sociological science

Echo Chambers Are Defined by Conflict, Not Isolation

Anna Keuchenius, Petter Törnberg, Justus Uitermark

University of Amsterdam

Abstract: The influential "echo chamber" hypothesis suggests that social media drive polarization through a mutual reinforcement between isolation and radicalization. The existence of such echo chambers has been a central focus of academic debate, with competing studies finding ostensibly contradictory empirical evidence. This article identifies a fundamental methodological limitation of these empirical studies: they do not differentiate between negative and positive interactions. To overcome this limitation, we develop a method to extract signed network representations of Twitter/X debates using machine learning. Applying our approach to a major Dutch cultural controversy, we show that the inclusion of negative interactions provides a new empirical picture of the dynamics of online polarization. Our findings suggest that conflict, not isolation, is at the heart of polarization.

Keywords: polarization; conflict; echo chambers; social media; signed networks

Reproducibility Package: The data and code underlying this article are available as part of our replication materials available at this link: https://doi.org/10.6084/m9.figshare.30948659.v1 or the DOI: 10.6084/m9.figshare.30948659. Due to Twitter/X's Terms of Service and ethical and legal obligations, we do not share tweet text, user IDs, or any data that could identify individuals or their stance in the debate. The dataset contains sensitive information, including political opinions, and releasing identifiable content would pose ethical risks to users and violate GDPR requirements. To support replication, we provide code, documentation, and non-identifying data sufficient to reproduce all analytical steps, with full analyses possible via rehydration of tweet IDs.

Citation: Keuchenius, Anna, Petter Törnberg, and Justus Uitermark. 2026. "Echo Chambers Are Defined by Conflict, Not Isolation" Sociological Science 13: 565-588.

Received: January 8, 2025

Accepted: January 12, 2026

Published: May 11, 2026

Editor(s): Arnout van de Rijt, Bart Bonikowski

DOI: 10.15195/v13.a22

POLITICS in many countries has in recent decades entered an era of unprecedented political polarization, with growing divides between political camps and harshening public discourse (Carothers and O'Donohue 2019; Harteveld 2021; Hobolt, Leeper, and Tilley 2021). Scholars have long discussed new media technology as a potential driver of this rise of polarization (Sunstein 1999, 2001a; Prior 2007; Iyengar and Hahn 2009; Sunstein and Vermeule 2008; Pariser 2011; Lelkes, Sood, and Iyengar 2017; Eady et al. 2019), with the so-called "echo chamber" as the most prominent causal link. According to this hypothesis, new media technology facilitates the formation of clusters of like-minded individuals (McPherson, Smith-Lovin, and Cook 2001). When such homogeneous groups are insulated from opposing perspectives, their views and biases are reinforced rather than moderated, resulting in polarization (Sunstein 1999; Mutz and Martin 2001; Lawrence, Sides, and Farrell 2010; Stroud 2010; Bakshy, Messing, and Adamic 2015; Nikolov et al. 2015; del Vicario et al. 2016; Schmidt et al. 2018). While this hypothesis is intuitive and has been widely adopted by scholars as well as the general public, a growing number of empirical studies find that social media users are in fact engaging in significant interaction across ideological divides (Goel, Mason, and Watts 2010; Yardi and boyd 2010; Conover et al. 2011a; Yoon and Park 2014; Gruzd and Roy 2014;

Colleoni, Rozza, and Arvidsson 2014; Bakshy et al. 2015; Vaccari et al. 2016; Barberá 2020). Recent literature questions the prevalence of online echo chambers and, by implication, the role of new media technology in driving polarization (Guess et al. 2018; Barberá 2020; Bail 2021).

This article points to a fundamental methodological limitation of many empirical studies on echo chambers: they do not differentiate between negative and positive interactions. Leveraging the use of networks to study social dynamics online (Lazer et al. 2009), user interactions in this line of research are represented as ties, and patterns in discussions are mapped through graphs (Conover et al. 2011a; Bakshy et al. 2015; Barberá et al. 2015). Typically, researchers do not distinguish between negative and positive interactions: ties can vary in strength, but they are generally assumed to be positive. Such a representation limits our understanding of the role of conflict in social media debates by either ignoring negative interactions or conflating them with positive interactions. In this study, we present a method for bringing conflict into these network representations. Acknowledging the positive or negative sentiment of interactions provides a richer empirical analysis and allows us to go beyond the echo chamber hypothesis toward a more nuanced examination of the driving forces of online polarization.

Taking the Dutch cultural conflict around Zwarte Piet on Twitter as a case, we use natural language processing to classify interactions as either positive or negative. We ask: how do groups emerge out of positive and negative interactions in online debates? The signed network analysis shows that most of the cross-ideological interaction is negative, involving direct attacks or forms of outgroup derogation. By comparing the results of our signed network analysis with traditional unsigned network approaches, we find that acknowledging negative interactions results in a qualitatively different understanding of online polarization, revealing different types of actors and different kinds of relationships. Moreover, we show that polarization is asymmetrical, with opposing groups displaying uneven levels of hostility.

Our results provide empirical evidence for an alternative theoretical understanding of echo chambers, by positing conflict, not isolation, as the driver of social media polarization. Although the traditional echo chamber hypothesis suggests a mutual reinforcement between isolation and radicalization as the core feedback process behind polarization (Sunstein 1999, 2001a), the findings of this article align with a recent model suggesting that social media may, in fact, drive polarization by intensifying conflictual interaction (Törnberg 2022). Such a perspective aligns with classic characterizations of the sociology of conflict, as developed by Simmel (1904a, 1904b, 1904c) and elaborated by Coser (1957) and Collins (2012). This provides the starting point for reorienting computational research on polarization to better account for conflict and develop a more complex understanding of the impact of social media on political life.

Echo chambers on social media

The notion of the "echo chamber" suggests that new media technologies enable us to avoid the discomfort of exposure to opposing ideas or opinions by letting us

choose to connect with like-minded people and seek out information that confirms our views. Scholars have argued that such "selective exposure" (Garrett 2009) has even become automated, as algorithms personalize our news and information environment, creating "filter bubbles" that shield us from dissenting views (Pariser 2011). Selective exposure can result in a polarization feedback loop, in which partisan media exposure strengthens beliefs, which in turn reinforces media selection (Stroud 2010; Slater 2007). This literature thus suggests that polarization fundamentally comes about through lack of exposure to alternative views, resulting in a breakdown of the foundations of democratic pluralism (Sunstein 1999, 2001b; Quattrociocchi et al. 2016; Törnberg 2018).

A significant strand of empirical research has set out to test the echo chamber hypothesis by examining the structure of interactions on social media, especially Twitter. Early studies employed network representations of retweets between users—generally understood as a sign of endorsement (Metaxas, 2015)—to examine the structure of political debates. These studies tend to find that users indeed are organized in separate network communities, with limited interconnections across political divides (Conover et al. 2011a; Barberá et al. 2015; Himelboim et al. 2016; Guerrero-Solé 2017; Recuero et al. 2019). In fact, retweet structures and political ideology are so strongly correlated that user ideology can be predicted from retweet networks (Conover et al. 2011b; Barberá 2015). These findings have been taken as support for the echo chamber hypothesis (Sunstein 2001a; McPherson et al. 2001; Mutz and Martin 2001; Sunstein and Vermeule 2008; Stroud 2010)—an idea that has become widely accepted and highly influential in the public debate (El-Bermawy 2016; Hooton 2016; D'Costa 2017).

However, studies that construct their networks based on mentions between users, rather than retweets, typically come to the opposite conclusion: they do not find a significant divide between opposing groups (Honeycutt and Herring 2009; Yardi and boyd 2010; Conover et al. 2011a; Barberá 2015; Williams et al. 2015) but suggest that "mentions form a communication bridge across which information flows between ideologically-opposed users" [47:6]. Interpreting mentions as neutral interaction, or as expressions of information exchange (Borge-Holthoefer et al. 2011), these studies have been used to question the prevalence of echo chambers—and at times even the link between social media and political polarization (Barberá et al. 2015; Bruns 2017, 2019, 2021; Guess et al. 2018; Bail 2021; Törnberg et al. 2021).

The assumption that interaction across political divides could be understood as neutral, or as "information flows" (Conover et al. 2011a), can, however, be questioned based on findings of qualitative research. For instance, Evolvi [60:396] finds that mentions are often used to "belittle others with different ideas rather than invite conversation." Studying climate change debates on Twitter, Moernaut, Mast, and Temerman (2020) find that climate change skeptics and believers do not engage in constructive debate but rather aim at delegitimizing and dehumanizing one another. Gruzd and Roy (2014) manually labeled tweets across party lines in Canadian politics, finding that roughly half are hostile. Williams et al. (2015) report evidence that mentions express positive ingroup and negative outgroup sentiments. These findings suggest an important limitation at the heart of the current approach to studying echo chambers in social media: this research has either left out negative interactions altogether (studies of retweets) or conflated positive and negative

interactions (studies of mentions), thus missing the conflictual dimension of online debates. This points to the need of developing methods that allow bringing conflict into the study of social media polarization.

Signed network analysis

Social network analysis has, in recent years, grown into one of the foremost means to quantitatively study social structures (Lazer et al. 2009). This body of research almost exclusively represents interactions between individuals as positive or neutral, treating highly connected groups of individuals as belonging to the same community (Harrigan et al. 2020). Recent years have seen the incipient growth of empirical research on signed networks, which goes beyond this representation by acknowledging the polarity of ties, with negative ties implying repulsion and positive ties attraction. Such networks have been employed to test social balance theory (Heider 1946; Cartwright and Harary 1956) and status theory (Guha et al. 2004; Leskovec, Huttenlocher, and Kleinberg 2010a; Hassan, Abu-Jbara, and Radev 2012a, 2012b; Zheng, Zeng, and Wang 2015; Sadilek, Klimek, and Thurner 2018), to measure the impact of negative ties in an organizational context (Labianca and Brass 2006), and to predict tie formation (Tang et al. 2016) (for an overview of research on signed social networks see Harrigan et al. [2020]). A few studies have employed signed networks to study polarized debates (Traag and Bruggeman 2009; Uitermark et al. 2016; Neal 2020). However, the potential of signed network analysis to study polarization on social media has remained underutilized, with a small number of studies that are relatively small in scope.

The limited use of signed network analysis is in large part a result of the relative difficulty involved in extracting signed network data from social media. Scholars have had to either manually classify relations as positive or negative (Gruzd and Roy 2014; Moernaut et al. 2020), focus on one of a small number of niche social media platforms that employ negative ties (Guha et al. 2004; Kunegis et al. 2009; Leskovec et al. 2010a, 2010b), or make strong assumptions on the sign of ties (de Stefano and Santelli 2019; Yardi and boyd 2010). None of these paths is passable for the application of signed networks to the study of large-scale dynamics of echo chambers on mainstream social media platforms such as Twitter.

This article presents a method for moving beyond this impasse by automatically extracting signed networks from large-scale social media debates by using natural language processing and machine learning. We manually code the sentiment expressed in relation to mentioned users for a large sample of tweets. Each user-to-user mention was classified as expressing agreement (positive), disagreement (negative), or as ambiguous/neutral toward the mentioned user. This data was used to train a machine learning algorithm to automatically classify mentions as positive or negative, resulting in a large signed social network of user relations. (See the Methods section for additional details.)

Conflictual debate about "Zwarte Piet"

"Zwarte Piet" ("Black Pete") is traditionally a key figure in the celebration of Sinterklaas, a Dutch variant of Christmas. While Sinterklaas looks somewhat like Santa

Claus (an old white man dressed in red, often riding a white horse), his helper Zwarte Piet is performed by white people wearing blackface, with exaggerated red lips, a black curly wig, and large golden earrings. Zwarte Piet has in recent years become a lightning rod in the Dutch culture wars, bringing about a protracted and intense national debate (Wekker 2016; Chauvin, Coenders, and Timo Koren 2018; Vliegenthart and Zuure 2020). Although the debate on Zwarte Piet is complex and sprawling, in essence, opponents of Zwarte Piet see the abolishing of the Zwarte Piet character as one key battle in the struggle against the legacy of colonialism and racism (Helsloot 2009;Rodenberg and Wagenaar 2016; Wekker 2016; Vliegenthart and Zuure 2020). In contrast, supporters of Zwarte Piet say that the character is not racist and perceive the suggestion to change Zwarte Piet as an attack on a valued Dutch tradition (Helsloot 2012; Rodenberg and Wagenaar 2016; Vliegenthart and Zuure 2020).

We use a dataset of tweets on Zwarte Piet to study this contentious debate. The data cover the period from December 2017 to May 2019, comprising roughly 430,000 tweets, from 81,700 unique users, with 296,881 unique mentions between users. To examine how the differentiation of negative, neutral, and positive ties changes the findings of network analysis, we compare the signed network with the unsigned retweet and mention networks that are generally employed to study polarization.

Results

Echo chambers through the lens of signed network analysis

The included nodes and edges constitute a first notable difference in how the retweet, mention, and signed networks represent the Twitter data: comparing the two unsigned networks in terms of nodes, we find that the retweet network includes more users (30,493; 90.1 percent of the users in the signed network) (see Figure 1 and Figure 2A) than the mention network (17,446; 52.0 percent of the users in the signed network) (see Figure 2A). Although the number of users that the mention network includes but the retweet network misses is relatively small (3,062; 9.1 percent of the users in the signed network), it includes prominent and important actors in the national debate (see Figure 2B), such as the politicians Lodewijk Ascher (@lodewijka), Klaas Dijkhoff (@dijkhoff), and Jesse Klaver (@jesseklaver). The signed network contains users from both the mention network and the retweet network (33,555). In terms of edges, the most striking feature of the retweet network is that it misses all negative user-to-user interactions. Additionally, it leaves out a substantial fraction of the positive relations (26,413; 13.4 percent; see Figure 3A). The mention network misses the majority of positive user-to-user relations (157,854; 80.2 percent of relations classified as positive; see Figure 2B) and misrepresents a substantial number of negative user-to-user relations (34,884; 30.2 percent of all mention relations; see Figure 2).

We now turn to examine what these differing network representations tell us about echo chambers. Focusing on retweets, it seems that the two sides of this debate operate in separate universes, with little cross-ideological interactions (see Figure 4), in line with studies that interpret such a structure as empirical support

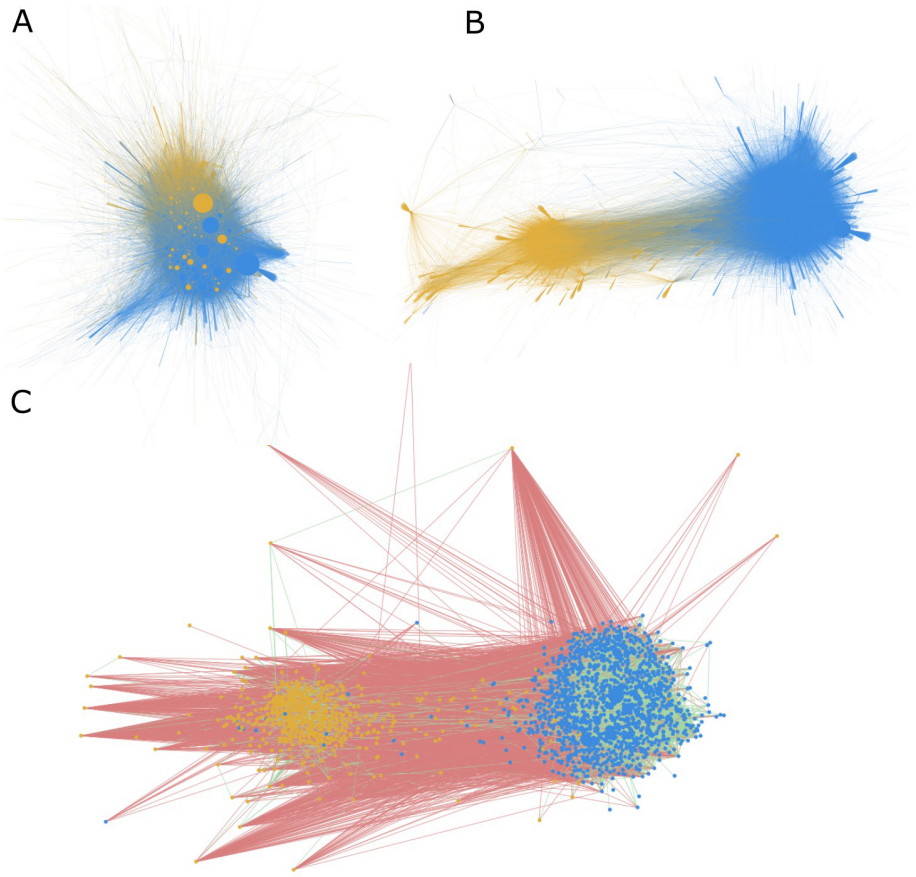

Figure 1: Visualizations of the mention network (A), retweet network (B), and signed network (C) constructed from the Zwarte Piet Twitter data. Users are positioned according to the Force Atlas algorithm and colored by their stance in the debate (proponents blue and opponents yellow). The positive and negative edges in the signed network are colored green and red, respectively. In the mention and retweet network, the Force Atlas algorithm (implemented in Gephi) draws connected users closer together, whereas in the signed network, the adjusted Force Atlas algorithm (implemented in Sigma.js) attracts nodes connected with a positive edge and repulses nodes connected with a negative edge. The figure shows that the mention network forms a seemingly cohesive whole, the retweet network is split into two opposing groups with little cross-interaction, and the signed network displays two groups with positive ingroup and negative outgroup interactions.

for the existence of echo chambers (Conover et al. 2011a, 2011b; Barberá et al. 2015; Guerrero-Solé 2017; Himelboim et al. 2016; Soares et al. 2019). The large majority of retweet relations connect internally either proponents (76.1 percent) or opponents (20.4 percent), and only a negligible fraction cuts across the divide (3.5 percent).

The mention network view on echo chambers, however, suggests a very different picture. This network is highly integrated, with frequent connections between the two opposing sides. In fact, a large majority (64.5 percent) of mention relations connect users with different stances (see Figure 4). This is in line with the research that has been seen as providing evidence to question the echo chamber hypothesis, finding that "mentions form a communication bridge across which information

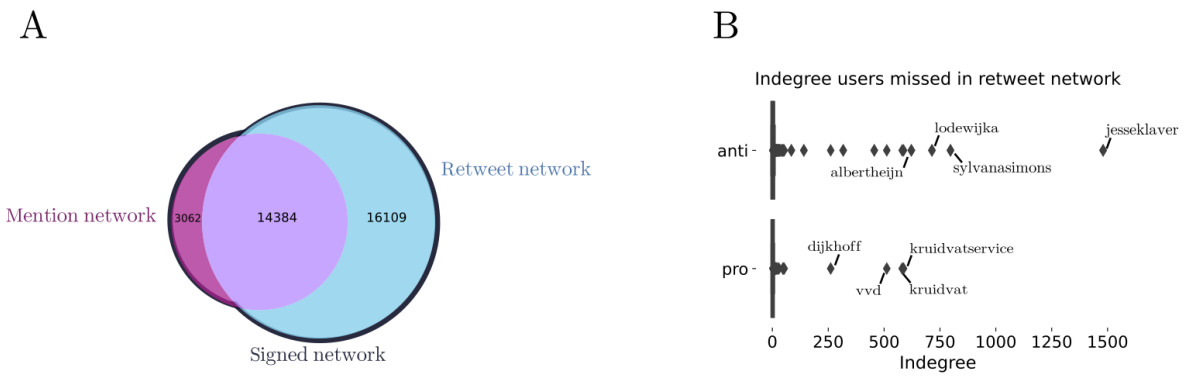

Figure 2: Differences and similarities between the sets of users identified in, respectively, the retweet network, the mention network, and the signed network. The Venn diagram of user sets (A) shows that the signed network contains the union of users in the mention and retweet network. The boxplots (B) give an impression of the prominence of users who are not included in the retweet network (because they did not send tweets or retweets on the topic). Although the retweet network includes most users (n = 30,493 or 90.1 percent), it misses some prominent users. For example, politician Jesse Klaver, an opponent of Zwarte Piet, received over 1.400 mentions.

flows between ideologically-opposed users" [48:6] (Honeycutt and Herring 2009; Yardi and boyd 2010; Conover et al. 2011a).

The signed network provides a third perspective, which posits an explanation for the contradictory findings and interpretations. The signed analysis shows that, although there are indeed many connections across the political divide, the majority of such intergroup relations (81.6 percent) are negative. Intragroup relations, in contrast, are almost exclusively positive (96.6 percent) (see Figure 4). This perspective provides a qualitatively different view from both the proponents and opponents of the echo chamber hypothesis: there are ideological groupings on social media, but they are not defined by the flow of information, but by active conflict. The sides are engaged in expressing support and solidarity for their ingroup while deriding the outgroup. Members of ideological groups use retweets and mentions to express support and solidarity for the ingroup, and condemnation for the outgroup. Rather than Sunstein's (Sunstein 1999) suggestion of homogeneity being the driver of polarization, this suggests that polarization results from a feedback cycle between external conflict and internal solidarity (Coser 1957; Collins 2012).

So far, we identified groups in the debate based on their issue position (are they in favor or against Zwarte Piet?). Another common method of defining groups is to infer them from interactions by means of community detection. Do our conclusions hold up when we use this method? To answer this question, we apply the Constant Potts Model (Reichardt and Bornholdt 2006) implemented in the Leiden algorithm (Traag and Bruggeman 2009; Traag, van Dooren, and Nesterov 2011; Traag, Waltman, and Jan van Eck 2019), which can use both positive edges (attracting nodes) and negative edges (repulsing nodes), see the Methods section for details. For the retweet network, we find a clear divide between two groups (CPM quality = 162,862), which aligns closely with the two sides of the debate (one group has 94 percent proponents, the other 95 percent opponents; see Figure 1). The mention network is instead structured as a single cohesive community, including

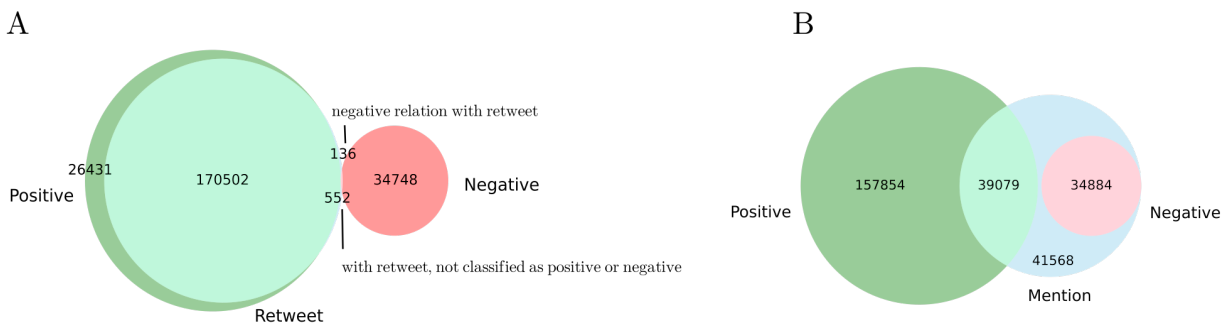

Figure 3: Differences and similarities between the sets of relations identified in, respectively, the retweet network, the mention network, and the signed network. The Venn diagrams demonstrate how the signed relations between users—positive (green) and negative relations (red)—compare to the user-to-user relations based on retweets (A) or mentions (B). Taking the signed relations as a baseline, the figure shows that retweet relations capture the large majority of positive relations but miss all negative relations between users. The mention relations include 19.8 percent of the positive relations and include all relations identified as negative but conflate these two types of ties. The mention relations that are not included in the signed relations (n = 41.468) are mostly based on only one interaction (85.9 percent).

users from both sides of the divide (CPM quality = 112,367, see Figure 1). The signed network again improves both these representations, showing two sides engaged in conflict, with relations between opposing groups being predominantly negative (85 percent; see Figure 1).

The signed network analysis thus suggests a fundamentally different picture of echo chambers: they are not isolated groups of the likeminded people nor are echo chambers non-existent due to connections across political divides. Instead, they are defined through the conflictual interactions across groups and expressions of support and solidarity within them. This perspective allows us to look deeper into the structure of intergroup conflict, bringing to the fore asymmetries between the two sides of the debate, the topic to which we now turn.

Structure of intragroup conflict

Signed network analysis allows us to cast light on asymmetrical polarization. Such asymmetry is an important theme within the literature on political polarization (Grossmann and Hopkins 2016; Barberá 2020), and signed network analysis enables us to examine asymmetries in the structure and levels of conflictual relations. As the opposition between supporters and opponents of Zwarte Piet maps onto the opposition between conservatives and progressives, we relate findings from our case study to broader literature on asymmetric polarization.

We should first note that there are remarkable differences in terms of activity, see Figure 5. On average, proponents of Zwarte Piet tweet 2.5 times more than opponents (14.7 vs. 5.7 tweets per user) and have roughly 2.5 times as many outgoing edges in the network (9.29 vs. 3.64). The proponents are also more confrontational: a third (33.3 percent) of the proponents has at least one negative

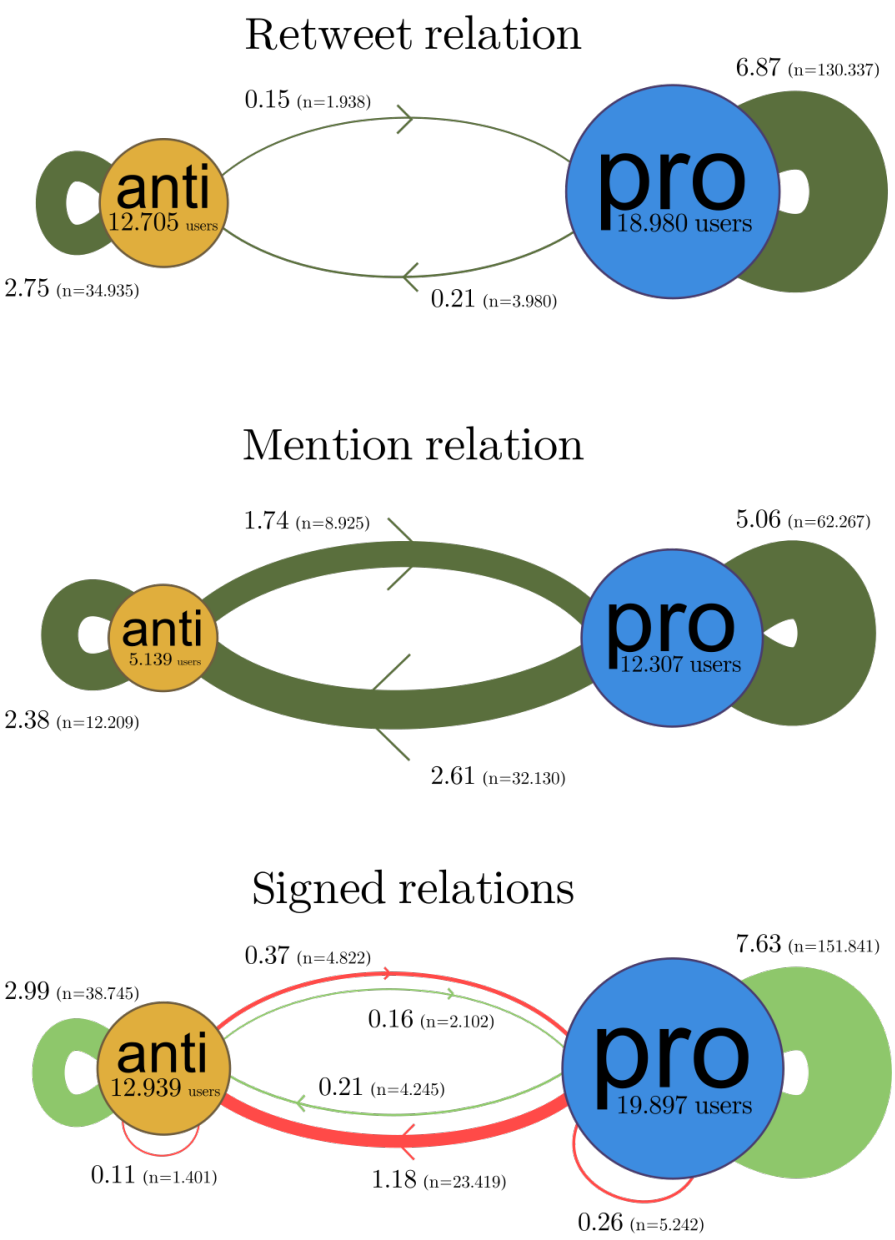

Figure 4: These figures display the retweet, mention, and signed relations between the two sides. The edges are weighted as the count of the type of relation divided by the size of the user-group, for example, the retweet pro–pro edge has a weight of 130,337/18,980 = 6.87, which can be interpreted as an average; for example, opponents retweet on average 6.87 other opponents. This figure shows that the aggregate interactions between proponents and opponents depend on which types of interaction (retweet, mentions, or signed) are considered.

outgoing link in the network, compared to only 18 percent of opponents, and the negative interaction rate (the number of outgoing negative edges divided by total outgoing edges per user) is twice as high for proponents than for opponents (0.07 vs. 0.14). As a result of their higher levels of activity and negativity, proponents

Average stats pro/anti users

Issue sentiment	anti	pro
Total (re)tweets	5.72	14.73
Negative links	0.48	1.44
Positive links	3.16	7.84
Total links	3.64	9.29
Positive links to anti	2.99	0.21
Positive links to pro	0.16	7.63
Negative links to anti	0.11	1.18
Negative links to pro	0.37	0.26
Negative interaction rate	0.07	0.14

Figure 5: Average statistics for outgoing links in the signed network for opponents (anti) and proponents (pro) of Black Pete. The negative interaction rate is calculated by the number of negative outgoing links divided by the total outgoing links of each user, and thereafter averaged over all users.

have, on average, three times more negative outgoing links than opposed users (1.44 vs. 0.48), most of which are directed toward the opposing side.

Looking at the distribution of negative links per user (Figure 6), we find that the distribution of negative links over users is highly skewed: the proponents of Zwarte Piet include a group of very active and highly confrontational users who account for a large proportion of the negative relations overall. As most of the negative tweets are directed at people holding a different position on Zwarte Piet, it is unsurprising that the top targets of attack tend to be opponents of Zwarte Piet. A few key figures receive most of the negative ties. Opponents of Zwarte Piet users represent 72.2 percent of the top-1 percent (320 users) most negatively referenced users and 87.1 percent of the top-0.1 percent (32 users).

These results suggest, first of all, that there is indeed asymmetrical polarization: supporters of Zwarte Piet are much more active in the debate and are more negative. This means not only that different sides to the debate behave differently but also that they receive different treatment: opponents of Zwarte Piet are much more likely to have negative tweets directed against them. Moreover, we find that a very active and highly negative group of supporters of Zwarte Piet accounts for much of this pattern. These findings suggest that (1) interactions across the political divide often take the form of criticism, derogation, or intimidation and (2) supporters of Zwarte Piet account for a much larger portion of the negativity than proponents. The results not only confirm asymmetrical polarization, but they are also in line with other recent social media research that shows conservatives interact more across partisan divides (Eady et al. 2019; Wu and Resnick 2021; Gaisbauer et al. 2021)—but adding the central point that this interaction more often tends to be confrontational.

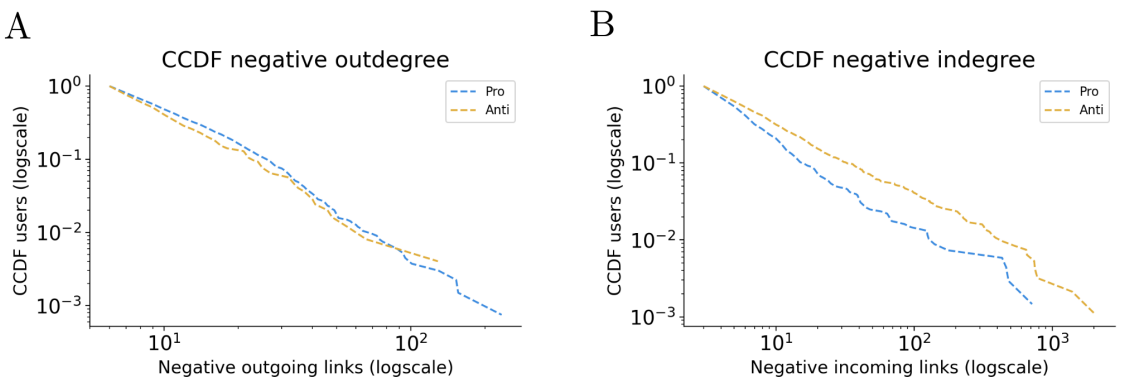

Figure 6: CCDF of negative outgoing links (A) and incoming links (B) per user. The highly skewed distributions of negative outdegree show that most users have few negative outgoing relations, but there are some users, more among the supporters of Zwarte Piet (blue) than among opponents (orange), with many negative outgoing edges. The different slope between the negative indegree distribution between proponents and opponents shows that the latter structurally receives more negativity.

Discussion

Studies of polarization on social media have been limited by their use of unsigned networks, which requires researchers to either leave out negative interaction (by, e.g., looking at Twitter retweets only) or to conflate negative and positive relations (by, e.g., encoding Twitter mentions as positive network ties). This has led to puzzling empirical findings with respect to the echo chamber hypothesis, as studies either find ideological echo chambers or substantive cross-ideological engagement, depending on whether they choose the former or the latter approach to representing user interaction.

This article has presented a method for moving beyond this methodological impasse by introducing a method for extracting signed networks from Twitter data and using this to throw new light on the nature of the echo chamber and interaction across the political divide on social media. We compare the picture of a Dutch cultural controversy on Twitter as represented by three forms of network representations: the mention network, the retweet network, and the signed network. In line with previous research, we find that the retweet network reveals two separate echo chambers, with users self-segregated into two isolated groups. The mention network, contrastingly, shows substantial communication across the ideological divide. These seemingly contradictory findings are resolved by using a signed network, which shows that most of the cross-ideological user interaction is negative, whereas ingroup interactions are almost exclusively positive. This presents a novel understanding of echo chambers, in which they are defined not by isolation, but by intergroup conflict.

The signed network analysis allows a clearer image of the structure of online polarization, revealing asymmetries between the political factions. While previous studies have taken the higher level of interaction to suggest that conservatives are less locked into echo chambers than progressives, the signed network representation

suggests that conservatives are more active and more prone to attack their outgroup. As some of the most popular targets of these attacks include organizations or individuals that have neither stake nor position in the debate—such as supermarkets or political parties—this suggests that the aim of these interactions is not always to convince the opposing group, but also to show allegiance with the ingroup through symbolic attacks against the outgroup. This aligns with research suggesting the interaction with users of different views may not trigger moderation—as the common version of the echo chamber suggests—but may instead further intensify polarization (Bail 2021; Bail et al. 2018). It furthermore contributes important empirical evidence for a recent conflict-driven model of social media polarization, which "turns the echo chamber on its head" (6) by suggesting that social media may be polarizing by increasing interaction across the political divide, rather than isolating political opponents (Törnberg 2022).

These findings reveal how methodological choices of data representation can come to have profound consequences for how we understand social phenomena. Unsigned network representations bring about a false dichotomy between isolation and interconnection by ignoring conflictual interactions. This erases the central role of conflict and solidarity in online polarization and results in a theoretical foundation in which isolation is taken to be the driver of polarization (Sunstein 1999; Sunstein and Vermeule 2008; Bishop 2009; Pariser 2011). When instead representing social media interaction as signed networks, the resulting understanding of online polarization shifts significantly, revealing echo chambers as defined not by isolation from information flows but through intragroup solidarity and intergroup antagonism. This points to a very different theoretical foundation for the dynamics of polarization, drawing on a long sociological tradition which puts conflict as the chief driver of polarization. Scholars in this tradition, such as Simmel, Coser, and Collins, suggest that polarization results from a feedback loop in which external conflict drives internal solidarity, and vice versa. In this framework, to the extent that social media facilitates polarization, it is not because it isolates opposing communities, but, on the contrary, because it faces them off in contentious confrontation. The approach introduced in this article captures the conflictual dimension of polarization, allowing more nuanced insight into the underlying social mechanisms that tear social media users apart or pull them together.

Methods

Twitter data

We gathered tweets on the Black Pete debate by keyword matching of various terms related to the debate, such as "Black Pete," "Zwarte Piet," and "KOZP" (abbreviation for "Kick Out Zwarte Piet"), using the Twitter Capture Analysis Toolset (Borra and Rieder 2014) and removed tweets that were not written in Dutch. The tweets in our dataset were published between December 4th, 2017, and May 7th, 2019, and include original tweets and unquoted retweets. In total, the dataset contains 418,421 tweets from 61,543 unique users, with 174,555 unique mentions between users.

Classifying users' stance

In large-scale studies on political polarization on Twitter, the ideological stance of users is typically inferred from their interactions (Barberá et al. 2015; Bail et al. 2018). In this study, we explicitly opt not to do this to investigate how users from the same and different stances interact with one another in the debate. Additionally, we depart from previous approaches by not limiting the analysis to the more active users is the debate which creates a bias toward the vocal minority to the detriment of the more silent majority (Mustafaraj et al. 2011). Instead, we select a method for inferring the user's stance that is as inclusive as possible. Our strategy, then, is to classify the position toward Black Pete that is expressed in all the (re)tweets of the user (pro, anti, or neutral/ambiguous) by examining the full tweet texts and use that to deduce a user's stance on Black Pete.

To classify the position expressed in tweets, we semi-manually classified a sample of tweets and thereafter trained a machine learning algorithm with this data. The sample data consist of two sets of tweets. First, 4,787 (2.7 percent) tweets were selected at random from the full set of unique tweets (n = 179,712). These tweets were manually labeled with the assistance of four fluent Dutch speakers. Each tweet was assigned one label: pro, anti, neutral, or ambiguous. The codebook instructions were conservative: if the stance toward Black Pete is not self-evident, the tweet was labeled as ambiguous. From the coding efforts, we learned that it was often difficult to distinguish neutral from ambiguous tweets, and we found few tweets (n = 512) that were coded as expressing a neutral issue sentiment. Therefore, we merged the neutral and ambiguous tweets into one category for subsequent classification purposes. The inter-coder agreement was measured by a Krippendorf alpha of 0.724.

The second set of the training data consists of tweets by prominent pro and anti users in the debate. We selected the top 1 percent accounts of users that were most active either in terms of retweets, mentions, or the number of unique users mentioned or retweeted. Similarly, we selected the top 1 percent of the accounts that received the most retweets or mentions, in terms of frequency and in terms of the number of unique source users. These top users (n = 329) were manually labeled as having a pro (n = 231), anti (n = 59), or neutral/ambiguous (n = 39) stance in the debate. The tweets of these top pro (anti) users that weren't also retweeted by the opposite side of the debate were then classified as pro (anti). This resulted in an extra 26,323 labeled tweets.

After splitting the data into a test set (40 percent) and a training set (60 percent), we downsampled the tweets with a pro label in the training set to equal the number of anti tweets in the training set (n = 2,607) to avoid biases in the classification. We resampled the test set to have the same distribution of pro, anti, and neutral/ambiguous tweets as the original full dataset, respectively 60, 12, and 28 percent. Next, we used this data to train the fastText algorithm (Joulin et al. 2017) with pretrained word embeddings on a Dutch Wikipedia Corpus (Bojanowski et al. 2017), maximizing the $F1$ score for all classes, thus attempting to predict all classes well, in both precision and recall. The fastText algorithm gives an indication of how certain the classification is (the softmax probability), valued between 0 and 1 for each prediction. We use this certainty indication to apply a simple rule: classify all

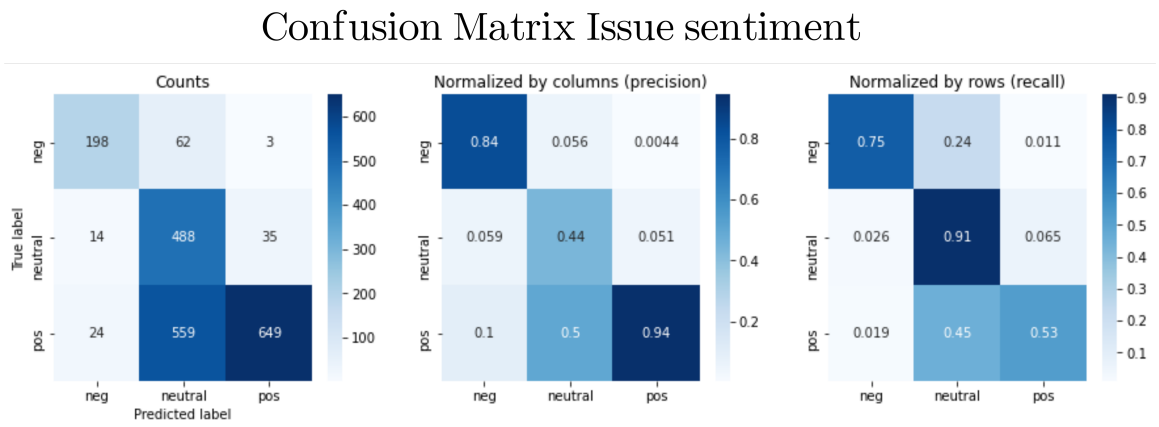

Figure 7: The results of the classifier (parameter values: epoch = 10, learning rate = 0.7, *n*-grams = 3) after applying the simple certainty rule (neutral if certainty <0.99): confusion matrix with counts (left), normalized by the true labels (middle), and normalized by the predicted labels (right). The values in the diagonals of the middle matrix are the precision rates, and the values on the diagonals of the right matrix are the recall rates. Recall rates here are reduced due to the certainty rule, but the most important errors (classify positive if the true value is negative and classify negative if the true value is positive) are reduced.

tweets with lower certainty (<0.99) as neutral/ambiguous. This procedure reduces the recall for the pro and anti classes but also, more importantly, reduces the errors we care most about: classifying pro tweets as anti and classifying anti tweets as pro.

The classifier—after applying the certainty rule—categorizes the issue stance with sufficiently high accuracy, see Figure 7. There are only three cases in which an anti tweet is misclassified as pro (0.011) times of all anti tweets and 0.0044 times all pro tweet classifications) and 24 cases in which a pro tweet is misclassified as anti (0.019 times of all pro tweets and 0.1 times of all anti classifications). Aggregating tweets per user and applying a simple majority rule results in 14,353 clear anti users (14 percent), 23,581 (38 percent) clear pro users, and 23,637 (38 percent) users that we couldn't unambiguously classify as pro or anti. Most of the users with an ambiguous stance have just one (re)tweet (75 percent), and we therefore left them out of the subsequent analysis without jeopardizing the robustness of our results.

Classifying signs of interaction

As the sign of interaction is an integral part of this study, we aim to measure this as accurately as possible. Previous studies have relied on heuristics to infer the sign of interaction, for example, using the balance theoretical notion that the enemy of my friend is my enemy (Heider 1946; Cartwright and Harary 1956). Instead, we analyze the text of tweets from one user directed toward another user by the use of mentions (e.g., @username). We classify for each mention whether the source user is expressing endorsement (positive), disagreement (negative), or an ambiguous (neutral) sentiment toward the mentioned user. This type of classification cannot be addressed with existing algorithms for sentiment analysis, because the sentiment toward the mentioned user is not necessarily aligned with the sentiment of the tweet: a tweet expressing a positive sentiment can contain a negative (hostile)

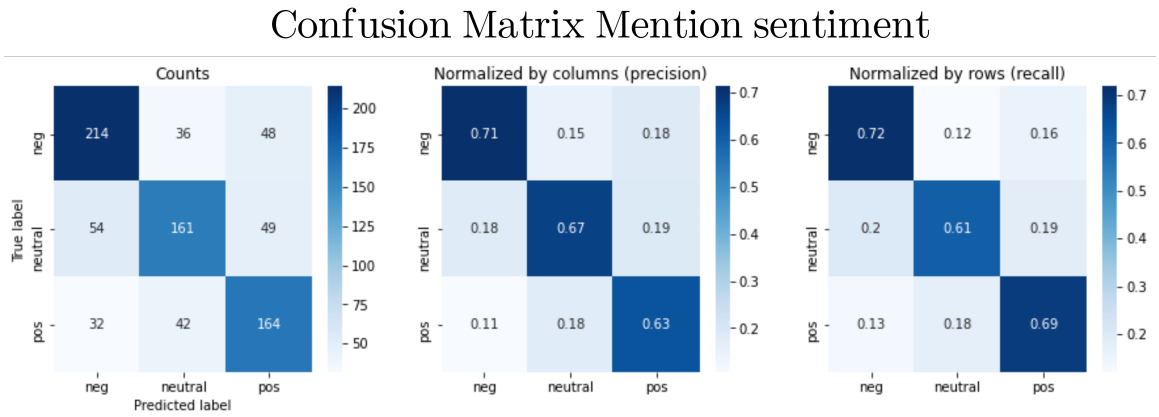

Figure 8: The results of the classifier (parameter values: epoch = 20, learning rate = 0.65, *n*-grams = 3): confusion matrix with counts (left), normalized by the true labels (middle), and normalized by the predicted labels (right). The values in the diagonals of the middle matrix are the precision rates, and the values on the diagonals of the right matrix are the recall rates. Recall rates here are reduced due to the certainty rule, but the most important errors (classify positive if the true value is negative and classify negative if the true value is positive) are reduced.

mention. Instead, our general strategy is to select a sample of mentions at random from the full dataset, label these manually, and with this sample train a machine learning algorithm. This method is inclusive with respect to different types of users, reducing the known bias in earlier research toward the vocal minority to the detriment of the more silent majority (Mustafaraj et al. 2011).

The random sample data (n = 6,056, 3.5 percent) contains unique user-to-user mentions from users that we identified as pro or anti. These tweets with mentions were then manually labeled with the assistance of four fluent Dutch speakers. Each mention was assigned one label: positive, negative, or neutral/ambiguous. The codebook instructions were conservative: if the interaction sentiment was not self-evident, the mention was labeled as neutral/ambiguous. The inter-coder agreement was measured by a Krippendorf alpha of 0.42.

After splitting the labeled data into a training set (70 percent) and a test set (40 percent), we removed all the mentions in the test set that occurred in a tweet that was also included in the training set (because one tweet can contain various mentions). We added features with (1) the predicted stance of the source user, (2) the predicted stance of the mentioned user, (3) whether the mention takes the form of "via @username," which are most often neutral, as they are automatically added by the webserver of the media outlet via which the tweet was posted, and (4) whether the mention is located at the start, body, or end of the tweet because that might correlate with the polarity of the mention.

Next, we used this input to train the fastText algorithm (Joulin et al. 2017). To teach the algorithm the basics of Dutch and Twitter language, we also provided fastText with a word embedding learned from a corpus of approximately 180 million Dutch Tweets posted in 2018 (see Supplementary material). We trained the algorithm to maximize the $F1$ score for all classes, thus attempting to predict all

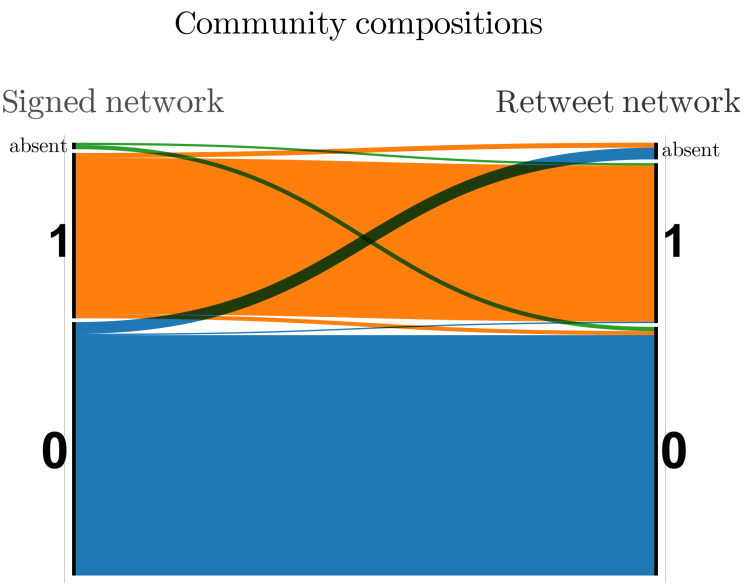

Figure 9: Alluvial graph illustrating the comparison of community compositions between the signed network (left) and retweet network (right). The thickness of lines corresponds to the number of users. The figure shows that the two communities in the networks are similar in composition.

classes well, in both precision and recall. The classifier categorizes the mention sentiment with sufficiently high accuracy, $F1$ score 0.67, see Figure 8.

We aggregated all the user-to-user interactions and labeled them based on a simple majority rule: if most of the user-to-user interactions were positive (negative), we classified the directed sign between these users as positive (negative). Retweets were considered a positive interaction from the retweeting to the retweeted user (Metaxas, 2015). This procedure classified most user relations as positive (n = 216,067, 75 percent), 12 percent as negative, and 14 percent of the relations as we could not classify with sufficient certainty. Most of the unclassified relations (86 percent) are based on one interaction only, and we therefore left them out of the subsequent analysis without jeopardizing the robustness of our results.

Community detection

Comparing the community structure of the retweet network, mention network, and signed network requires a method for community detection that can detect communities based on positive ties as well as negative ties. Additionally, we would like to compare the three networks at the same level of resolution. We therefore used a generalization of the Constant Potts Model (Reichardt and Bornholdt 2006) implemented in the Leiden algorithm (Traag and Bruggeman 2009; Traag et al. 2011, 2019). This method for community detection considers the sign of ties by maximizing positive ties within communities and minimizing negative ties within communities and allows to look at different granular scales of the community structure in the network.

We detected the community structure in the retweet network, mention network, and signed network with a resolution parameter gamma set to 0.0001. As reported

in the result section in more detail, the users in the mention network are all (99.6 percent) part of one community on this level, but both the retweet network and signed network feature two communities. These communities in the retweet and signed network are very similar in their composition, see Figure 9. As reported in the Results section, some small differences between the community compositions have far-reaching implications, because some of the users that the retweet network excludes are frequently attacked.

Data ethics

The data collection process has been carried out exclusively through the Twitter API, which is publicly available, and for the analysis, we used publicly available data (users with privacy restrictions are not included in the dataset). We abided by the terms, conditions, and privacy policies of Twitter. As this content is publicly published and is frequently discussed in mass media, we regard the debates as a public domain that does not require individual consent for inclusion in research, based on the ethical guidelines for internet research provided by The Association of Internet Researchers (Franzke et al. 2020) and by the British Sociological Association (BSA 2017). We only report on aggregates and limit reporting on details of individuals to user accounts that belong to public figures or institutions, or that have more than 4,000 followers. The data published along with this research do not include user IDs nor the classification of the sentiment on the Black Pete discussion, because this is part of a special category of personal data, formerly known as sensitive data.

References

Bail, Christopher A. 2021. Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing. Princeton: Princeton University Press. https://doi.org/10.1515/9780691216508

Bail, Christopher A., Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, et al. 2018. "Exposure to Opposing Views on Social Media Can Increase Political Polarization." *Proceedings of the National Academy of Sciences* 115(37):9216–221. https://doi.org/10.1073/pnas.1804840115

Bakshy, Eytan, Solomon Messing, and Lada A. Adamic. 2015. "Exposure to Ideologically Diverse News and Opinion on Facebook." *Science* 348(6239):1130–132. https://doi.org/10.1126/science.aaa1160

Barberá, Pablo. 2015. "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23(1):76–91. https://doi.org/10.1093/pan/mpu011

Barberá, Pablo. 2020. "Social Media, Echo Chambers, and Political Polarization." Pp. 34–55 in *Social Media and Democracy: The State of the Field, Prospects and Reform*, edited by N. Persily and J. A. Tucker. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108890960.004

Barberá, Pablo, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2015. "Tweeting from Left to Right: Is Online Political Communication More than

an Echo Chamber?" *Psychological Science* 26(10):1531–42. https://doi.org/10.1177/0956797615594620

Bishop, Bill. 2009. The Big Sort: Why the Clustering of Like-Minded America Is Tearing Us Apart. New York: Mariner Books.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. "Enriching Word Vectors with Subword Information." *Transactions of the Association for Computational Linguistics* 5:135–146. https://doi.org/10.1162/tacl_a_00051

Borge-Holthoefer, Javier, Alejandro Rivero, Iñigo García, Elisa Cauhé, Alfredo Ferrer, Darío Ferrer, et al. 2011. "Structural and Dynamical Patterns on Online Social Networks: The Spanish May 15th Movement as a Case Study." *PLOS ONE* 6(8):1–8. https://doi.org/10.1371/journal.pone.0023883

Borra, Erik, and Bernhard Rieder. 2014. "Programmed Method: Developing a Toolset for Capturing and Analyzing Tweets." *Aslib Journal of Information Management* 66(3):262–278. https://doi.org/10.1108/AJIM-09-2013-0094.

British Sociological Association (BSA). 2017. Statement of Ethical Practice.

Bruns, Axel. 2017. "Echo Chamber? What Echo Chamber? Reviewing the Evidence." Paper presented at the 6th Biennial Future of Journalism Conference (FOJ17), September 15. https://doi.org/10.64628/AA.t5x9eq6ue

Bruns, Axel. 2019. *Are Filter Bubbles Real?* Cambridge: Polity Press.

Bruns, Axel. 2021. "Echo Chambers? Filter Bubbles? The Misleading Metaphors That Obscure the Real Problem." Pp. 33–48 in *Hate Speech and Polarization in Participatory Society*. London: Routledge. https://doi.org/10.4324/9781003109891-4

Carothers, Thomas, and Andrew O'Donohue. 2019. Democracies Divided: The Global Challenge of Political Polarization. Washington, DC: Brookings Institution Press. https://doi.org/10.5040/9780815750819

Cartwright, Dorwin, and Frank Harary. 1956. "Structural Balance: A Generalization of Heider's Theory." *Psychological Review* 63(5):277–93. https://doi.org/10.1037/h0046049

Chauvin, Sébastien, Yannick Coenders, and Timo Koren. 2018. "Never Having Been Racist: Explaining the Blackness of Blackface in the Netherlands." *Public Culture* 30(3):509–26. https://doi.org/10.1215/08992363-6912163

Colleoni, Elena, Alessandro Rozza, and Adam Arvidsson. 2014. "Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data." *Journal of Communication* 64(2):317–32. https://doi.org/10.1111/jcom.12084

Collins, Randall. 2012. "C-Escalation and D-Escalation: A Theory of the Time-Dynamics of Conflict." *American Sociological Review* 77(1):1–20. https://doi.org/10.1177/0003122411428221

Conover, Michael D., Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. 2011a. "Political Polarization on Twitter." Proceedings of the 5th International AAAI Conference on Weblogs and Social Media. https://doi.org/10.1609/icwsm.v5i1.14126

Conover, Michael D., Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. 2011b. "Predicting the Political Alignment of Twitter Users." Proceedings of the 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, 192–9. https://doi.org/10.1109/PASSAT/SocialCom.2011.34

Coser, Lewis A. 1957. "Social Conflict and the Theory of Social Change." *British Journal of Sociology* 8(3):197–207. https://doi.org/10.2307/586859

D'Costa, Krystal. 2017. "A Nation Divided by Social Media." Scientific American Blog Network.

Eady, Gregory, Jonathan Nagler, Andrew Guess, Jan Zilinsky, and Joshua A. Tucker. 2019. "How Many People Live in Political Bubbles on Social Media? Evidence from Linked Survey and Twitter Data." *Sage Open* 9(1):1–21. https://doi.org/10.1177/2158244019832705

El-Bermawy, Mostafa M. 2016. "Your Filter Bubble Is Destroying Democracy." WIRED, November 18. Retrieved June 29, 2022 (https://www.wired.com/2016/11/filter-bubble-destroying-democracy/).

Evolvi, Giorgia. 2019. "#Islamexit: Inter-Group Antagonism on Twitter." *Information, Communication and Society* 22(3):386–401. https://doi.org/10.1080/1369118X.2017.1388427

Franzke, Aline Shakti, Anja Bechmann, Michael Zimmer, Charles Ess, and the Association of Internet Researchers. 2020. Internet Research: Ethical Guidelines 3.0.

Gaisbauer, Felix, Albert Pournaki, Sven Banisch, and Eckehard Olbrich. 2021. "Ideological Differences in Engagement in Public Debate on Twitter." *PLOS ONE* 16(3):e0249241–18. https://doi.org/10.1371/journal.pone.0249241

Garrett, R. Kelly. 2009. "Politically Motivated Reinforcement Seeking: Reframing the Selective Exposure Debate." *Journal of Communication* 59(4):676–99. https://doi.org/10.1111/j.1460-2466.2009.01452.x

Goel, Sharad, Winter Mason, and Duncan J. Watts. 2010. "Real and Perceived Attitude Agreement in Social Networks." *Journal of Personality and Social Psychology* 99(4):611–21. https://doi.org/10.1037/a0020697

Grossmann, Matthew, and David A. Hopkins. 2016. Asymmetric Politics: Ideological Republicans and Group Interest Democrats. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190626594.001.0001

Gruzd, Anatoliy, and Jeffrey Roy. 2014. "Investigating Political Polarization on Twitter: A Canadian Perspective." *Policy and Internet* 6(1):28–45. https://doi.org/10.1002/1944-2866.POI354

Guerrero-Solé, Frederic. 2017. "Community Detection in Political Discussions on Twitter: An Application of the Retweet Overlap Network Method to the Catalan Process toward Independence." *Social Science Computer Review* 35(2):244–61. https://doi.org/10.1177/0894439315617254

Guess, Andy, Benjamin Lyons, Brendan Nyhan, and Jason Reifler. 2018. Avoiding the Echo Chamber about Echo Chambers: Why Selective Exposure to Like-Minded Political News Is Less Prevalent Than You Think. Miami: Knight Foundation.

Guha, Ramayya, Prabhakar Raghavan, Ravi Kumar, and Andrew Tomkins. 2004. "Propagation of Trust and Distrust." Proceedings of the Thirteenth International World Wide Web Conference (WWW2004), pp. 403–12. https://doi.org/10.1145/988672.988727

Harrigan, Nicholas M., Giuseppe (Joe) Labianca, and Filip Agneessens. 2020. "Negative Ties and Signed Graphs Research: Stimulating Research on Dissociative Forces in Social Networks." *Social Networks* 60:1–10. https://doi.org/10.1016/j.socnet.2019.09.004

Harteveld, Eelco. 2021. "Fragmented Foes: Affective Polarization in the Multiparty Context of the Netherlands." *Electoral Studies* 71:102332. https://doi.org/10.1016/j.electstud.2021.102332

Hassan, Ahmed, Amr Abu-Jbara, and Dragomir Radev. 2012a. "Detecting Subgroups in Online Discussions by Modeling Positive and Negative Relations among Participants." Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 59–70.

Hassan, Ahmed, Amr Abu-Jbara, and Dragomir Radev. 2012b. "Extracting Signed Social Networks from Text." Proceedings of the TextGraphs-7 Workshop at ACL, 6–14.

Heider, Fritz. 1946. "Attitudes and Cognitive Organization." *Journal of Psychology* 21(1):107–112. https://doi.org/10.1080/00223980.1946.9917275.

Helsloot, John. 2009. "Is Zwarte Piet uit te Leggen?" Sinterklaas Verklaard, 76–85.

Helsloot, John. 2012. "Zwarte Piet and Cultural Aphasia in the Netherlands." *Quotidian: Journal for the Study of Everyday Life* 3:1–20.

Himelboim, Itai, Kaye D. Sweetser, Shannon F. Tinkham, Kirk Cameron, Matthew Danelo, and Kate West. 2016. "Valence-Based Homophily on Twitter: Network Analysis of Emotions and Political Talk in the 2012 Presidential Election." *New Media and Society* 18(7):1382–400. https://doi.org/10.1177/1461444814555096

Hobolt, Sara Binzer, Thomas J. Leeper, and James Tilley. 2021. "Divided by the Vote: Affective Polarization in the Wake of the Brexit Referendum." *British Journal of Political Science* 51(4):1476–93. https://doi.org/10.1017/S0007123420000125

Honeycutt, Courtenay and Susan C. Herring. 2009. "Beyond Microblogging: Conversation and Collaboration via Twitter." Proceedings of the 42nd Annual Hawaii International Conference on System Sciences, pp. 1–10.

Hooton, Christopher. 2016. "Social Media Echo Chambers Gifted Donald Trump the Presidency." The Independent, November 10. Retrieved June 29, 2022 (https://www.independent.co.uk/voices/donald-trump-president-social-media-echo-chamber-hypernormalisation-adam-curtis-protests-blame-a7409481.html).

Iyengar, Shanto, and Kyu S. Hahn. 2009. "Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use." *Journal of Communication* 59:19–39. https://doi.org/10.1111/j.1460-2466.2008.01402.x

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. "Bag of Tricks for Efficient Text Classification." Proceedings of the 15th Conference of the European Chapter of the *Association for Computational Linguistics* 2:427–31. https://doi.org/10.18653/v1/E17-2068

Kunegis, Jérôme, Andreas Lommatzsch, and Christian Bauckhage. 2009. "The Slashdot Zoo: Mining a Social Network with Negative Edges." Proceedings of the 8th International World Wide Web Conference (WWW '09), 741–50. https://doi.org/10.1145/1526709.1526809

Labianca, Giuseppe, and Daniel J. Brass. 2006. "Exploring the Social Ledger: Negative Relationships and Negative Asymmetry in Social Networks in Organizations." *Academy of Management Review* 31(3):596–614. https://doi.org/10.5465/amr.2006.21318920

Lawrence, Eric, John Sides, and Henry Farrell. 2010. "Self-Segregation or Deliberation? Blog Readership, Participation, and Polarization in American Politics." *Perspectives on Politics* 8(1):141–57. https://doi.org/10.1017/S1537592709992714

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, et al. 2009. "Computational Social Science." *Science* 323(5915):721–23. https://doi.org/10.1126/science.1167742

Lelkes, Yphtach, Gaurav Sood, and Shanto Iyengar. 2017. "The Hostile Audience: The Effect of Access to Broadband Internet on Partisan Affect." *American Journal of Political Science* 61(1):5–20. https://doi.org/10.1111/ajps.12237

Leskovec, Jure, Daniel Huttenlocher, and Jon Kleinberg. 2010a. "Predicting Positive and Negative Links." Proceedings of the International World Wide Web Conference, 641–50. https://doi.org/10.1145/1772690.1772756

Leskovec, Jure, Daniel Huttenlocher, and Jon Kleinberg. 2010b. "Signed Networks in Social Media." Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1361–70. https://doi.org/10.1145/1753326.1753532

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27:415–44. https://doi.org/10.1146/annurev.soc.27.1.415.

Metaxas, P., Mustafaraj, E., Wong, K., Zeng, L., O'Keefe, M., and Finn, S. (2015). "What do retweets indicate? Results from user survey and meta-review of research". In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 9, No. 1, pp. 658–661). https://doi.org/10.1609/icwsm.v9i1.14661

Moernaut, Ruben, Jan Mast, Martina Temerman, and Marcel Broersma. 2020. "Hot Weather, Hot Topic. Polarization and Sceptical Framing in the Climate Debate on Twitter." *Information, Communication and Society* 1–20. https://doi.org/10.1080/1369118X.2020.1834600

Mustafaraj, Eni, Samantha Finn, Carolyn Whitlock, and Panagiotis Takis Metaxas. 2011. "Vocal Minority versus Silent Majority: Discovering the Opinions of the Long Tail." Proceedings of the 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, 103–10. https://doi.org/10.1109/PASSAT/SocialCom.2011.188

Mutz, Diana C., and Paul S. Martin. 2001. "Facilitating Communication across Lines of Political Difference: The Role of Mass Media." *American Political Science Review* 95(1):97–114. https://doi.org/10.1017/S0003055401000223

Neal, Zachary P. 2020. "A Sign of the Times? Weak and Strong Polarization in the U.S. Congress, 1973–2016." *Social Networks* 60:103–112. https://doi.org/10.1016/j.socnet.2018.07.007

Nikolov, Dimitar, Diego F.M. Oliveira, Alessandro Flammini, and Filippo Menczer. 2015. "Measuring Online Social Bubbles." *PeerJ Computer Science* 2015(12):1–14. https://doi.org/10.7717/peerj-cs.38

Pariser, Eli. 2011. The Filter Bubble: What the Internet Is Hiding from You. London: Penguin UK. https://doi.org/10.3139/9783446431164

Prior, Markus. 2007. Post-Broadcast Democracy: How Media Choice Increases Inequality in Political Involvement and Polarizes Elections. New York: Cambridge University Press. https://doi.org/10.1017/CBO9781139878425

Quattrociocchi, Walter, Antonio Scala, and Cass R. Sunstein. 2016. "Echo Chambers on Facebook." arXiv:1411.2893. https://doi.org/10.2139/ssrn.2795110

Recuero, Raquel, Gabriela Zago, and Fabricio Soares. 2019. "Using Social Network Analysis and Social Capital to Identify User Roles on Polarized Political Conversations on Twitter." Social Media + *Society* 5(2):205630511984874. https://doi.org/10.1177/2056305119848745

Reichardt, Jörg, and Stefan Bornholdt. 2006. "Statistical Mechanics of Community Detection." *Physical Review E* 74:1–14. https://doi.org/10.1103/PhysRevE.74.016110

Rodenberg, Jeroen, and Pieter Wagenaar. 2016. "Essentializing 'Black Pete': Competing Narratives Surrounding the Sinterklaas Tradition in the Netherlands." *International Journal of Heritage Studies* 22(9):716–28. https://doi.org/10.1080/13527258.2016.1193039

Sadilek, Martin, Peter Klimek, and Stefan Thurner. 2018. "Asocial Balance: How Your Friends Determine Your Enemies. Understanding the Co-Evolution of Friendship and Enmity Interactions in a Virtual World." *Journal of Computational Social Science* 1(1):227–239. https://doi.org/10.1007/s42001-017-0010-9

Schmidt, Anna Lisa, Fabiana Zollo, Antonio Scala, Cornelia Betsch, and Walter Quattrociocchi. 2018. "Polarization of the Vaccination Debate on Facebook." *Vaccine* 36(25):3606–3612. https://doi.org/10.1016/j.vaccine.2018.05.040

Simmel, Georg. 1904a. "The Sociology of Conflict. I." *American Journal of Sociology* 9(4):490–525. https://doi.org/10.1086/211234

Simmel, Georg. 1904b. "The Sociology of Conflict. II." *American Journal of Sociology* 9(5):672–89. https://doi.org/10.1086/211248

Simmel, Georg. 1904c. "The Sociology of Conflict. III." *American Journal of Sociology* 9(6):798–811. https://doi.org/10.1086/211272

Slater, Michael D. 2007. "Reinforcing Spirals: The Mutual Influence of Media Selectivity and Media Effects and Their Impact on Individual Behavior and Social Identity." *Communication Theory* 17(3):281–303. https://doi.org/10.1111/j.1468-2885.2007.00296.x

Soares, Fabrício Benevenuto, Raquel Recuero, and Gabriela Zago. 2019. "Asymmetric Polarization on Twitter and the 2018 Brazilian Presidential Elections." Proceedings of the International Conference on Social Media and Society (SMSociety), 67–76. https://doi.org/10.1145/3328529.3328546

de Stefano, Domenico, and Francesco Santelli. 2019. "Combining Sentiment Analysis and Social Network Analysis to Explore Twitter Opinion Spreading." Proceedings of the 2019 28th International Conference on Computer Communications and Networks (ICCCN), 1–6. https://doi.org/10.1109/ICCCN.2019.8846911

Stroud, Natalie Jomini. 2010. "Polarization and Partisan Selective Exposure." *Journal of Communication* 60(3):556–76. https://doi.org/10.1111/j.1460-2466.2010.01497.x

Sunstein, Cass R. 1999. "The Law of Group Polarization." John M. Olin Program in Law and Economics Working Paper No. 91.

Sunstein, Cass R. 2001a. Designing Democracy: What Constitutions Do. Oxford: Oxford University Press. https://doi.org/10.2139/ssrn.199668

Sunstein, Cass R. 2001b. *Echo Chambers: Bush v. Gore, Impeachment, and Beyond*. Princeton: Princeton University Press. https://doi.org/10.1093/oso/9780195145427.001.0001

Sunstein, Cass R., and Adrian Vermeule. 2008. "Conspiracy Theories." *Public Law and Legal Theory Research Paper Series No. 199 and No. 387*. https://doi.org/10.2139/ssrn.1084585

Tang, Jiliang, Yi Chang, Charu Aggarwal, and Huan Liu. 2016. "A Survey of Signed Network Mining in Social Media." *ACM Computing Surveys* 49(3):1–37. https://doi.org/10.1145/2956185

Törnberg, Petter. 2018. "Echo Chambers and Viral Misinformation: Modeling Fake News as Complex Contagion." *PLOS ONE* 13(9):e0203958. https://doi.org/10.1371/journal.pone.0203958

Törnberg, Petter. 2022. "How Digital Media Drive Affective Polarization through Partisan Sorting." *Proceedings of the National Academy of Sciences* 119(42):e2207159119. https://doi.org/10.1073/pnas.2207159119

Törnberg, Petter, Claes Andersson, Kristian Lindgren, and Sven Banisch. 2021. "Modeling the Emergence of Affective Polarization in the Social Media Society." *PLOS ONE* 16(10):e0258259. https://doi.org/10.1371/journal.pone.0258259

Traag, Vincent A., and Jeroen Bruggeman. 2009. "Community Detection in Networks with Positive and Negative Links." *Physical Review E* 80(3):036115. https://doi.org/10.1103/PhysRevE.80.036115

Traag, Vincent A., Paul van Dooren, and Yurii Nesterov. 2011. "Narrow Scope for Resolution-Limit-Free Community Detection." *Physical Review E* 84(1):1–9. https://doi.org/10.1103/PhysRevE.84.016114

Traag, Vincent A., Ludo Waltman, and Nees Jan van Eck. 2019. "From Louvain to Leiden: Guaranteeing Well-Connected Communities." *Scientific Reports* 9(1):1–12. https://doi.org/10.1038/s41598-019-41695-z

Uitermark, Justus, Vincent A. Traag, and Jeroen Bruggeman. 2016. "Dissecting Discursive Contention: A Relational Analysis of the Dutch Debate on Minority Integration, 1990–2006." *Social Networks* 47:107–15. https://doi.org/10.1016/j.socnet.2016.05.006

Vaccari, Cristian, Augusto Valeriani, Pablo Barberá, John T. Jost, Jonathan Nagler, and Joshua A. Tucker. 2016. "Of Echo Chambers and Contrarian Clubs: Exposure to Political Disagreement among German and Italian Users of Twitter." *Social Media and Society* 2(3):1–24. https://doi.org/10.1177/2056305116664221

del Vicario, Michela, Gianna Vivaldo, Alessandro Bessi, Fabiana Zollo, Antonio Scala, Guido Caldarelli, et al. 2016. "Echo Chambers: Emotional Contagion and Group Polarization on Facebook." *Scientific Reports* 6(37825):1–12. https://doi.org/10.1038/srep37825

Vliegenthart, Rens, and Jasper Zuure. 2020. *Doen, Durven of de Waarheid? Democratie in Digitale Tijden*, edited by J. de Ridder, R. Vliegenthart, and J. Zuure. Amsterdam: Amsterdam University Press.

Wekker, Gloria. 2016. White Innocence: Paradoxes of Colonialism and Race. Durham, NC: Duke University Press. https://doi.org/10.1515/9780822374565

Williams, Hywel T.P., James R. McMurray, Tim Kurz, and F. Hugo Lambert. 2015. "Network Analysis Reveals Open Forums and Echo Chambers in Social Media Discussions of Climate Change." *Global Environmental Change* 32:126–38. https://doi.org/10.1016/j.gloenvcha.2015.03.006

Wu, Siqi, and Paul Resnick. 2021. "Cross-Partisan Discussions on YouTube: Conservatives Talk to Liberals but Liberals Don't Talk to Conservatives." Proceedings of the International AAAI Conference on Web and Social Media.

Yardi, Sarita, and danah boyd. 2010. "Dynamic Debates: An Analysis of Group Polarization over Time on Twitter." *Bulletin of Science, Technology and Society* 30(5):316–27. https://doi.org/10.1177/0270467610380011

Yoon, Han Woo, and Han Woo Park. 2014. "Strategies Affecting Twitter-Based Networking Pattern of South Korean Politicians: Social Network Analysis and Exponential Random Graph Model." *Quality and Quantity* 48(1):409–23. https://doi.org/10.1007/s11135-012-9777-1

Zheng, Xin, Daniel Zeng, and Fei-Yue Wang. 2015. "Social Balance in Signed Networks." *Information Systems Frontiers* 17(5):1077–95. https://doi.org/10.1007/s10796-014-9483-8

Financial Disclosure: This research has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 732942, project ODYCCEUS.

Anna Keuchenius: Sociology, University of Amsterdam, E-mail: anna@keuchenius.com.
Petter Törnberg: Computational Social Science, University of Amsterdam, E-mail: p.tornberg@uva.nl
Justus Uitermark: Human Geography Planning and International Development Studies, University of Amsterdam, E-mail: j.l.uitermark@uva.nl