow the division of the network into opposed groups, while our metric of *Antagonism* captures the overall tendency towards negative interaction in the network regardless of groups. By considering both these measures, we can provide a more comprehensive picture of polarization than when these two concepts are not explicitly distinguished. Figure 1 shows how these two metrics capture various polarization scenarios given a partition of the network into groups and the positive and negative interactions in the system. A network with low Alignment and low Antagonism has few negative interactions and no strong division into groups, corresponding to a situation with the weakest polarization. The lower right part of the space, where Alignment is high but Antagonism is low, corresponds to an *echo chamber* case in which most interactions are positive but happen between like-minded individuals and not across groups. The upper left cases are networks with high Antagonism but low Alignment, capturing scenarios where disagreement exists but not necessarily following the division of the network into groups. This can happen when everyone is against everyone or where other divisions exist but do not follow the general ideological separation of the network into groups. And finally, the upper right part of the space corresponds to cases where polarization is high, as both Antagonism and Alignment are high. In this high-polarization case, there is a strong cleavage between groups such that positive interactions are confined within groups while frequent negative interactions happen mostly across groups.

Results

Approximating Alignment in Birdwatch

In this section, we evaluate our methodology and its performance based on the results obtained from the two Birdwatch datasets. We use Birdwatch for two key factors. Firstly, as described in Section Computational methods, we can run the exact method for small networks, while for large networks we have to run the approximate algorithm due to the complexity of the problem. The size of BW1 allows us to run both the exact and approximate algorithms and compare the solutions to estimate the difference in signed networks of this kind. The results of both algorithms are very similar in BW1, with the approximated SAI_G being 84% of the SAI_G obtained with the exact method and an average partition overlap coefficient [62] of 0.89.

For both BW1 and BW2, we find that the optimal number of groups is $k^* = 2$, and the largest groups contain roughly twice the number of users of their smaller counterparts (see SI Appendix, Table S2 for more details). Figure 2 shows the signed network of relations obtained from BW1. Previous literature focused on Birdwatch suggests that the platform is characterized by two opposing factions, corresponding to Republican- and Democrat-leaning users, who attach notes to tweets following behaviors of *counter-partisan policing* and *inner-partisan cheer-leading* [47]. By building on the ideology score extracted from the tweets, we test whether the groups identified through our method reproduce this behavior, thereby evaluating the coherence of our approach with other metrics of political alignment.

When we retrieve the notes that users from each of these partitions have given to tweets, we find evidence of these policing-cheerleading patterns, as our largest group - which we denote as *inferred Democrats* - is strongly biased towards tagging Republican-leaning tweets as misleading. Contrarily, the smaller partition - *inferred Republicans* - consistently rates like-minded tweets as not misleading (see Figure 2).

Evolution of Alignment in Birdwatch

The lower part of Figure 2 shows the time series of $SAI_{R(t)}$ in BW1. The fluctuations in the measure over time indicate whether the level of Alignment among interactions increased or decreased during that particular period. The time series of

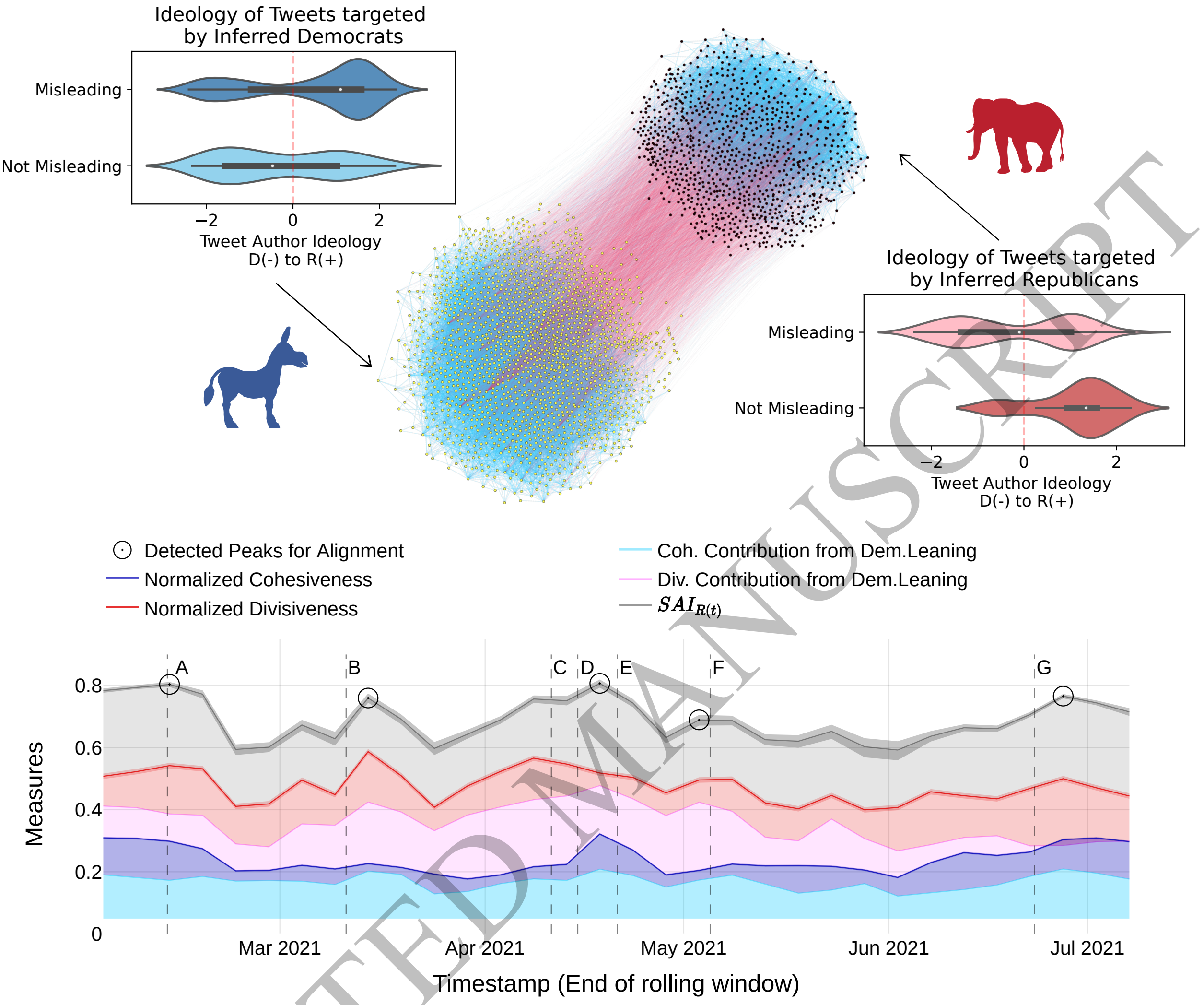

Fig. 2. (Upper figure) Signed network visualization of Birdwatch. Network of signed relationships for the BW1 dataset, comprising a total of 2,676 users and around 25,562 edges, negative colored red and positive blue. Node color corresponds to their group membership as identified by the exact method. Nodes belonging to the largest (smallest) group are depicted in yellow (black). Negative edges tend to connect different groups, while positive edges predominantly connect nodes within groups, demonstrating a considerable degree of balance.

Insets: Inferred ideology of the targeted tweet's author separated by which group targeted the tweet and the nature of the note. We can only retrieve a score for tweet authors that have connections to political actors (∼ 60% of the users that posted tweets targeted in Birdwatch). The larger group gives misleading notes with more probability to tweets authored by Republican users, i.e. counter-partisan policing, with a slightly higher tendency to give not misleading notes to tweets by Democrat users. Thus, we identify the larger group as Democrat-leaning. The smaller group is much more likely to give not misleading notes to tweets authored by Republican users, showing a pattern of cheer-leading within Republicans and thus being identified as Republican-leaning.

(Lower figure) Timeline of Alignment, Cohesiveness and Divisiveness in Birdwatch (BW1). The time series of each metric is calculated over a rolling window of ten days with increases of 5 days, with values allocated on the right of each window. The shaded area around Alignment time series shows 95% Confidence Intervals calculated against $10,000$ instances of the null model. Divisiveness is shown in red and Cohesiveness is shown in blue, with lighter areas showing the contribution of Democrat-leaning users to each metric and the remaining area above showing the contribution of Republican-leaning users. Bootstrapping intervals in Divisiveness and Cohesiveness are obtained for 10,000 bootstrap samples with replacement. The Alignment measure, $SAI_{R(t)}$, oscillates around a mean value of 0.65. Divisiveness stays consistently above Cohesiveness, showing that negative interactions are the main driver of Alignment. Detected peaks in $SAI_{R(t)}$ are marked with circles and notable political events in the US are marked with vertical dashed lines for reference. For each peak, a summary text analysis of tweets in that period is shown in SI Appendix, Table S4, which can be further contextualized as increases in Cohesiveness, Divisiveness, or both.

normalized Cohesiveness and Divisiveness contextualize these movements, as they show whether peaks are due to higher cohesion within groups or higher division between groups, and what is the contribution of each group to these metrics. In these time series, Antagonism and Alignment have a low correlation, which emphasizes the need to consider them as two different measures (more details can be found in SI Appendix, Fig. S5).

We applied a peak detection algorithm and identified five local maxima of $SAI_{R(t)}$ that are marked in Figure 2. To understand the context of the tweets on the day of each peak, we generated wordshift diagrams [63] for each peak in comparison to the rest of the tweets. Details on the wordshift diagrams can be found in SI Appendix, Fig. S7, Fig. S8 and Fig. S9. Our analysis shows that peaks of Alignment happen around controversial topics in the US. For example, we see that the second peak, associated with events related to COVID-19 vaccination (B), is driven by an increase in Divisiveness, especially from Democrat-leaning users. Alternatively, the third detected peak, which is associated with events about police shootings (C,D,E) , has a stronger contribution of Cohesiveness, especially within the Republican-leaning users. The other three peaks (1st, 4th and 5th) are driven by a mix of Cohesiveness and Divisiveness. The keywords and events at those time periods point towards discussions regarding the US Government and its policies (G), Donald Trump and 2020 election results (F), and other relevant events such as the Capitol insurrection or the Texas Power Crisis (A). A list of keywords and identified relevant events can be found in SI Appendix, Table S4.

Results for DerStandard

Our approach to detecting groups in the DerStandard network shows that this network has an optimal k^* of two groups, as in Birdwatch. The size of these groups is slightly more similar, with the largest one comprising 62% of the nodes. Even though the DerStandard dataset spans a much longer period and contains more users than the Birdwatch datasets, the Global Alignment of the network is substantially high ($SAI_G = 0.3955$), showing that alignment can appear across different sizes and time scales. Divisiveness (0.2899) is still considerably higher than Cohesiveness (0.1409), also mirroring the results for Birdwatch. More details on these results can be found in SI Appendix, Table S3.

Given the classification of news in DerStandard, we can measure Alignment and Antagonism on the set of user ratings focused on specific topics, thus locating issues in the space of network structures shown in Fig. 1. The scatter plot for Alignment and Antagonism of DerStandard topics is shown in Figure 3, where the spread of values allows for all four combinations outlined by our approach. Alignment and Antagonism have a low correlation across topics ($r = -0.0016, p = 0.981, 95\%$ CI $[-0.134, 0.131]$), suggesting that these two concepts should not be conflated into a general dimension of polarization. By inspecting the topics falling into each quadrant of the plot, we find their distribution agrees with intuitive expectations. For example, topics with a high conflict potential such as migration, COVID-19 politics, gender politics, climate change, and elections are on the high range of Antagonism, whereas lifestyle, sports, and culture topics such as movies, family, travel, art market or international football are located in the low ranges of Antagonism. With regard to the dimension of Alignment, we find that conflicting topics such as national elections, abortion, military service, or climate change are more aligned than migration or COVID-19 politics. These last two were indeed issues that did not divide the Austrian population clearly into left and right. Note that these patterns cannot be explained by the number of ratings, posts, or articles on each topic, as shown more in detail in the SI Appendix, section 3B.

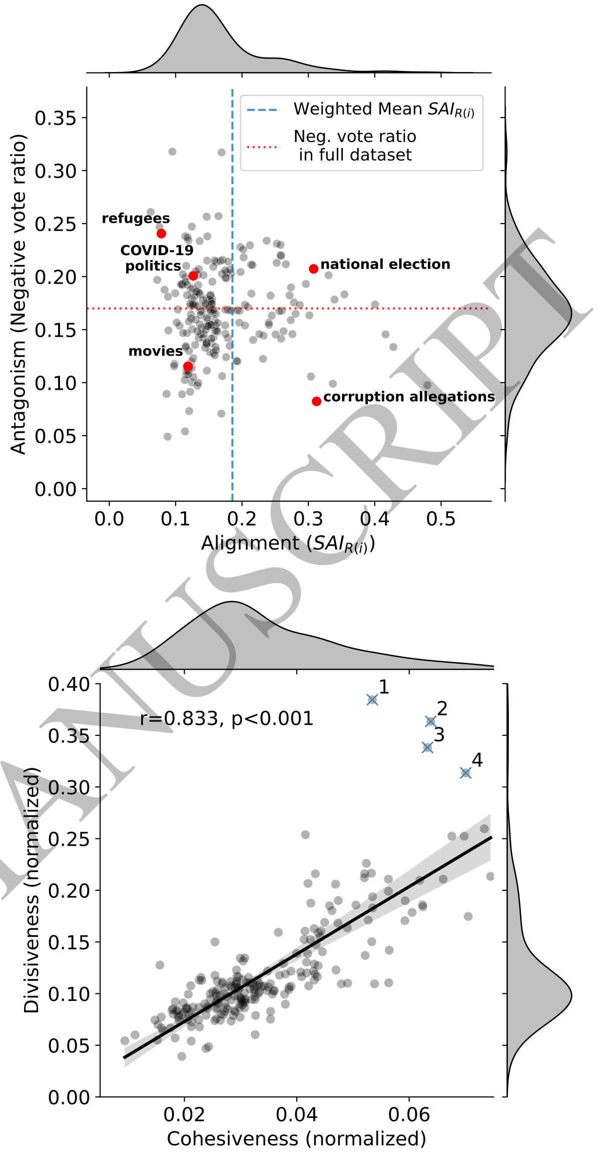

Fig. 3. Alignment versus Antagonism and Cohesiveness versus Divisiveness across DerStandard topics. The upper figure shows Antagonism and Alignment of the ratings of each news topic in DerStandard. Topics have been selected based on the topic/subtopic tags associated with the articles located above the postings (e.g., sports, climate change, etc.). Dashed lines show the mean values of each metric to identify the quadrants depicted in Figure 1. An interactive version of this figure can be found at https://emmafrax.github.io/scatter.html. The lower figure shows the scatterplot of Divisiveness versus Cohesiveness for DerStandard rating sub-sets based on topics. These two measures, which account for two different mechanisms that define Alignment, have a significant correlation across topics of 0.8. The highlighted outliers correspond to: (1) BVT (Austrian counterterrorism agency), (2) Abortion, (3) Scheuba (Austrian comedian) and (4) ÖVP (Political Party)

We highlight a few examples within each quadrant of Figure 3 to better illustrate how Alignment and Antagonism relate to each other. While *Refugees* and *COVID-19 politics* are identified as conflicting topics, resulting in higher levels of Antagonism, they do not align precisely with the primary division line. During the crucial years for those topics of 2015/16 and 2020/21, we have seen some unexpected

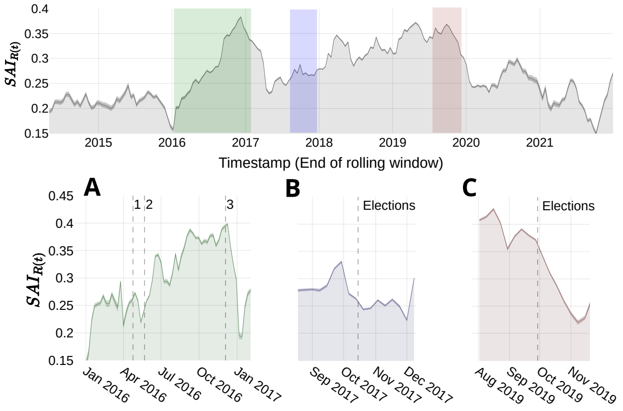

Fig. 4. Alignment timeline in DerStandard ratings sub-set of political topics, with detailed fluctuations in election periods. Upper timeline figure shows the Alignment measure obtained using a rolling window of 120 days of width and a step of 14 days. The features of the rolling window are selected so that the trends in Alignment through the eight years are visible, e.g. the change in trend at the start of 2016. In the lower figures we show more detailed changes of Alignment, with a rolling window of 30 days of width and a step of 7 days, around the three repetitions of the 2016 Presidential elections (A: 1,2 and 3) and the 2017 and 2019 Legislative Elections (B and C).

political alliances that do not follow from a classical left-right spectrum. These include common platforms between the anti-migration left and right-wing populists or the anti-statist right and anti-vaccine parts of rather left-wing Green parties. These agreements on certain issues between otherwise ideologically distant parties have historically been described by the term ”Querfront” (”cross-front”) [64]. Conversely, the tag *National elections* exhibits both Antagonism and Alignment, indicating a combination that favors polarization. This can be explained by federal elections in a representative democracy to lead to more discussion along traditional party lines. Additionally, *Corruption allegations* pertains to specific events involving some of the political parties in Austria. Although it demonstrates Alignment, these particular events did not generate substantial conflict within the platform. This could potentially be due to a limited number of defenders of those specific parties that have been covered much in the news in a corruption context (FPÖ and ÖVP, resulting from their joint government coalition), as DerStandard is historically considered a more left-liberal-leaning newspaper. As expected, a more offtopic tag such as *Movies* exhibits low levels of both Alignment and Antagonism.

While Antagonism and Alignment across topics are weakly correlated, Cohesiveness and Divisiveness are strongly correlated, as shown on the lower panel of Fig 3. This is expected, as the affective component of polarization captured by Alignment implies a correlation between out-group animosity and in-group support. Nevertheless, there are topics that deviate from the association between Cohesiveness and Divisiveness by having substantially higher Divisiveness: BVT (Institution), Abortion, Scheuba (Austrian comedian) and ÖVP (Political Party) (see lower panel of Fig 3), while this pattern is not mirrored for Cohesiveness. As with the time series of Alignment on Birdwatch, measuring Cohesiveness and Divisiveness is informative even though they both form part of the same phenomenon of Alignment.

The time series of Alignment in DerStandard reveals how cleavages become salient around politically-relevant events. Figure 4 shows the time series $SAI_{R(t)}$ for all DerStandard discussions in news on three topics: national elections, parties, and the federal president. This highlights political discussions from other, less-contentious topics as identified above. There is a clear change in the trend of Alignment at the beginning of 2016, showing steady growth up to the beginning of 2017. This falls into the time period of the so-called ”2015 European migrant crisis” [65] when migrants arrived in Europe in numbers that were unprecedented since World War Second. While migration started before 2016, the rise in Alignment starts right after the reporting of sexual assaults during New Year's Eve 2015-2016 celebrations in Cologne, Germany[66], which were widely covered in German-speaking media and debated over the following year.

Political events can also drive decreases in Alignment, especially if we consider that Austria has a multi-party system. After an election, the political climate changes toward building government coalitions with multiple parties, thus predicting lower Alignment as suggested by the case of online networks of Swiss politicians [67]. This can be observed in the time series of Alignment in DerStandard if we zoom in to recent elections. Panel A of Figure 4 shows the timeline of Alignment during 2016, where the increase in Alignment that year accelerates after the result of a presidential election was overturned by the Supreme Court of Justice. This controversial decision lead to a period of increased Alignment towards the repetition of the election, to then quickly reset to earlier levels of Alignment as soon as the repeated election took place and a candidate won by a large margin.

Panel B of Figure 4 shows a decrease in Alignment that happened shortly before the 2017 legislative elections, which was called early since there were clear favorite parties to form a coalition in pre-election polls. The effect of the legislative elections in 2019 (panel C) showed a sharp decrease in Alignment afterward, as the result was not as clearly expected as in 2017 and which led to a new government coalition with a party that was not involved in the previous government.

Conclusions

We have successfully factored online polarization into two dimensions: Antagonism, representing the level of hostility or disagreement in an online discussion, and Alignment, determined by the tendency of individuals to position themselves in a discussion according to their belonging to a group (or for that matter, their positioning across the ”other” group(s)). These two measures, although both contributing to polarization, have distinct characteristics and are weakly correlated across topics. We discovered that large-scale online political discussions exhibit an underlying polarized structure, which becomes more prominent when examining discussions centered around aligned topics. An essential takeaway is that online polarization is a dynamic, responsive phenomenon deeply influenced by contemporary political and societal events. It exhibits rapid responses, but through an examination spanning a sufficiently extended time frames, such as in the case of DerStandard, we can discern overarching trends alongside specific peaks.

Particularly, in terms of insights drawn from the study of Birdwatch, we found that changes in polarization can arise from different mechanisms (i.e. Cohesiveness or Divisiveness) within one or both of the groups. Additionally, the identification of Republicans and Democrats provides valuable insights into the status of each topic and positioning in relevant online discussions. Our findings on Birdwatch, as a platform dedicated to crowd-sourced fact-checking which has now been extended globally, can be beneficial to understanding the dynamics and effectiveness of using a wisdom-of-the-crowds approach to combat misinformation.

On the other hand, through our comprehensive analysis of DerStandard, we have uncovered that topics such as COVID-19 politics and Refugees, despite their contentious and relevant nature in online discussions, do not align with Austria's general Left-Right divide. This finding sheds light on the political divisions within Austria and serves as evidence that our methodology is capable of identifying cross-cutting cleavages, as these are topics with high antagonism but lower alignment. Furthermore, through an analysis of the temporal trends of Alignment pertaining to politically relevant topics, our findings demonstrate coherence with expected behaviors given the context of the respective time frames.

FAULTANA, our proposed pipeline, is agnostic in terms of political system (is applicable to multi-party as well as two-party), language, or issue dimensions, and can be extended to other use cases as long as positive and negative interaction information is available. It can also be tuned to platform-specific features, for example choosing the prior distribution for user relations. In the specific cases of DerStandard and Birdwatch, we were able to retrieve a division in the ideological spectrum (left .vs. right), but it is possible that other platforms' main divisions fall between other social, demographic or ideological positionings. Therefore, it allows us to study the main cleavage in a platform's community without the need of classifying users by their opinions *a priori*.

Our work is subject to several limitations, the main one being that we have to use approximated methods to find (near)-optimal partitions for large scale networks. However, even in that situation, we still capture significant values for our metrics and our approximated results are comparable to the exact results for the BW1 dataset, which brings us to the conclusion that we are still measuring what we aimed to, even if not at the highest accuracy possible. Moreover, even in the exact solution it is not possible to ensure a unique single optimal partition, since the method only ensures a unique solution for the minimum amount of frustrated edges, and several partitions can satisfy that requirement [57].

On the other hand, a relevant assumption is on the fixed belonging of users to a group defined by the optimal partition. We are assuming there is a global clustering to which users are aligned. This is not too far-fetched given the fact that there tends to exist issue alignment in society [68, 1] and we consider different numbers of clusters that lead to lower alignment. However, for long time scales or unprecedented global phenomena such as the COVID-19 pandemic, there could be relevant changes in these structures.

We leave for future work the inclusion of a tracing system that assesses the partition quality through time and updates it accordingly. With access to longitudinal data for large populations, our method could be very useful to automatically detect shifts in the main lines of division, which would provide more accurate pictures of polarization in terms of its components of alignment and antagonism, how it manifests across topics, and how it evolves over time .

Acknowledgments

We thank Samin Aref for insightful discussions on the methodology.

Ethics Statement

The data utilized in this study consists of publicly available records of social media data. Specifically, Birdwatch data (now known as Community Notes) is freely accessible at Community Notes. The re-hydration of these notes with tweet data was conducted using academic access granted by the former Twitter API. Data from DerStandard was extracted with explicit permission from the newspaper, who approved us to scrape content automatically. All data shared for reproducibility has been anonymized to protect the privacy of users in the data.

Supplementary Material

This article has an accompanying supplementary file.

Funding

D.G. and M.P. acknowledge funding from the Vienna Science and Technology Fund (WWTF) [10.47379/VRG16005]. E.F. acknowledges funding from the Maria de Maeztu Units of Excellence Programme CEX2021-001195-M, funded by MICIU/AEI /10.13039/501100011033.

Author contributions statement

D.G, V.G, E.F, M.P., S.S designed research; M.P. retrieved, prepared and contextualized one of the datasets; E.F. analyzed data and generated the manuscript; D.G., V.G. provided continuous supervision and original ideas; all authors contributed to writing the final version of the manuscript.

Previous presentation

These results were (partially) previously presented at IC2S2 2022 and NetSci 2023.

Preprints

A preprint of this article is published at
https://doi.org/10.48550/arXiv.2307.06571.

Data availability

The code and anonymized network data used in this article are available in the repository Unpacking Polarization, DOI 10.17605/OSF.IO/EY5CT.

References

1. Daniel DellaPosta, Yongren Shi, and Michael Macy. Why do liberals drink lattes? *American Journal of Sociology*, 120(5):1473–1511, 2015.
2. Stein Rokkan. Geography, religion, and social class: Crosscutting cleavages in norwegian politics. *Party systems and voter alignments*, 367:379–86, 1967.

3. Peter Michael Blau and Joseph E Schwartz. *Crosscutting Social Circles: Testing a Macrostructural Theory of Intergroup Relations.* Academic Press, 1984.
4. Lilliana Mason. A cross-cutting calm: How social sorting drives affective polarization. *Public Opinion Quarterly*, 80(S1):351–377, 2016.
5. Eli J Finkel, Christopher A Bail, Mina Cikara, Peter H Ditto, Shanto Iyengar, Samara Klar, Lilliana Mason, Mary C McGrath, Brendan Nyhan, David G Rand, et al. Political sectarianism in america. *Science*, 370(6516):533–536, 2020.
6. Seymour Martin Lipset, Seymour M Lipset, and Stein Rokkan. *Party systems and voter alignments: Cross-national perspectives*, volume 7. New York: Free Press, 1967.
7. Mark N Franklin. The decline of cleavage politics. *Electoral change: Responses to evolving social and attitudinal structures in Western countries*, pages 383–405, 1992.
8. Hanspeter Kriesi, Edgar Grande, Romain Lachat, Martin Dolezal, Simon Bornschier, and Timotheos Frey. *West European politics in the age of globalization.* Cambridge University Press, 2008.
9. Robert Ford and William Jennings. The changing cleavage politics of western europe. *Annual review of political science*, 23:295–314, 2020.
10. Liesbet Hooghe and Gary Marks. Cleavage theory meets europe's crises: Lipset, rokkan, and the transnational cleavage. *Journal of European public policy*, 25(1):109–135, 2018.
11. Stefano Bartolini and Peter Mair. *Identity, competition and electoral availability: the stabilisation of European electorates 1885-1985.* ECPR Press, 2007.
12. Andreas C Goldberg. The evolution of cleavage voting in four western countries: Structural, behavioural or political dealignment? *European Journal of Political Research*, 59(1):68–90, 2020.
13. Pedro Calais Guerra, Wagner Meira Jr, Claire Cardie, and Robert Kleinberg. A measure of polarization on social media networks based on community boundaries. In *Seventh international AAAI conference on weblogs and social media*, 2013.
14. Anna Keuchenius, Petter Törnberg, and Justus Uitermark. Why it is important to consider negative ties when studying polarized debates: A signed network analysis of a dutch cultural controversy on twitter. *PloS one*, 16(8):e0256696, 2021.
15. Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10):1531–1542, 2015.
16. Fritz Heider. *The Psychology of Interpersonal Relations.* John Wiley & Sons Inc, Hoboken, 1958.
17. Dorwin Cartwright and Frank Harary. Structural balance: a generalization of heider's theory. *Psychological review*, 63(5):277, 1956.
18. Shiwei Che, Wu Yang, and Wei Wang. A memetic algorithm for community detection in signed networks. *IEEE Access*, 8:123585–123602, 2020.
19. Yansen Su, Bangju Wang, Fan Cheng, Lei Zhang, Xingyi Zhang, and Linqiang Pan. An algorithm based on positive and negative links for community detection in signed networks. *Scientific reports*, 7(1):10874, 2017.
20. Pouya Esmailian and Mahdi Jalili. Community detection in signed networks: the role of negative ties in different scales. *Scientific reports*, 5(1):14339, 2015.
21. Yang Li, Bo Yang, Xuehua Zhao, Zhejian Yang, and Hechang Chen. Ssbm: A signed stochastic block model for multiple structure discovery in large-scale exploratory signed networks. *Knowledge-Based Systems*, 259:110068, 2023.
22. Samin Aref and Zachary P Neal. Identifying hidden coalitions in the us house of representatives by optimally partitioning signed networks based on generalized balance. *Scientific reports*, 11(1):1–9, 2021.
23. Samin Aref, Ly Dinh, Rezvaneh Rezapour, and Jana Diesner. Multilevel structural evaluation of signed directed social networks based on balance theory. *Scientific reports*, 10(1):1–12, 2020.
24. Giuseppe Facchetti, Giovanni Iacono, and Claudio Altafini. Computing global structural balance in large-scale signed social networks. *Proceedings of the National Academy of Sciences*, 108(52):20953–20958, 2011.
25. Vincent A Traag and Jeroen Bruggeman. Community detection in networks with positive and negative links. *Physical Review E*, 80(3):036115, 2009.
26. Mihai Cucuringu, Peter Davies, Aldo Glielmo, and Hemant Tyagi. Sponge: A generalized eigenproblem for clustering signed networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1088–1098. PMLR, 2019.
27. Samin Aref and Zachary Neal. Detecting coalitions by optimally partitioning signed networks of political collaboration. *Scientific reports*, 10(1):1506, 2020.
28. Georges Andres, Giona Casiraghi, Giacomo Vaccario, and Frank Schweitzer. Reconstructing signed relations from interaction data. *arXiv preprint arXiv:2209.03219*, 2022.
29. David Garcia and Dorian Tanase. Measuring cultural dynamics through the eurovision song contest. *Advances in Complex Systems*, 16(08):1350037, 2013.
30. Zachary Neal. The backbone of bipartite projections: Inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors. *Social Networks*, 39:84–97, 2014.
31. Zeynep Tufekci. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 505–514, 2014.
32. Ramanthan Guha, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Propagation of trust and distrust. In *Proceedings of the 13th international conference on World Wide Web*, pages 403–412, 2004.
33. Jérôme Kunegis, Andreas Lommatzsch, and Christian Bauckhage. The slashdot zoo: mining a social network with negative edges. In *Proceedings of the 18th international conference on World wide web*, pages 741–750, 2009.
34. Robert West, Hristo S Paskov, Jure Leskovec, and Christopher Potts. Exploiting social network structure for person-to-person sentiment analysis. *Transactions of the Association for Computational Linguistics*, 2:297–310, 2014.
35. John Pougué-Biyong, Valentina Semenova, Alexandre Matton, Rachel Han, Aerin Kim, Renaud Lambiotte, and Doyne Farmer. Debagreement: A comment-reply dataset for (dis) agreement detection in online debates. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

36. John Pougué-Biyong, Akshay Gupta, Aria Haghighi, and Ahmed El-Kishky. Learning stance embeddings from signed social graphs. *arXiv preprint arXiv:2201.11675*, 2022.
37. Silviu Maniu, Bogdan Cautis, and Talel Abdessalem. Building a signed network from interactions in Wikipedia. In *Databases and Social Networks on - DBSocial '11*, pages 19–24, Athens, Greece, 2011. ACM Press.
38. Jürgen Lerner and Alessandro Lomi. The free encyclopedia that anyone can dispute: An analysis of the micro-structural dynamics of positive and negative relations in the production of contentious wikipedia articles. *Social Networks*, 60:11–25, 2020.
39. Carla Intal and Taha Yasseri. Dissent and rebellion in the house of commons: A social network analysis of brexit-related divisions in the 57th parliament. *Applied Network Science*, 6(1):36, 2021.
40. Angela Fontan and Claudio Altafini. A signed network perspective on the government formation process in parliamentary democracies. *Scientific reports*, 11(1):5134, 2021.
41. Arthur Capozzi, Alfonso Semeraro, and Giancarlo Ruffo. Analyzing and visualizing polarization and balance with signed networks: the us congress case study. *Journal of Complex Networks*, 11(4):cnad027, 2023.
42. Patrick Doreian and Andrej Mrvar. Structural balance and signed international relations. *Journal of Social Structure*, 16(1):1–49, 2015.
43. Zeev Maoz, Lesley G Terris, Ranan D Kuperman, and Ilan Talmud. What is the enemy of my enemy? causes and consequences of imbalanced international relations, 1816–2001. *The Journal of Politics*, 69(1):100–115, 2007.
44. Fernando Diaz-Diaz, Paolo Bartesaghi, and Ernesto Estrada. Network theory meets history. local balance in global international relations. *arXiv preprint arXiv:2303.03774*, 2023.
45. Ernesto Estrada. Rethinking structural balance in signed social networks. *Discrete Applied Mathematics*, 268:70–90, 2019.
46. Nicolas Pröllochs. Community-based fact-checking on twitter's birdwatch platform. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 794–805, 2022.
47. Jennifer Allen, Cameron Martel, and David G Rand. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in twitter's birdwatch crowdsourced fact-checking program. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.
48. Taha Yasseri and Filippo Menczer. Can crowdsourcing rescue the social marketplace of ideas? *Communications of the ACM*, 66(9):42–45, 2023.
49. Ryan Champaigne. Birdwatch and the polarization of the crowds. 2022.
50. Mohammed Saeed, Nicolas Traub, Maelle Nicolas, Gianluca Demartini, and Paolo Papotti. Crowdsourced fact-checking at twitter: How does the crowd compare with experts? In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1736–1746, 2022.
51. Chiara Patricia Drolsbach and Nicolas Pröllochs. Believability and harmfulness shape the virality of misleading social media posts. In *Proceedings of the ACM Web Conference 2023*, pages 4172–4177, 2023.
52. Chiara Patricia Drolsbach and Nicolas Pröllochs. Diffusion of community fact-checked misinformation on twitter. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–22, 2023.
53. Garfield Benjamin. Who watches the birdwatchers? sociotechnical vulnerabilities in twitter's content contextualisation. In *International Workshop on Socio-Technical Aspects in Security*, pages 3–23. Springer, 2021.
54. Stefan Wojcik, Sophie Hilgard, Nick Judd, Delia Mocanu, Stephen Ragain, MB Hunzaker, Keith Coleman, and Jay Baxter. Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation. *arXiv preprint arXiv:2210.15723*, 2022.
55. Isaiah Jones, Brent Hecht, and Nicholas Vincent. Misleading tweets and helpful notes: Investigating data labor by twitter birdwatch users. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing*, pages 68–71, 2022.
56. T Niederkrotenthaler, Z Laido, S Kirchner, M Braun, H Metzler, T Waldhör, MJ Strauss, D Garcia, and B Till. Mental health over nine months during the sars-cov2 pandemic: Representative cross-sectional survey in twelve waves between april and december 2020 in austria. *Journal of affective disorders*, 296:49–58, 2022.
57. Samin Aref and Mark C Wilson. Measuring partial balance in signed networks. *Journal of Complex Networks*, 6(4):566–595, 2018.
58. Samin Aref, Andrew J Mason, and Mark C Wilson. A modeling and computational study of the frustration index in signed networks. *Networks*, 75(1):95–110, 2020.
59. Patrick Doreian and Andrej Mrvar. Partitioning signed social networks. *Social Networks*, 31(1):1–11, 2009.
60. David Schoch. *signnet: An R package to analyze signed networks*, 2020.
61. James A Davis. Clustering and structural balance in graphs. *Human relations*, 20(2):181–187, 1967.
62. MK Vijaymeena and K Kavitha. A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, 3(2):19–28, 2016.
63. Ryan J Gallagher, Morgan R Frank, Lewis Mitchell, Aaron J Schwartz, Andrew J Reagan, Christopher M Danforth, and Peter Sheridan Dodds. Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts. *EPJ Data Science*, 10(1):4, 2021.
64. Wikipedia. Querfront — wikipedia, die freie enzyklopädie, 2023. [Online; Stand 12. September 2023].
65. 2015 European migrant crisis, June 2023. Page Version ID: 1159102024.
66. 2015–16 New Year's Eve sexual assaults in Germany, June 2023. Page Version ID: 1159999875.
67. David Garcia, Adiya Abisheva, Simon Schweighofer, Uwe Serdült, and Frank Schweitzer. Ideological and temporal components of network polarization in online political participatory media. *Policy & internet*, 7(1):46–79, 2015.
68. Delia Baldassarri and Andrew Gelman. Partisans without constraint: Political polarization and trends in american public opinion. *American Journal of Sociology*, 114(2):408–446, 2008.