Garland et al. *EPJ Data Science* (2022) 11:3
https://doi.org/10.1140/epjds/s13688-021-00314-6

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Impact and dynamics of hate and counter speech online

Joshua Garland[1*†], Keyan Ghazi-Zahedi[2†], Jean-Gabriel Young[3], Laurent Hébert-Dufresne[3] and Mirta Galesic[1,4,5]

*Correspondence:
joshua@santafe.edu
[1] Santa Fe Institute, 1399 Hyde Park Road, 87501, Santa Fe, NM, USA
Full list of author information is available at the end of the article
†Equal contributors

Abstract

Citizen-generated counter speech is a promising way to fight hate speech and promote peaceful, non-polarized discourse. However, there is a lack of large-scale longitudinal studies of its effectiveness for reducing hate speech. To this end, we perform an exploratory analysis of the effectiveness of counter speech using several different macro- and micro-level measures to analyze 180,000 political conversations that took place on German Twitter over four years. We report on the dynamic interactions of hate and counter speech over time and provide insights into whether, as in 'classic' bullying situations, organized efforts are more effective than independent individuals in steering online discourse. Taken together, our results build a multifaceted picture of the dynamics of hate and counter speech online. While we make no causal claims due to the complexity of discourse dynamics, our findings suggest that organized hate speech is associated with changes in public discourse and that counter speech—especially when organized—may help curb hateful rhetoric in online discourse.

Keywords: Hate speech; Counter speech; Twitter; Natural language processing; Time-series analysis

1 Introduction

Hate speech is rampant on many online platforms and manifests in many different forms, e.g., insulting or intimidating, encouraging exclusion, segregation, and calls for violence, as well as spreading harmful stereotypes and disinformation about a group of individuals based on their race, ethnicity, gender, creed, religion, or political beliefs [1–10]. While it is widely accepted that hate speech is a growing problem on online platforms, what to do about it is a point of contention. One proposal is to more or less automatically detect and remove hateful content. This approach would have to overcome several challenges, however, including the nuanced and constantly evolving nature of hate speech, societal and legal norms about free speech, and the possibility of merely moving hate to other platforms rather than eliminating it [11].

An emerging alternative is citizen-driven *counter speech*, whereby users of online platforms themselves generate responses to hateful content in order to stop it, reduce its consequences, discourage it, as well as to support the victim and fellow counter speakers [12–

15], and ultimately increase civility and deliberation quality of online discussions [16, 17]. Like hate, counter speech can take many forms, including providing facts, pointing to logical inconsistencies in hateful messages, attacking the perpetrators, supporting the victims, spreading neutral messages, or flooding a discussion with unrelated content [18–26].

The effectiveness of counter-speech in curbing the spread of hatred online is not well understood. In fact, until recently, there had been a lack of longitudinal, large-scale studies of counter speech [3, 27]. One reason had been the difficulty of designing automated algorithms for discovering counter speech in large online corpora. Past studies have provided insightful analyses of the effectiveness of counter speech, but were limited in scope as they relied on relatively small, hand-coded examples of discourse [16, 25, 28–31]. Recently, however, two self-labeling groups engaged in organized online hate and counter speech on German twitter. One, called Reconquista Germanica (RG), aimed to spread hate and disinformation about immigrants, while the other, called Reconquista Internet (RI), tried to actively resist this discourse using organized counter speech (see Sect. 2.1 for more details and [25] for a qualitative analysis and in-depth description of the two groups).

In previous work, we used the existence of these two groups to develop an automated classifier that could identify typical speech from these groups on a very large textual corpus [32]—orders of magnitude bigger than any other study of online hate and counter speech to date. This classification system achieved high accuracy scores on out-of-sample data and was also verified against human judgment. Here, we use this classification system to perform a longitudinal study on more than 180,000 fully-resolved (out-of-sample) Twitter conversations that took place from 2015 to 2018 within German political discourse to study the *dynamics* of hate and counter speech and to gain descriptive insight into the potential effectiveness of counter speech.

We anticipate that both organized counter and hate speech will be more effective than independent individual efforts. We develop this theoretical expectation based on two relevant lines of literature. One is the literature on bullying, a phenomenon that shares some of its causes and manifestations with hate speech [2]. The presence of peers is important for both bullying and bully-opposing behaviors. Bullies often seek an approving crowd, because crowd's attention justifies their behavior and helps stifle resistance, while also enabling them to achieve visibility and social status they seek [33]. Bystanders, as well, often look to others when deciding whether to actively oppose the bullying and help the victim. The presence of such peers increases the chance that an opposing individual will receive peer support and protection [34, 35]. These observations translate to the world of online hate and counter speech, where individuals who engage in either kind of speech can become targets of online hate themselves [26], and may even be threatened by physical violence. Seeing that one is a lone countering voice in a flood of similar messages can discourage individuals from engaging in opposing speech, as the effort put in exposing one's own view might seem futile [26].

The other line of literature supporting the expectation that organized efforts will be more effective is the research on social norms. Numerous studies have demonstrated that perceived beliefs and behaviors of others influence people's own political attitudes [36–38], university evaluations [39], pro-environmental behaviors [40], financial decisions [41], voting behavior [42], food choice [43], and health-related behaviors [44]. Similarly, the presence of opposing voices in online discussions can be viewed as a signal of how widespread these views are in the overall population [45, 46]. These perceptions can guide people's reactions to and acceptance of hate and counter discourse. When opposition is organized so that proponents of a particular view come together to post comments in the same discussion (a behavior that is one of the basic operating principles for both RG and RI), this view will become more visible to others. The resulting change in descriptive norm can encourage others with similar views to become more vocal about their positions, further reinforcing the norm. On the other hand, this heightened visibility of a specific discourse (hate or counter) can also mobilize the opposing group to post more countering comments [26].

While we cannot make causal claims about the impact these groups may have had on the broader German society or vice versa, our data provides a unique opportunity to perform an exploratory analysis of the interplay of organized hate and counter speech in a longitudinal online setting.

2 Background, data and methods

2.1 Background on reconquista germanica (RG) and reconquista Internet (RI)

Here we provide a brief overview of the two focal groups involved in our case study. For a curious reader we recommend Ref. [25] for an in-depth analysis of both groups. Reconquista Germanica (RG) is a highly organized far-right troll army which officially formed two months prior to the German federal election with the purpose to influence the election to benefit Germany's radical-right party the Alternative für Deutschland (AfD) in 2017. However, recruitment for RG and related hate groups started earlier in late 2016 [47]. RG was organized hierarchically, and prospects had to prove their political views in a series of online interviews, before they were granted access to the Discord server. Once admitted, people could climb the ranks by proving themselves in coordinated actions, which were given in daily orders to the entire group. Actions were then coordinated between multiple social media outlets and each member was usually responsible for multiple accounts. It is unclear how many active members RG had and how many accounts they controlled, but at its peak, RG might have had around 5000 active members [25].

Reconquista Internet (RI) was established in late April 2018 shortly after the exposure of RG by a group of journalists [48, 49], upon the encouragement of German TV host Jan Böhmermann during the popular German satire news show Neo Magazin Royal. While RI was originally a satirical response to RG it quickly grew to be the largest counter speech group to date in Germany [25]. RI was fully functional by May 2018. RI's goal were to spread positive messages, restore civil discourse and to directly counter the hate being spread by RG with "love and reason [50]." Like RG, RI was organized into groups which formed around specific social media platform, i.e., there was a group for Twitch, Facebook, Twitter, and so on [50]. Each group had two leaders, who organized activities within their group. To communicate with one another, call for help etc., several Discord channels were maintained, e.g., a channel for sharing links to tweets which were selected for engagement, under attack etc. At its peak around May 10, 2018, RI had 62,000 registered users on its Discord server. This number quickly decreased to 4–5000 active members during the first few months, followed by further splintering into smaller independent groups by the end of July 2018. The details presented here about the internal workings, organization and strategies of RI are a synthesis of private correspondences with active members of RI, as well as Reconquista Internet's Codex [50] and finally the information presented about RI in Ref. [25].

One difference between these two groups was their organizational structure and tactics. As mentioned previously, RG, for example, would take part in highly-coordinated cyber-hate attacks where they were instructed who to attack and how. With instructions [48, 49] like (translated from German): *Young women who came straight from university make classic victims...Offend. And then draw every register. Do not leave anything out. Weak point are often the family. Always have a repertoire of insults that you can adapt to the respective opponent...As soon as you see them (media/politicians) spilling their lies and their poison into the world again, tell them our opinion, engage them in discussions, mark their lies as #fakenews and troll the **** out of them.* In contrast, RI was more loosely organized [25]. Instead of giving direct orders of who/what to respond to and how, RI responded to calls for protection via Discord channels, and instead of being told what to say, RI followed a few general principles when responding, outlined in their codex [50], e.g, *Human dignity is inviolable. We are not against 'them.' We want to solve problems together and act guided by mutual respect, love, and reason, We do not wage war but seek conversation.*

While both RG and RI were active on several social media platforms, for our analysis, we focused on a sample of their Twitter presence, focused around several prominent German news organizations and public figures which enabled us to study the structure of the resulting conversations and the dynamic interplay between the two groups over time.

2.2 Data

For the analysis reported in this manuscript we performed two independent data collection phases:

1. *Classifier Training Data Collection:* We collected millions of Tweets originating from approximately 3700 known RG and RI members to train a classification system to identify hate and counter speech typical of these groups, as well as neutral speech not typical of either group. This data is described in Sect. 2.2.1 and the classification system as well as its accuracy and verification is described in [32] and Sect. 2.3.
2. *Reply Tree Data Collection:* We collected a longitudinal sample of German political conversations (or "reply trees") over a 4-year period during which RG and/or RI were active. For this we specifically targeted root accounts (initiators of conversations) where organized activity regularly occurred by both RG and RI. We collected trees before, during and after these groups were active, to study if there were quantifiable differences in overall political discourse. Details about the collection of this data and choices that were made are discussed in Sect. 2.2.2.

Table 1 summarizes the two resulting data sets. The following two sections provide details on how we went about collecting this data.

2.2.1 Classifier training data collection

To train our classification algorithm, we collected a balanced corpus of more than 9 million relevant tweets originating from known RG accounts (4,689,294 tweets) or RI accounts (4,323,881 tweets). For a full discussion of how we identified these members see [32]. But briefly, for RG members we used the list of hate accounts known as the "Böhmermann Liste," which was promoted as a list of accounts that spread hateful rhetoric, promote radical-right propaganda, or engage in hateful speech. This was a list compiled by investigative journalists [49], promoted by Jan Böhmermann and verified by several RI members. For RI we began with a hand-curated list of 103 accounts comprised of the

Table 1 A summary of all data that was used in this manuscript. The top table summarizes the tweets that were used in [32] to train the classifier used in this manuscript. This is broken into two categories: Reconquista Germanica and Reconquista Internet and summarizes how many members were identified from each group and how many tweets were collected from those respective groups. The top table summarizes the Reply Tree Data, including the total number of trees, the total tweets among all trees, the total number of accounts which served as root nodes and the time frame when these reply trees occurred

Summary of Classifier Training Data		
Category	Members Identified	Total Tweets Collected
Reconquista Germanica (Hate)	2120	4,689,294
Reconquista Internet (Counter)	1575	4,323,881

Summary of Reply Tree Data				
Category	Total Trees	Total Tweets in Trees	Total Root Accounts	Time Frame
Complete	203,711	1,649,137	9933	Jan. 2013–Dec. 2018
Filtered	181,370	1,222,240	23	Jan. 2015–Dec. 2018

core RI Twitter team which we received from RI directly. We then expanded this set by looking at the ego network of these 103 core RI members using the Twitter API and then only included users that appeared in at least 5 of the ego networks of core RI members. This list was then further reduced by analyzing their screen names and bios for known RI badges (see [32] for more details). Tweets that came from RG accounts were labeled as "hate speech" and tweets that came from RI accounts were labeled as "counter speech." This data allowed us to build a classifier to identify speech similar to that of RG and RI members. We discuss the ramifications of this in more detail in Sect. 2.3 and we also point the reader to [32] for a complete description on exactly how these tweets were selected, processed and used.

2.2.2 *Reply tree collection*

To perform an exploratory analysis of the potential impacts of organized hate and counter speech on German political discourse over time it was necessary to collect a longitudinal dataset of German political conversations. This collection involved three main challenges: selecting appropriate conversation initiators ("root accounts") to monitor over time, selecting an appropriate time frame to monitor the root accounts and finally collecting as many of the reply trees as possible for each root account across the desired time frame.

The first challenge was to identify a consistent source of conversations that would collectively represent a balanced sample of German political discourse during our time frame of study. Additionally, we wanted to ensure these conversations contained activity by both RG and RI. Hence, we identified conversations by selecting root accounts that had a large number of followers, tweeted frequently, existed for an extended and continuous period of time, and whose tweets and subsequent conversations were political in nature. Perhaps most importantly, we selected root accounts that we knew, from private correspondences with RI, were often targeted by RG. As a starting point, we focused our data collection efforts on conversations that grew in response to @tagesschau. We chose to start with Tagesschau because it is widely considered as a reliable and balanced news source [51], has existed for a long time, and has many followers on Twitter. As our study continued, we expanded the list of root accounts to include more accounts where RG/RI activity was taking place regularly. These included additional mainstream news outlets, journalists and bloggers who were targeted frequently by RG and finally, politicians and other relevant public figures. Each account added was done so based on advice by members of RI as to where the most relevant organized actives were occurring.

Our next challenge was to select a time frame to study German political discourse which included the time of RG and RI activity but also included enough of a lead up to these organized groups that we could monitor changes in discourse over time. To this end, we chose to start collecting reply trees from January 2013, the point preceding the founding of the Alternative für Deutschland (AfD). We monitored the discussions continuously until the end of 2018, when Twitter made fundamental changes to their website which made it impossible to collect trees in the manner we describe below. As such, our dataset ends in December of 2018. Nevertheless, the time frame from January of 2013 to December of 2018 allowed us to study discourse leading up to the formation of both organized groups, including a long stretch of time where both groups existed and interacted.

Our final challenge was to collect a sample of fully-resolved conversations from these root accounts across our study period. As the original Twitter API did not provide functionality to collect reply trees in their entirety, we developed two types of custom scrapers to collect these reply trees. The first type of scraper identified (at random) a one month contiguous sample of conversations originating from a single root account of interest which was missing from our collection of conversations. For example, these scrapers would randomly select one root account of interest, e.g., @tagesschau, as well as a one month period between January 2013 the present date, e.g., 1st of March 2016 to 31st of March 2016. The scraper would then systematically find all top-level urls[1] for tweets by that account in that time period, and stored these in a file. The second type of scraper would then randomly select a top-level url from this file and manually scrape the resulting conversation in its entirety from the top-level urls' raw html. A log of previously scraped reply trees ensured that each tree was only scraped once. The scraper ran continuously from June to December of 2018. In total, we collected 203,711 conversations (reply trees) that grew in response to tweets of 9,933 root accounts between January of 2013 to December of 2018.

The data gathering process stopped in early 2019 following abrupt changes to Twitter's website at that time. Our system had no explicit bias or preference as to which reply trees were collected at any given time in the collection process. Consequently, accounts that were added to the list later were sampled less often and, conversely, accounts that were added earlier had a larger number of conversations. As this could potentially create bias in the resulting sample, for the analysis reported here we limited the dataset to 181,370 conversations, containing 1,222,240 tweets, which grew in response to tweets from 23 accounts for which we had continuous coverage throughout the period of January 2015 (the beginning of the migrant crisis in Europe) to December 2018.

The final 23 root accounts included mainstream[2] German news organizations (*@diezeit, @derspiegel and @spiegelonline, @faznet, @morgenmagazin, @tagesschau, @tagesthemen, @zdfheute*), well-known journalists and bloggers (*@annewilltalk, @augstein, @dunjahayali, @janboehm, @jkasek, @maischberger, @nicolediekmann*), and politicians (*viz.,*

[1] These urls had the format https://www.twitter.com/<screen_name>/status/<tweet_id>, where <screen_name> and <tweet_id> refer to the name of the Twitter account and the unique id of the tweet.

[2] At the time of the study these news organizations appeared in lists of the most important news outlets, e.g., [51]. In addition, we learned from private conversations with RI which of these accounts had organized activity from both sides, which helped us narrow the list to the most relevant news organizations.

@cem_oezdemir, @c_lindner, @goeringeckardt, @heikomaas, @olafscholz, @regsprecher, @renatekuenast). Each of these accounts have witnessed organized activity by either RG or RI or both. The number of trees per day from each type of account was stable across the sampled period, see Fig. S1. In analyses reported below, to ensure the type of root account (e.g., news, journalist, politician) did not bias the results, we investigated the sensitivity of all results to statistical controls for the type of root account, and found that they had no discernible effect.

While our team had to go to great lengths to obtain the fully-resolved Twitter conversations used for this analysis, Twitter is now making this kind of analysis easier. With the advent of the second version of the Twitter API as well as the academic product track, academic researchers will now be able to directly obtain fully-resolved historical conversations. However, at the time of submitting this paper, the data that could be extracted from the API is incomplete for most past conversations—and much of the signal of extreme forms of discourse has been deleted due to suspensions, user deleted accounts and user deleted tweets (see Additional file 1, Section S1.1 for more details).

2.3 Classification of hate, counter and neutral speech

To study how RG and RI may have impacted German political discourse in our reply tree dataset it was necessary to first build a classification system which could recognize the hateful rhetoric of RG, the counter speech of RI as well as neutral political discourse. In [32] we built such a classification system using the data described in Sect. 2.2.1. The focus of the present work is applying that system to the dataset described in Sect. 2.2.2, but we describe that classification system for completeness. In addition, we discuss the results verifying that this classifier performed as expected e.g., agreeing with human intuition.

The classification pipeline we used to classify tweets in reply trees consisted of two stages [52]: extraction of features from text and classification using those features. To extract the features, we pre-processed the data and then constructed paragraph embeddings, also known as doc2vec models [53], using the standard gensim implementation [54]. We performed a parameter sweep following standard practice and the guidelines of [55]. This sweep included the analysis of several doc2vec parameters e.g., maximum distance between current and predicted words, "distributed-memory" vs. "distributed bag of words" frameworks, three levels of stop word removal and five different document label types (see [32] for more details). Each version of the doc2vec model was trained on five different but partially overlapping training sets. Each training set included 500,000 randomly selected tweets originating from RG accounts and another 500,000 coming from RI accounts.

These trained doc2vec models allowed us to extract features, i.e., infer a feature vector, from a given tweet. To classify each tweet as either hate or counter we then coupled each doc2vec model with a regularized logistic regression classifier, as implemented by scikit-learn [56]. These logistic regression functions were trained on the same training set as the corresponding doc2vec model.

Note that the doc2vec models and logistic regression functions used only the tweet content. Features such as, badges, screen names etc. were not used in this phase explicitly. However, the initial labeling of the accounts did consider these features, so one could view RG and RI badges as secondary features used by the classification scheme.

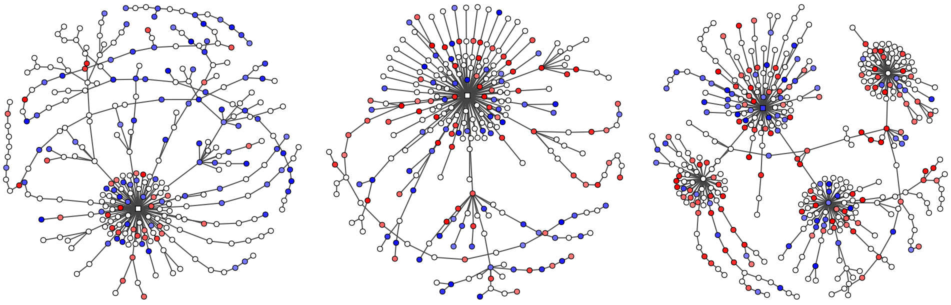

Figure 1 Examples of reply trees (Twitter conversations). Nodes are tweets and edges denote a "replied to" relationship. Node colors denote whether the content of a tweet was hate (red), counter (blue) or neutral speech (white), with intensity of colors representing the magnitude of the hate score according to the classifiers presented in [32]

By pairing a trained doc2vec model and a logistic regression function, we were able to assign a probability $p_h(t)$ to each tweet that it belongs to either the hate or counter class. We then recoded this probability to a hate score $S_h(t)$ ranging from –1 to 1, using

$$S_h(t) = 2p_h(t) - 1, \tag{1}$$

where negative values mean that a tweet is similar to prototypical RI counter speech and positive values mean that it is similar to prototypical RG hate speech. For the results reported here we used an ensemble learning approach where 25 independent doc2vec, logistic regression pairings voted on each tweet in a given tree. The tweet's final hate score $S_h(t)$ was then defined as the average hate score assigned to that tweet across all 25 pairings.

For final classification of tweets, we used a confidence voting system with thresholding to assign each tweet a label. For the majority of the analyses reported in this paper (unless stated otherwise), if $S_h(t) \geq 0.4$, t was labeled hate, and if $S_h(t) \leq -0.4$, t was labeled counter. If $-0.4 < S_h(t) < 0.4$ then the tweet was labeled neutral. These scores effectively mean that the ensemble of classifiers was at least 70% confident in labeling a tweet as either hate or counter speech. Figure 1 shows examples of reply trees after the final classification system was applied to them. Here, we colored the nodes (tweets) t in each tree red for hate, blue for counter, and white for neutral based on $S_h(t)$ using the thresholds just mentioned.

Depending on the threshold being used, this classification system achieved F1 scores ranging from 0.762 to 0.97 on balanced test sets containing equal proportion of hate and counter speech. See Table 2 for a full summary of classification performance metrics. For the threshold used in this paper $\gamma = 0.40$ the classification pipeline achieved F1 scores of 0.877. This accuracy exceeds previous studies that used smaller unbalanced data sets and achieved F1 scores ranging from 0.49 to 0.77 [28, 29, 57].

While these F1 scores are impressive compared to other similar studies, a careful reader will notice that our classifier is not actually directly classifying hate and counter speech, but that instead we are using *proxies* for hate and counter speech. Indeed, as we have mentioned in the Introduction, classifying hate speech remains a difficult challenge to this day. Instead, because of the way this training data was originally labeled, i.e., as an RG or RI tweet, we are actually classifying speech that is typical of RG vs. RI members. In a sense, we are trading off between potentially noisy labels and the scale of the labeled

Table 2 Classification scores for the panel of experts used to label the Reply Tree tweets using the top 25 experts. γ is a confidence threshold on $S_h(t)$, viz., $|S_h(t) = 2p_h(t) - 1| \geq \gamma$. "Percent Labeled" is the percentage of examples in the test set that were labeled as either hate or counter at a confidence level of γ. The top row represents traditional accuracy measures, which compare favorably to previous studies that used smaller unbalanced data sets and achieved F1 scores ranging from 0.49 to 0.77 [32]. The remaining rows are more nuanced as not all examples in the tests sets are being labeled correctly or incorrectly and thus these rows need to be viewed with cautious optimism; see [32] for an in-depth discussion of these issues

Accuracy Scores for Classification System				
γ	Precision	Recall	F1	Percent Labeled
0.0	0.763	0.762	0.762	100%
0.20	0.827	0.827	0.827	76%
0.40	**0.877**	**0.876**	**0.877**	**57%**
0.60	0.917	0.917	0.917	41%
0.80	0.958	0.958	0.958	25%
0.90	0.977	0.977	0.977	15%

training corpus, which comes with some risk. Some of the tweets in the training data set which were labeled as hate or counter might have included some other type of speech. At the same time, using these labels allowed us to conduct analyses on a large scale, capturing a wide breadth of both hate and counter speech patterns. (See Ref. [32] for more details on classification trade offs and the care that was taken in selecting the accounts we used to train the classifiers.)

We next describe three steps we took to investigate the validity of our automated classification system. First, we conducted a qualitative analysis of the speech labeled as hate and counter by our classifier. To this end, we counted the frequency of all tokenized words in the whole corpus, and determined how indicative each word is for tweets labeled as either hate or counter speech [58]. Figure 2 shows the top 100 most frequently used words that are specific to each corpus, i.e., these are words that occur the most frequently in hate/counter speech, but which simultaneously occur very infrequently in counter/hate speech. The typical words used by RG (Fig. 2(A)), which we label as hate, focus on Merkel and variations of the word migrant, immigrant, asylum seekers (their typical targets of attack). The typical words used by RI (Fig. 2(B)) focus on nazis, major neonazi rallies such as Chemnitz and the so-called "radical right" (Rechtsradikal). All of these tokens align well with our notion of hate and counter speech.

Second, three of the authors with a working knowledge of the German language read a large sample of classified tweets to confirm that the classifications made sense. We provide a few example tweets typical of RG (hate speech) and RI members (counter speech). Tweets with high hate scores (typical RG speech) typically discuss purported crimes committed by refugees against Germans, as well as criticize politicians, other public figures, and fellow citizens who are perceived as being too weak or overly welcoming towards refugees and the rising diversity of the German society. Typical examples of these tweets (translated from German) are *"Disgusting: Afghan molested 7 school girls (10–12) Germany's 'cuddling' justice has struck again. A pedophile Afghan that taxpayers have to feed for life was just what was missing in Merkel's 'colorful country'!"* and *"Call for help from the mother of the raped and murdered 14-year-old Susanna to chancellor Merkel. If Susanna and her mother were refugees, Merkel would have reacted. But Merkel did not say a word about any of the many German girls murdered by refugees."*

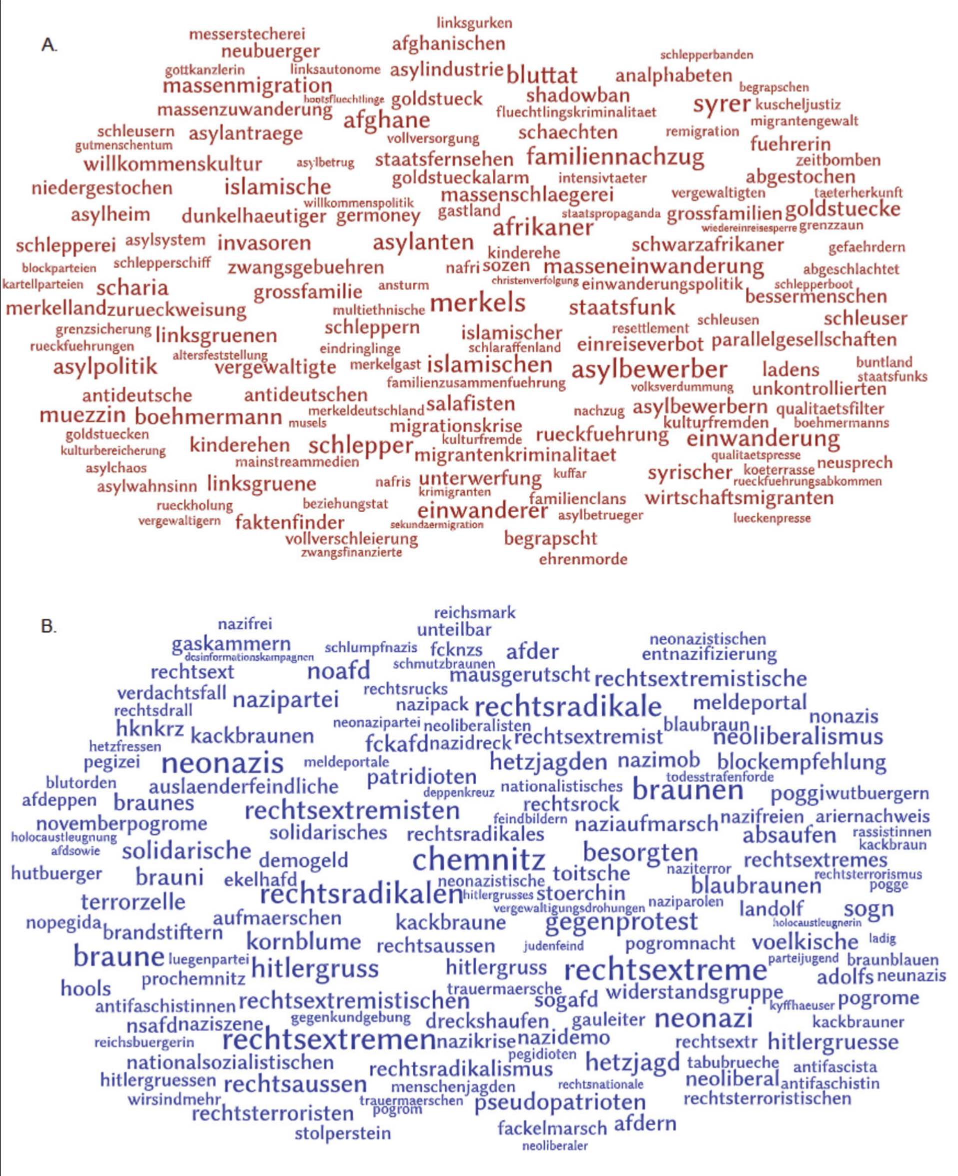

Figure 2 Words that are most frequently used in tweets with high vs. low hate scores. (**A**): most frequent words appearing in hate tweets that were rarely used in counter tweets. (**B**): The most frequent words used in counter tweets that did not appear in hate tweets

Tweets with high counter scores often point to the similarities of the German radical-right movement and the Nazis, and highlights various ways in which radical-right movements influence German society, calling for counter protests. Typical examples of these tweets (translated from German) are *I find it very deceiving when voters of the #AFD claim they are not Nazis. Whoever goes to vote should find out more about that party. AFD voters vote for Nazis and thus a right-wing extremist Nazi party. Distancing is not possible. Nazis thrive on mindless followers/supporters.* and *AfD marches with neo-Nazis and right-wing radicals through Chemnitz, Maaßen, Seehofer and Kretschmer make all migrants a problem and relativize mobs and hunts [targeting refugees]. In this way the hatred creeps from* *the right towards the middle and becomes normal.* Further qualitative analysis of a subset of discussions between RG and RI can be found in [25].

Third, to verify that our potentially noisy classifier did indeed accurately classify hate and counter speech we conducted a crowdsourcing study to check whether the hate scores derived through this automated pipeline corresponded to human judgment. In our study, human judges evaluated some of the same tweets evaluated by the automated system using the content of the tweets and nothing else. We selected 28 German-speaking raters to evaluate 5000 randomly selected tweets evenly spread across the whole range of scores $-1 \leq S_h(t) \leq 1$. Raters ranked tweets on a scale of 1 to 5, from "very likely counter speech" to "very likely hate speech," with 3 corresponding to neutral content. Each tweet was evaluated by at least 2 different raters(see [32] for details). The results suggest that classifier hate scores aligns well with human judgment, with overall correlation between hate scores and human judgments being $r = 0.94$. The correlation was somewhat lower for tweets classified as counter speech ($r = 0.75$) than for those classified as hate ($r = 0.96$). As expected, classifier scores around 0.5 received intermediate hate scores from human judges. It should be noted that the labels assigned by these human classifiers were not used in the labeling process at all. Rather these labelings were only used to validate the classification system's output. The only thing used to label the tweets was the group affiliation of the tweet, i.e., tweeted by an RG member or coming from an RI member and nothing else.

2.4 Statistical analyses

To investigate whether counter speech had an effect on various indicators of discourse dynamics (such as relative prevalence of hate and counter speech over time, impact of hate and counter speech on subsequent discourse, as well as likes of and discussions initiated by tweets containing hate and counter speech) one would ideally like to perform some flavor of statistical causal inference analysis. However, due to a variety of conflating factors, the reply tree time series presented in this paper are difficult to analyze from a causal perspective. The conversations are part of an intricate social system in which it is difficult to define and isolate a reasonable 'control group', rendering analyses such as "difference in differences" [59] inapplicable in this case. Furthermore, because of unquantifiable outside influences, causal inference methods such as Granger Causality [60] or Transfer Entropy [61] are of limited value. The lack of evidence for determinism and stationarity in our datasets made state-space [62] based causal inference, e.g., Convergent Cross Mapping [63], not applicable.

2.4.1 Change point analysis

While we could not perform rigorous causal inference due to fundamental limitations of the system, we nevertheless put effort into conducting a principled study of the changes in this dynamics over time—even if the analysis could only identify correlations and suggest changes in descriptive statistics such as overall trends.

In particular, to investigate changes in the proportions of hate and counter speech over time and in association with different events (Fig. 3), we conducted both exploratory and confirmatory statistical analyses. Exploratory change point analyses were conducted using the *findchangepts algorithm* available in Matlab [64, 65]. The algorithm searched for changes in linear trends, always using a standard deviation of the empirical trend as the threshold value. Many change point algorithms have been developed (see [65] for a review) to identify changes in either some summary statistic or trends, e.g., mean, variance

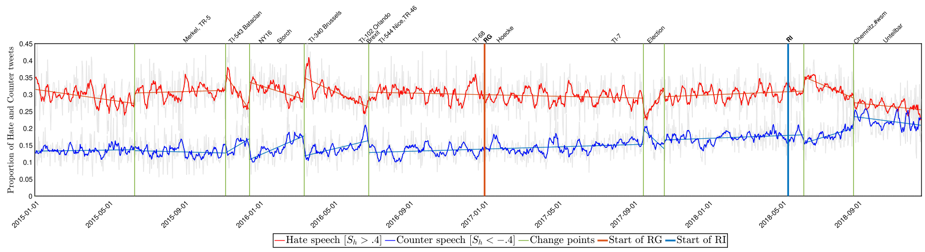

Figure 3 Proportion of hate and counter speech over time. Organized counter speech (RI, blue vertical line) is followed by changes in proportions of hate and counter speech tweeted within 181,370 reply trees to 23 prominent twitter accounts, from January 2015 to December 2018. Each data point is a daily average and trends are smoothed over a one-week window. Exploratory change point analysis [64] identifies the changes in linear trends shown by green vertical lines (see text for further confirmatory statistical tests). Event labels: Merkel = prime minister pronounces 'Wir schaffen das' in support of immigrants, TR(TI)–x = right-wing (Islamic) terrorist attacks where x is the number of dead or injured people (location included when attack was outside Germany), NY16 = New Year's Eve assaults on women in Cologne's main train station, Storch = AfD politician calls for use of arms against immigrants at the boarder, Brexit = referendum date, RG = Reconquista Germanica and associated groups start organizing, Hoecke = AfD politician calls for a change in how Germany remembers the Holocaust, Election = German Federal Election, RI = Reconquista Internet becomes functional, Chemnitz = large anti-immigrant protests, Unteilbar—large counter rally in Berlin

or information production. However, we were interested in studying changes in overall *trends* in discourse which made [64] an ideal approach. For confirmatory analyses, we tested differences in trends before and after occurrence of Reconquista Germanica and Reconquista Internet.

2.4.2 Prediction

To investigate the relationship between the current and future proportion of hate and counter speech, we used a Vector Autoregression (VAR) analysis [66]. For this analysis, we accounted for non-stationarities in each time series in the usual way by taking first differences of the time series shown in Fig. 3. Dickey–Fuller tests [67] showed that the resulting trends were stationary in all three periods. We use these transformed stationary time series for the following VAR analysis.

2.4.3 Longitudinal mixed linear model

To investigate the effectiveness of different types of speech to steer a conversation depending on its hate score (Fig. 6), we corrected the impact of focal tweets using a longitudinal mixed linear model. In this model, focal tweet impact was predicted by its score and a host of other factors that could have affected the impact in addition to the score. The focal tweet's position in the tree, tree size, average hate score of the tree, week, and type of Twitter account that started the reply tree (news outlets, politicians, or journalists/bloggers) were included as fixed effects, and variations by tree and week, tree size were allowed by including them as random effects (see Table S1). These corrections did not change the original results much (see top panel of Fig. S3 for uncorrected results). Note that each square in (Fig. 6) represents the average of all focal tweets with a particular hate score posted in a specific week. Different squares include different number of such tweets, as shown in bottom panel of Fig. S3, and a different number of trees. These and other factors have been statistically accounted for by the model described above. We used similar models to correct all of the other trends shown in the paper. The resulting changes were minimal, so we chose to present the raw data in all other figures.

3 Results

We explore several different macro- and micro-level measures of effectiveness, providing complementary views from different angles. We aim to gain insights into the effectiveness of counter speech by measuring how hate and counter speech interact with each other on macro- and micro-levels. On the *macro-level*, we study changes in the composition of online discourse, reflected in the prevalence of hate, counter, and neutral speech over time. This analysis helps assess the overall deliberation value and civility of discussions [16, 30, 68]. On the *micro-level*, we study interactions between hate and counter speakers, including how they support and reply to each other and how each particular utterance changes the overall subsequent discourse [29].

3.1 Macro-level effectiveness measures

Figure 3 shows a time series of the proportions of counter and hate speech that occurred in our corpus of reply trees over time. Prior to May of 2018, when RI became functional, the relative proportions of both hate and counter speech are largely in a steady state. Roughly 30% of the discourse in our sample of political conversations during this time is hateful and only around 13% was counter speech. However, both of these time series are quite noisy, and there are several clear deviations reflecting the ebbs and flows of political discourse. Many of these deviations coincide with major events such as large-scale terrorist attacks, political rallies or speeches, and elections. After each deviation, however, the proportion of hate and counter speech reverts to the earlier equilibrium, suggesting that these were so-called shock events in an otherwise steady-state system.

After the formation of RI in early spring of 2018, there was a notable change: a decreasing trend in the proportion of hate speech and an increasing trend in the proportion of counter speech. These trends continued until approximately September 2018, at which point they more or less stabilized. That point coincided with large alt-right rallies in Chemnitz, Germany, followed by a large counter rally in Berlin. Around that time the proportion of hate and counter speech stabilized at similar levels, with counter speech rising to approximately 21–22% and hate falling to around 25%. Both proportions continued to decline (although very slightly) through the remainder of the time period studied.

To complement these qualitative observations, we conducted three additional analyses of the interactions of hate and counter speech: an exploratory change point detection to detect changes in trends over time, a confirmatory analysis of trends before and after the occurrence of RG and RI, and a vector autoregressive analysis to study the relationship between present and future proportions of hate and counter speech.

First, the exploratory change point detection helped to identify regions of the time series with similar overall trends in proportion of hate and counter speech (see Methods and Refs. [64, 65] for details). These regions are generally separated by "change points" or important events (e.g., terrorists attacks or political events), which shocked the system and impacted the discourse dynamics (sometimes just temporarily). The automatically identified trends are shown in Fig. 3 as straight, thin red and blue lines, accompanied by green vertical lines which signify the change points selected by the algorithm. This analysis largely confirmed the descriptive analysis above. For large portions of time, e.g., from the summer of 2016 to a month before the German federal election in September of 2017, the trends in hate speech and counter speech were roughly constant. Around the time of the German federal election there is a lot of volatility, with major influxes of counter and hate speech. This reflects the political debate between mainstream and far right political views, with the sharp rise in hate speech leading up to the elections being in line with RG's stated goal of swaying the 2017 election toward the AfD, a so called "radical right" ("rechtsradikal") party. After the election and until the formation of RI there is again a constant trend in the proportion of hate and counter speech. Following the formation of RI, there is a noticeable increase in the proportion of counter speech and decrease in the proportion of hate speech, with these trends stabilizing around the time of Chemnitz and Berlin rallies. The analysis also identifies several other periods with decreasing hate and increasing counter speech. However, these periods were effectively relaxation periods following large shock events such as the Islamic terrorist attacks in Paris and Brussels, which both caused an increase in hate speech and then a relaxation as these events faded from the public focus.

Second, confirmatory statistical analysis helped us to estimate the impact of RG and RI on hate and counter speech trends. We fit linear trends with slope m to time series of 1, 3 and 6 months taken before and after the formation of RG and RI, and analyzed the differences in the before and after slopes. If RG had an effect, we would expect a positive change in slope of hate speech trends, i.e., a positive $\Delta_m = m_{\text{after}} - m_{\text{before}}$ around the formation of RG. We would expect similar results for counter speech trends around the formation of RI. We would also expect negative differences in slopes for hate speech around the formation of RI, but not necessarily vice versa because organized counter speech was not yet developed around the time when RG formed. The difference in slopes for the proportion of hate was indeed positive 1 month after formation of RG ($\Delta_m = 0.003$, $CI = (0.002, 0.004)$), and it remained so when comparing slopes 3 and 6 months before and after RG. However, there was a strong increase and decrease in hate speech in the months before the formation of RG, linked to a large Islamic terrorist attack that may be biasing this analysis. The change in slope Δ_m for counter speech before and after RG was less consistent: Δ_m was slightly positive 1 month after the occurrence of RG ($\Delta_m = 0.0005$, $CI = (-0.0003, 0.001)$) but decreased 3 and 6 months later (-0.00008, $CI = (-0.0002, -0.00002)$). During this time there was no organized counter speech, so this could suggest an initial spontaneous backlash to the emergence of RG that faded after 3 to 6 months.

After the formation of RI, Δ_m for counter speech was indeed positive both when comparing slopes 1 month before and after ($\Delta_m = 0.001$, $CI = (0.0005, 0.002)$) as well as 6 months before and after ($\Delta_m = 0.0003$, $CI = (0.0003, 0.0004)$), although not for 3 months before and after ($\Delta_m = -0.00002$, $CI = (-0.0001, 0.0002)$). When looking at hate proportions around the time of RI, we see a consistently negative Δ_m suggesting that hate declined in proportion around the time RI became organized on 1- ($\Delta_m = -0.0002$), 3- ($\Delta_m = -0.0002$), and 6-month ($\Delta_m = -0.0005$) scales.

Third, we used Vector Autoregression (VAR) models to study the relationship between the current proportion of hate and counter speech and their future values over time during three different periods: i) before the establishment of RG (January 2015–end of 2016), ii) after RG but before RI was established (January 2017–April 2018), and iii) after the establishment of RI (May–December 2018).

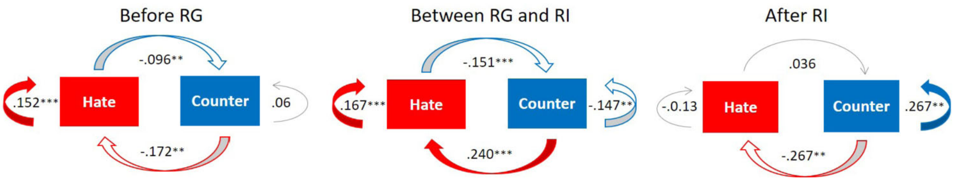

Figure 4 Predictive analysis of hate and counter speech. The relationships between current and future proportions of hate and counter speech for three periods of study. Shown are the results of VAR models for the periods before RG, between RG and RI, and after RI. Red (blue) arrows show statistically reliable effects (** and *** represent $p < 0.01$, and $p < 0.001$ respectively) on hate (counter) speech, and gray show nonreliable effects ($p > 0.05$). Full arrows show positive relationships, and empty arrows negative relationships. Before RG was established, the effects of hate and counter speech were relatively balanced. After the occurrence of RG and before RI, past hate and counter speech became related to more hate speech going forward. After RI was established, past counter speech now negatively predicted future hate and positively predicted future counter speech

Figure 4 shows the relationships between current and future proportions of hate and counter speech for the three time periods of study based on the VAR model coefficients. This analysis suggests statistically significant changes in the day to day relationships between the proportion of hate and counter speech after the establishment of RG and RI. When discussing these results we will refer to a "positive"/"negative" relationship when the discussed quantity is expected to increase/decrease in the future, respectively. It is important to note that this analysis cannot point to causality of these relationships.

Before RG, the VAR model suggests that there was a positive relationship between the proportion of hate speech and itself, i.e., the presence of hate speech today would suggest a higher proportion of hate the next day. However, during this time there was a negative relationship between counter and hate speech, i.e., more counter speech today would mean less hate speech tomorrow—acting as a dampening factor on hate speech. Also during this time, hate had a negative relationship with counter—more hate speech today implied less counter speech the next day. As discussed in the Introduction, both of these trends might have been a consequence of the fact that both hate and counter speech were conducted by isolated individuals, who might have been intimidated by the opposition they received. Finally, during this period, counter speech did not have a significant relationship with itself, i.e., more or less counter speech today did not imply anything about the future values of counter speech. After RG was established but before RI, counter speech no longer seemed to dampen hate speech and instead dampened counter speech. In particular, the presence of counter speech on one day would result in *more* hate speech and less counter speech the subsequent days—perhaps suggesting that hate speakers were organizing to target counter speakers, while counter speakers, who were still acting as individuals, were discouraged by the hate they received. During this time, hate speech also had a positive relationship with itself. All of these relationships changed after RI was established, again as expected theoretically given that counter speakers now had a much stronger peer support. In particular, counter speech led to more counter speech while also dampening hate speech and hate speech no longer had a significant association with either future hate or counter speech.

Taken together, the macro level analysis suggests that the emergence of the organized counter speech group RI might have pushed the German political discourse on Twitter into a new state, with a more balanced presence of both hate and counter speech.

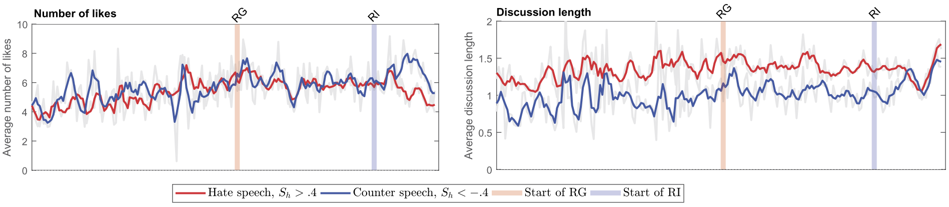

Figure 5 Impressions of hate and counter speech. Impact of hate and counter speech messages over time as quantified by the average number of likes and length of conversation they initiate. The emergence of organized counter speech (RI, blue vertical line). Results are for 181,370 reply trees from January 2015 to December 2018. Each data point is a week average and trends are smoothed over a month-long window. The timeline on the *x*-axis is the same as in other figures but was omitted for space, except for markers of the emergence of RG and RI

3.2 Micro-level effectiveness measures

However, the effectiveness of this counter speech group cannot be studied in a societal vacuum. The time following the formation of RI was characterized by large political rallies and extensive discussions on both sides of the political spectrum, and it is not possible to make causal inferences about the effects of either organized hate or counter groups on Twitter. Therefore, we continue with micro level measures of effectiveness to understand the associated changes on a more nuanced descriptive level.

We begin by analyzing the support that hate and counter speech receives through likes, subsequent replies, and retweets. We then investigate how each hate and counter tweet steers the subsequent discussion in reply trees towards more of the same or towards opposing speech.

3.2.1 Showing support

At the most basic level, participants in Twitter conversations can show their support (or lack thereof) for tweets by affording them their likes and engaging in discussions with them. Engaging in discussion however is not necessarily a sign of support and could be voicing dissent. Figure 5 shows that throughout most of the studied period hate and counter speech received a similar number of likes, but that hate tweets tended to attract longer discussions (measured as the total number of messages in the subtree that a tweet initiated). However, after the emergence of organized counter speech, counter tweets started receiving more likes and attracting longer discussions. The emergence of organized hate speech was not associated with similar changes.

3.2.2 Steering a conversation

A more fine-grained measure of effectiveness is whether individual tweets can steer the subsequent conversation at the level of a single discussion item. We calculate the change in average discourse after a tweet is posted in a reply tree, and measure the impact I of a tweet as the difference between the average hate score S_h (defined in Eq. (1)) of all tweets following and preceding it. More formally, let t_i be the ith tweet that occurred in a tree with N total tweets. Then the impact, or the ability for a tweet t_i to steer a conversations direction, is defined as:

$$I(t_i) = \frac{1}{i-2}\sum_{j=1}^{i-1} S_h(t_j) - \frac{1}{N-(i+1)}\sum_{j=i+1}^{N} S_h(t_j), \tag{2}$$

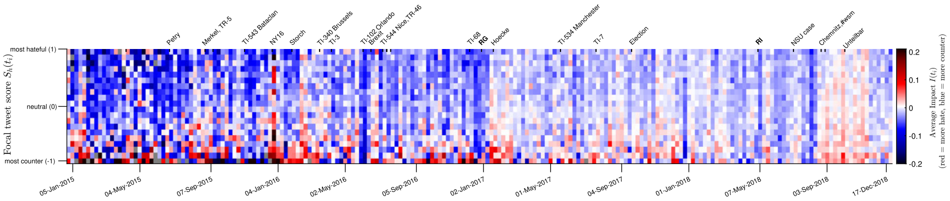

Figure 6 Impact of hate and counter focal tweets on subsequent discourse over time. Each colored square represents the average difference between tweets following and preceding a focal tweet posted in a reply tree in a particular week (Eq. (2)). Results are computed for 82,132 reply trees with at least 3 tweets, over time (horizontal axis), and are binned by hate-score of the focal tweet (vertical axis). Red (blue) squares signify that the focal tweet was followed by a change in the subsequent conversation towards more hateful (counter) speech, with color saturation corresponding to the size of the difference. Event labels are the same as in Fig. 3

where $S_h(t_j)$ is the hate score associated with the jth tweet, t_j. We may refer to t_i as a "focal" tweet when convenient as it is the focus of the computation. We correct the impacts for factors that could have affected it, e.g., discussion length or average hate-score, using a longitudinal mixed linear model (see Methods and Table S1 for details).

Figure 6 shows the results of this analysis. Time is binned into one week segments and shown on the horizontal axis, while the impact of focal tweets are shown on the vertical axis and binned based on their score, going from -1 (most counter) to 1 (most hate). Each square represents the average impact $I(t_i)$ of all tweets occurring in that week with $S_h(t_i) \in [x - 0.05, x + 0.05]$. This plot allows us to study the average ability for tweets with a given hate score to steer conversations in that week. For this analysis, we needed to ensure that a full-blown conversation occurred, so we restricted the sample of reply trees to 82,132 trees that contained at least three tweets (the initial or root tweet, and at least two replies), for a total of 943,822 tweets.

The results shown in Fig. 6 suggest that, in general, hateful tweets were followed by counter speech no matter of their hate scores. Counter tweets tended to be followed by hate speech if they had very low hate scores, but by more counter speech when they had moderately low scores. Analyses using the longitudinal mixed linear model mentioned above further suggested that tweets near the end of a tree were especially likely to attract more opposition (Table S1). Tweets in larger trees, and in trees that were already biased towards a particular kind of speech, tended to receive more support.

Throughout most of the studied period, political conversations tended to include somewhat more hate than counter speech (Fig. 3), but overall contained a large amount of neutral speech (average hate score across all trees was $S_h = 0.14 \pm 0.048$). However, the large blue areas in Fig. 6 indicate that as conversations progressed, the average scores of the conversations shifted towards more neutral or even the counter side of the speech spectrum. There were only a few exceptional weeks in which discussions tended to be steered toward hate no matter the score of the focal tweet. An example is the week of 2016's New Year's Eve, when numerous sexual assaults on women in Cologne's main train station were blamed on Syrian and North African refugees and asylum seekers.

There was a notable difference in discourse dynamics from January 2017 onward, as seen by the transition visible around that time in Fig. 6. This transition corresponds to the time when RG began organizing. Unlike before, hateful conversations in this period did not drift towards counter speech, but stayed neutral or even drifted towards more hateful discourse. Of note, this transition cannot be explained by the classifier picking up on speech not present before 2017 because, as can be seen in Fig. 3, the overall proportion of hate speech, as identified by the classifier, is stable around that time. Training data selection bias is also unlikely to have played a role, as shown by Fig. S2. The shift in dynamics is also not explained by any known change in the Twitter interface or algorithms.[3]

Another transition occurred around the time when RI started to organize in the Spring of 2018. Counter speech started to again dampen hate and pull conversations away from hateful rhetoric. This period was followed by several months of backlash during which almost all tweets were associated with more subsequent hate. The backlash likely reflected the broader societal situation in the Summer of 2018 when large alt-right rallies took place throughout Germany, e.g., in Chemnitz. Finally, the Fall of 2018 was characterized by a return of a more effective counter speech, likely bolstered by large counter rallies occurring in October of 2018, such as "Unteilbar" in Berlin.

3.2.3 Unpacking the pivots

To understand the dynamics unfolding in Fig. 6 in more detail, we unpack how focal tweets pivot the conversations towards either type of speech. We study four possible reactions to each focal tweet: the subsequent discourse can appear to be *supporting* the focal tweet (i.e., counter focal tweet is followed by more counter speech, and vice versa for hate), *opposing* the focal tweet (counter focal tweet is followed by more hate speech), *polarizing* whereby previously neutral discourse becomes more similar to hate or counter speech after the focal tweet, or *ignoring* the focal tweet (discourse remains unchanged before and after the focal tweet). This analysis also reveals the reasons for changes in macro level indicators of hate and counter speech shown in Fig. 3.

Figure 7 shows the relative proportion of different types of reactions that occurred after hate, counter, or neutral focal tweets. In general, hate focal tweets were more often followed by supporting, opposing, or ignoring reactions than counter speech. These trends were relatively stable but could be temporarily altered by notable events, such as the attack on an Orlando night club in June 2016, when counter speech received a burst of support. These results suggest that the notable change in discourse dynamics from January 2017 onward, visible in Fig. 6, occurred primarily because neutral speech became more polarized in the direction of hate speech around that time (panel (c)). In other words, neutral reply trees started to be steered towards more hateful speech. In addition, hate speech became a bit more supported around that time (panel (a)).

This analysis also sheds light on the reasons for relative decrease in frequency of hate speech after the emergence of RI (Fig. 3), as well as for the fact that conversations were more often steered towards counter speech around that time (Fig. 6). Figure 7 suggests that this occurred because of an increase in support for counter tweets (panel (a)) and, for a short while, an increased polarization of neutral speech towards counter speech (panel (c)). However, as suggested in (Fig. 6), a period of hate speech backlash followed the early successes of RI. According to Fig. 7, this occurred mostly because opposing discourse became more frequent after counter speech than after hate speech (panel (b)), and neutral speech became more polarized towards hate again (panel (c)).

[3] https://github.com/igorbrigadir/twitter-history/blob/master/changes.csv

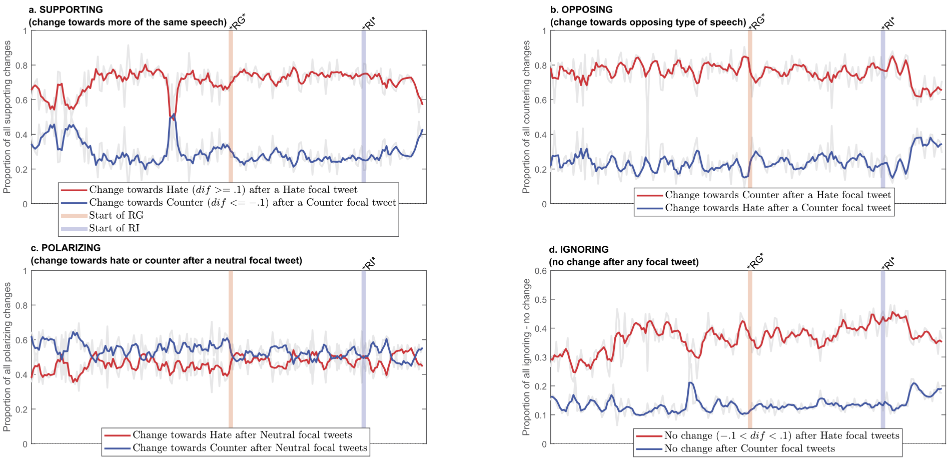

Figure 7 Reactions to tweets of different types. Proportion of focal tweets followed by different reactions, for 82,132 reply trees with at least 3 tweets. Shown are: numbers of focal tweets each week that are followed by (**a**) supporting, or changes towards the same type of speech, (**b**) countering, or changes towards the opposing type of speech, (**c**) polarizing, or changes towards hate or counter after a neutral focal tweet, and (**d**) ignoring the focal tweet. Each data point is a week average and trends are smoothed over a month-long window

4 Discussion

We present a large-scale longitudinal study of the dynamics of hate and counter speech in political discussions online. By analyzing hundreds of thousands of Twitter conversations through different lenses at both macro and micro levels, we contribute a nuanced picture of these dynamics. Across a number of different indicators, we find that organized counter speech appears to contribute to a more balanced public discourse. After the emergence of the organized counter group Reconquista Internet (RI) in the late Spring of 2018, the relative frequency of counter speech increased while that of hate speech decreased (Fig. 3). Counter speech became more related to less hate and more counter speech in the future (Fig. 4). The number of likes and the length of discussions after counter tweets increased after RI was founded (Fig. 5). Counter speech became more effective in steering conversations when it organized through RI, primarily by providing more support to counter tweets and by steering neutral discourse towards more counter speech (Figs. 6 and 7). While this was met with a backlash initially, the relative frequency of hate speech stabilized into a new, lower proportion of overall speech (Fig. 3). These findings suggest that, like in 'traditional' bullying settings, the presence of supporting peers (in this case, other individuals willing to engage in counter speech) motivates people to themselves oppose hate speech and defend its targets.

According to our results, citizens wishing to engage in counter speech would likely increase the effectiveness of their efforts if they organized and participated in discussions in a coordinated way. For example, they could organize around a central platform where members can communicate and strategize (e.g., RI members communicated via Discord server). As an organization they could then steer hateful conversations to a more neutral or even positive ground, by supporting the victims, voicing dissent to hateful positions, or even simply by "liking" counter messages so that they are more broadly visible. This is in line with similar results that were found qualitatively in [26].

Although we see an association of hate and counter speech and subsequent discourse according to a number of indicators, we cannot draw causal conclusions about the effectiveness of organized hate or counter speech alone. German society has been undergoing significant self-examination about its values and political directions throughout the studied period and beyond. This has likely influenced the ongoing discourse beyond organized counter speech. Regardless, the multifaceted approach we took here suggests that organized counter speech may be a promising strategy in combating the spread of hate online.

There are several critical areas of future research on counter speech.[4] One is the analysis of which counter speech strategies are particularly effective in curbing hate. A few recent studies have begun to look at this problem qualitatively. For example, [26] interviewed more than 25 active counter speakers who were members of the so-called "I am here" (#jagärhär) counter speech movement. Asking important questions like, why they took part, what the counter speaker learned or got out of the experience, and perhaps most importantly what strategies did they find effective or detrimental in combating online hate. This kind of work is crucial because it allows for a first-hand account of the strategies being used but unfortunately does not lend itself to any quantitative analysis on these speakers strategies. Additionally, Keller & Askanius [25] used the same situation between RG and RI to perform a qualitative content analysis to understand the strategies being used by both groups. To this end, they examined the internal documents of both groups as well as 709 comments generated in response to 3 articles posted on Facebook. They concluded that "organized, communication- and exchange-oriented, rule-guided efforts can be effective to combat cyberhate," but also state that in some circumstances counter speech can cause further polarization of beliefs. Again, this analysis did not allow for any causal analysis of the effects of either group. An exciting avenue for future research would be to combine the tools and lessons of these qualitative studies with the large corpus we have used here to drill into and identify which of the many strategies being used by these groups was effective.

Another important open question is how events such as terrorist attacks and political rallies shape hate and counter speech trends. In our results, we observed some patterns echoing previous results on the relationship between group threat and extreme attitudes and behaviors [69–71], but without a clear baseline we cannot derive any strong causal conclusions. Future research could try to find possibilities for including exogenous events that could enable such analyses [5].

As it is common for users to report accounts of the opposing faction for policy violations, yet another open question is how hate and counter speech relate to longevity of user accounts, in particular what percent of different accounts survive over time. Unfortunately, out data does not allow for these analyses *post hoc* as it is impossible to say whether an account was deleted by the user, suspended temporarily or permanently for policy violations. Our data also does not allow us to detect "shadow banning", periods when an account appeared to be active but its posts were not visible to others. Future research could try to record these indicators in real time or with the help of the platform provider. Similarly, it would be interesting to study the percentage of removed tweets that were flagged as either hate or counter by our classification algorithm. But, again drawing any conclusions about *why* (e.g., removed for policy violations, deleted by user, originated from suspended

[4] For a comprehensive review of past research on counter speech see [15].

account, etc.) a tweet was removed would be challenging but could provide insight into the longevity of hate and counter speech on online platforms.

Recently, there has been a lot of interesting work (e.g., [72–74]) on detecting coordinated activity on social media, especially around propagation of misinformation. While this work is adjacent to our work it actually solves the converse problem. These studies focus on detecting suspected coordinated activity, while we already knew there was coordinated activities and were interested in studying the dynamics of these activities in our dataset. However, given the organized nature of these groups, it would be interesting to study whether automation such as bots or manipulation of trending topics were used by these groups. In the present work, we were interested in the nature and dynamics of this discourse no matter how they were generated i.e., via bot or human. However, future research should certainly adapt tools such as Botometer [75, 76], Botslayer [77], and similar tools to be applicable to Twitter reply trees. These analyses would provide valuable information on the mechanics of coordinated activities online.

In August of 2020 Twitter announced that in the second iteration of the Twitter API, conversations which occurred in the previous week are now obtainable directly from the API using the new "conversation_id" query feature. In January of 2021, the one-week restriction on conversation queries was relaxed when Twitter announced the "Academic Research product track" (or Academic API). With it, academics are now granted the ability to perform full-archival search of Twitter's history using any of the API's many endpoints and search queries. Hence, going forward academic researchers will be able to do longitudinal analysis similar to that which we are reporting here. However, there are a few caveats which we feel are worth briefly discussing. First, at the time of writing this, the full-archival search functionality was still in early access and many older reply trees were still unsearchable. During revisions of this paper we attempted to reconstruct our conversation dataset using this new API but were unsuccessful. In our experience, as of August 31, 2021, only 0.02% of the requested trees were partially returned and the rest were not found (see Additional file 1, Sect. S1.1 for a full discussion of this analysis). Second, in recent years Twitter has worked to remove content from their platform which violated its Terms of Service. While this is beneficial to the Twitter community at large, this means that when researchers query the new API for older conversations, the API's response may have a significant portion of the more extreme tweets removed. This effectively washes out the signal that was of most interest in the present research. See Fig. S4 for an example. In this figure, half of the conversation has been removed because the person participating in the conversation has been suspended. In other conversations we explored (not shown), we see similar messages involving tweets removed by users, removed for violation of Terms of Service etc. As such, while full-archival search of conversations will be possible going forward, compiling Twitter conversations in real time (as much as possible) will continue to be necessary when studying behavior that violates the platform's Terms of Service such as hate speech. We hope this study provides a template for future longitudinal studies of conversation dynamics.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-021-00314-6.

Additional file 1. Supplementary information (PDF 921 kB)

Acknowledgements
The authors would like to thank the anonymous reviewers for their valuable feedback.

Funding
J.G. was partially supported by an Omidyar and an Applied Complexity Fellowship at the Santa Fe Institute as well as NSF EAGER 1807478. J.-G.Y. was supported by a James S. McDonnell Foundation Postdoctoral Fellowship Award. L.H.-D. and J-G.Y. were supported by Google Open Source under the Open-Source Complex Ecosystems And Networks (OCEAN) project. M.G. was partially supported by NSF-DRMS 1757211. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Google Open Source or the NSF.

Availability of data and materials
The models and trends that support the findings of this study are available from the corresponding author upon reasonable request. However, for ethical reasons as well as to protect the identity of counterspeakers raw data will not be released. Python and Matlab codes used for different analyses are available from the corresponding author upon request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
JG and KGZ designed the study, collected/scraped the Twitter data used in this study, as well as developed the classification pipeline. All authors contributed to analysis and interpretation of the data and to the writing of the manuscript. All authors read and approved the final manuscript.

Author details
[1]Santa Fe Institute, 1399 Hyde Park Road, 87501, Santa Fe, NM, USA. [2]Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, 04103, Leipzig, Germany. [3]Department of Computer Science and Vermont Complex Systems Center, University of Vermont, 05405, Burlington, VT, USA. [4]Complexity Science Hub Vienna, Josefstädter Straße 39, 1080, Vienna, Austria. [5]Vermont Complex Systems Center, University of Vermont, 05405, Burlington, VT, USA.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 26 April 2021 Accepted: 21 December 2021 Published online: 24 January 2022

References

1. Bakalis C (2015) Cyberhate: an Issue of Continued Concern for the Council of Europe's Anti-racism Commission. Council of Europe
2. Blaya C (2019) Cyberhate: a review and content analysis of intervention strategies. Aggress Violent Behav 45:163–172
3. Gagliardone I, Gal D, Alves T, Martinez G (2015) Countering online hate speech. Unesco Publishing
4. Hawdon J, Oksanen A, Räsänen P (2017) Exposure to online hate in four nations: a cross-national consideration. Deviant Behav 38(3):254–266
5. Müller K, Schwarz C (2019) Fanning the flames of hate: social media and hate crime. Available at SSRN 3082972
6. Oksanen A, Kaakinen M, Minkkinen J, Räsänen P, Enjolras B, Steen-Johnsen K (2018) Perceived societal fear and cyberhate after the november 2015 paris terrorist attacks. Terrorism Polit Violence 1–20
7. Weber A (2009) Manual on hate speech. Council Of Europe
8. YouTube: Hate speech policy. https://support.google.com/youtube/answer/2801939. Accessed: 2020-02-28
9. Twitter: Hateful conduct policy. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy. Accessed: 2020-02-28
10. Facebook: Hate speech. https://www.facebook.com/communitystandards/hate_speech. Accessed: 2020-02-28
11. Chandrasekharan E, Pavalanathan U, Srinivasan A, Glynn A, Eisenstein J, Gilbert E (2017) You can't stay here: the efficacy of reddit's 2015 ban examined through hate speech. In: Proceedings of the ACM on human-computer interaction 1(CSCW), pp 1–22
12. Benesch S, Ruths D, Dillon K, Saleem H, Wright L (2016) Considerations for successful counterspeech. Dangerous Speech Proj. Report available at https://dangerousspeech.org
13. Rieger D, Schmitt JB, Frischlich L (2018) Hate and counter-voices in the Internet: introduction to the special issue. SCM Stud Commun Media 7(4):459–472
14. Wachs S, Wright MF, Sittichai R, Singh R, Biswal R, Kim E-M, Yang S, Gámez-Guadix M, Almendros C, Flora K et al (2019) Associations between witnessing and perpetrating online hate in eight countries: the buffering effects of problem-focused coping. Int J Environ Res Public Health 16(20):3992
15. Buerger C, Wright L (2019) Counterspeech: a literature review
16. Ziegele M, Jost P, Bormann M, Heinbach D (2018) Journalistic counter-voices in comment sections: patterns, determinants, and potential consequences of interactive moderation of uncivil user comments. SCM Stud Commun Media 7(4):525–554
17. Habermas J (2015) Between facts and norms: contributions to a discourse theory of law and democracy. Wiley, New York
18. Brassard-Gourdeau É, Khoury R (2018) Impact of sentiment detection to recognize toxic and subversive online comments. arXiv preprint 1812.01704

19. Burnap P, Rana OF, Avis N, Williams M, Housley W, Edwards A, Morgan J, Sloan L (2015) Detecting tension in online communities with computational Twitter analysis. Technol Forecast Soc Change 95:96–108
20. Burnap P, Williams ML (2016) Us and them: identifying cyber hate on Twitter across multiple protected characteristics. EPJ Data Sci 5(1):11
21. MacAvaney S, Yao H-R, Yang E, Russell K, Goharian N, Frieder O (2019) Hate speech detection: challenges and solutions. PLoS ONE 14(8)
22. Ribeiro MH, Calais PH, Santos YA, Almeida VA, Meira W Jr (2018) Characterizing and detecting hateful users on Twitter. In: Twelfth international AAAI conference on web and social, Media
23. Zhang Z, Luo L (2019) Hate speech detection: a solved problem? The challenging case of long tail on Twitter. Semant Web 10(5):925–945
24. Zimmerman S, Kruschwitz U, Fox C (2018) Improving hate speech detection with deep learning ensembles. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)
25. Keller N, Askanius T (2020) Combatting hate and trolling with love and reason? A qualitative analysis of the discursive antagonisms between organized hate speech and counterspeech online. SCM Stud Commun Media 9(4):540–572
26. Buerger C (2020) The anti-hate brigade: how a group of thousands responds collectively to online vitriol. Available at SSRN 3748803
27. Gaffney H, Farrington DP, Espelage DL, Ttofi MM (2019) Are cyberbullying intervention and prevention programs effective? A systematic and meta-analytical review. Aggress Violent Behav 45:134–153
28. Mathew B, Kumar N, Goyal P, Mukherjee A (2018) Analyzing the hate and counter speech accounts on Twitter. arXiv preprint 1812.02712
29. Mathew B, Saha P, Tharad H, Rajgaria S, Singhania P, Maity SK, Goyal P, Mukherjee A (2019) Thou shalt not hate: countering online hate speech. In: Proceedings of the international AAAI conference on web and social media, vol 13, pp 369–380
30. Stroud NJ, Scacco JM, Muddiman A, Curry AL (2015) Changing deliberative norms on news organizations' Facebook sites. J Comput-Mediat Commun 20(2):188–203
31. Wright L, Ruths D, Dillon KP, Saleem HM, Benesch S (2017) Vectors for counterspeech on Twitter. In: Proceedings of the first workshop on abusive language online, pp 57–62
32. Garland J, Ghazi-Zahedi K, Young J-G, Hébert-Dufresne L, Galesic M (2020) Countering hate on social media: large scale classification of hate and counter speech. In: Proceedings of the fourth workshop on online abuse and harms. Association for Computational Linguistics, Online, pp 102–112. https://doi.org/10.18653/v1/2020.alw-1.13. https://www.aclweb.org/anthology/2020.alw-1.13
33. Salmivalli C (2014) Participant roles in bullying: how can peer bystanders be utilized in interventions? Theory Pract 53(4):286–292
34. Gini G, Albiero P, Benelli B, Altoe G (2008) Determinants of adolescents' active defending and passive bystanding behavior in bullying. J Adolesc 31(1):93–105
35. Salmivalli C, Voeten M, Poskiparta E (2011) Bystanders matter: associations between reinforcing, defending, and the frequency of bullying behavior in classrooms. J Clin Child Adolesc Psychol 40(5):668–676
36. Huckfeldt RR, Sprague J (1995) Citizens, politics and social communication: information and influence in an election campaign. Cambridge University Press, Cambridge
37. Lazer D (2011) Networks in political science: back to the future. PS Polit Sci Polit 44(1):61–68
38. Sinclair B (2012) The social citizen: peer networks and political behavior. University of Chicago Press, Chicago
39. Brown GD, Wood AM, Ogden RS, Maltby J (2015) Do student evaluations of university reflect inaccurate beliefs or actual experience? A relative rank model. J Behav Decis Mak 28(1):14–26
40. Farrow K, Grolleau G, Ibanez L (2017) Social norms and pro-environmental behavior: a review of the evidence. Ecol Econ 140:1–13
41. Banerjee A, Chandrasekhar AG, Duflo E, Jackson MO (2013) The diffusion of microfinance. Science 341(6144)
42. Bond RM, Fariss CJ, Jones JJ, Kramer AD, Marlow C, Settle JE, Fowler JH (2012) A 61-million-person experiment in social influence and political mobilization. Nature 489(7415):295–298
43. Croker H, Whitaker K, Cooke L, Wardle J (2009) Do social norms affect intended food choice? Prev Med 49(2–3):190–193
44. Christakis NA, Fowler JH (2009) Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives. Little, Brown Spark
45. Álvarez-Benjumea A, Winter F (2018) Normative change and culture of hate: an experiment in online environments. Eur Sociol Rev 34(3):223–237
46. Matias JN (2019) Preventing harassment and increasing group participation through social norms in 2190 online science discussions. Proc Natl Acad Sci 116(20):9785–9789
47. Davey J, Ebner J (2017) The fringe insurgency. Connectivity, convergence and mainstreaming of the extreme right. Institute for Strategic Dialogue
48. Gensing P (2018) Information war by all means. Tagesschau. tagesschau.de/faktenfinder/inland/organisierte-trolle-101.html
49. Anders R (2018) Lösch Dich! So organisiert ist der Hate im Netz I Doku über Hater und Trolle. Funk net Documentary. youtube.com/watch?v=zvKjfWSPI7s
50. Reconquista Internet Codex. https://www.reddit.com/r/ReconquistaInternet/wiki/index#wiki_reconquistainternet_wiki
51. News Portals from Germany. https://www.deutschland.de/en/topic/knowledge/news
52. Schmidt A, Wiegand M (2017) A survey on hate speech detection using natural language processing. In: Proceedings of the fifth international workshop on natural language processing for social media, pp 1–10
53. Le Q, Mikolov T (2014) Distributed representations of sentences and documents. In: International conference on machine learning, pp 1188–1196
54. Řehůřek R, Sojka P (2010) Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks. ELRA, Valletta, pp 45–50. http://is.muni.cz/publication/884893/en

55. Lau JH, Baldwin T (2016) An empirical evaluation of doc2vec with practical insights into document embedding generation. arXiv preprint 1607.05368
56. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
57. Ziems C, He B, Soni S, Kumar S (2020) Racism is a virus: anti-asian hate and counterhate in social media during the covid-19 crisis. arXiv preprint 2005.12423
58. Jaki S, De Smedt T (2019) Right-wing german hate speech on Twitter: analysis and automatic detection. arXiv preprint 1910.07518
59. Abadie A (2005) Semiparametric difference-in-differences estimators. Rev Econ Stud 72(1):1–19
60. Granger CW (1969) Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 424–438
61. Schreiber T (2000) Measuring information transfer. Phys Rev Lett 85(2):461
62. Takens F (1981) Detecting strange attractors in turbulence. In: Dynamical systems and turbulence. Springer, Warwick, pp 366–381
63. Sugihara G, May R, Ye H, Hsieh C-H, Deyle E, Fogarty M, Munch S (2012) Detecting causality in complex ecosystems. Science 338(6106):496–500
64. Killick R, Fearnhead P, Eckley IA (2012) Optimal detection of changepoints with a linear computational cost. J Am Stat Assoc 107(500):1590–1598
65. Aminikhanghahi S, Cook DJ (2017) A survey of methods for time series change point detection. Knowl Inf Syst 51(2):339–367
66. Stock J, Watson MW (2001) Vector autoregressions. J Econ Perspect 15(4):101–116
67. Dickey DA, Fuller WA (1979) Distribution of the estimators for autoregressive time series with a unit root. J Am Stat Assoc 74(366a):427–431
68. Ziegele M, Jost PB (2016) Not funny? the effects of factual versus sarcastic journalistic responses to uncivil user comments. Commun Res 0093650216671854
69. Jahnke S, Abad Borger K, Beelmann A (2021) Predictors of political violence outcomes among young people: a systematic review and meta-analysis. Polit Psychol
70. Jost JT, Glaser J, Kruglanski AW, Sulloway FJ (2003) Political conservatism as motivated social cognition. Psychol Bull 129(3):339
71. Turner ME, Pratkanis AR, Probasco P, Leve C (1992) Threat, cohesion, and group effectiveness. J Pers Soc Psychol 63(5):781
72. Alizadeh M, Shapiro JN, Buntain C, Tucker JA (2020) Content-based features predict social media influence operations. Sci Adv 6(30):5824
73. Pacheco D, Hui P-M, Torres-Lugo C, Truong BT, Flammini A, Menczer F (2021) Uncovering coordinated networks on social media: methods and case studies. In: Proceedings of the fifteenth international AAAI conference on web and social media, vol 15, pp 455–466
74. Pacheco D, Flammini A, Menczer F (2020) Unveiling coordinated groups behind white helmets disinformation. In: Companion proceedings of the web conference 2020, pp 611–616
75. Davis CA, Varol O, Ferrara E, Flammini A, Menczer F (2016) Botornot: a system to evaluate social bots. In: Proceedings of the 25th international conference companion on world wide web, pp 273–274
76. Varol O, Ferrara E, Davis C, Menczer F, Flammini A (2017) Online human-bot interactions: detection, estimation, and characterization. In: Proceedings of the international AAAI conference on web and social media, vol 11
77. Hui P-M, Yang K-C, Torres-Lugo C, Monroe Z, McCarty M, Serrette BD, Pentchev V, Menczer F (2019) Botslayer: real-time detection of bot amplification on Twitter. J Open Sour Softw 4(42):1706