Measuring the prevalence of online hate speech, with an application to the 2016 U.S. election

Alexandra A. Siegel[1], Evgenii Nikitin[2], Pablo Barberá[3], Joanna Sterling[4], Bethany Pullen[5], Richard Bonneau[6], Jonathan Nagler[7], and Joshua A. Tucker[8*]

October 2018

[1] Department of Politics and NYU SMaPP laboratory, New York University, New York, New York, USA.

[2] Department of Politics and NYU SMaPP laboratory, New York University, New York, New York, USA.

[3] Methodology Department, London School of Economics, London, UK.

[4] Woodrow Wilson School, Princeton University, Princeton, New Jersey, USA.

[5] NYU SMaPP laboratory, New York University, New York, New York, USA.

[6] Department of Biology and NYU SMaPP laboratory, New York University, and Simons Foundation, New York, New York, USA.

[7] Department of Politics and NYU SMaPP laboratory, New York University, New York, New York, USA.

[8] Department of Politics and NYU SMaPP laboratory, New York University, New York, New York, USA.

[*]Corresponding author email: joshua.tucker@nyu.edu

Abstract

Despite a growing body of research devoted to defining and detecting online hate speech and extremist rhetoric, the existing scientific literature lacks a systematic framework for assessing how the content and popularity of these harmful messages change over time. We offer a new approach to measuring the real-time prevalence of online hate, using both context-specific data and data produced by a large random sample of users; employing multiple methods of text classification; and measuring not only the volume, but also the proportion, and number of unique users producing it. Here we apply our framework to test the widely-held proposition that Donald Trump's divisive 2016 campaign and election has popularized online hate speech and white nationalist rhetoric in the American Twittersphere. Highlighting the need for such a systematic approach—contrary to the conventional wisdom—our analysis of over one billion tweets demonstrates that online hate did not become more popular on Twitter either over the course of the campaign or in the aftermath of Trump's election.

One Sentence Summary

Offering a new framework for measuring the prevalence of online hate speech over time, an application of our method analyzing over 1 billion tweets demonstrates that hate speech and white nationalist language did not systematically increase on Twitter over the course of the 2016 U.S. election campaign or following Trump's election.

Main Text

As popular social media platforms have increased the visibility of online hate speech, it has surged to the forefront of scientific, legal, and policy-making agendas. From targeted anti-Semitic attacks on Jewish journalists (*1*) to reports of social media's role in mobilizing ethnic violence in a variety of global contexts (*2*), the offline consequences of online hate speech have received increased attention. However, despite a growing body of research defining and detecting online hate speech, the existing scientific literature lacks a systematic framework for assessing how the volume and content of online hate changes over time (*3, 4*). Although almost no empirical work has explicitly measured the overall prevalence or temporal dynamics of harmful speech on popular social media sites (*5*), governments and online platforms have increasingly proposed and adopted policy interventions to combat online hate speech (*4, 6, 7*).

Here, we take a first step towards rectifying this knowledge gap by presenting a research framework for systematically investigating changes in the real-time popularity of online hate speech and extremist rhetoric. This approach has three components. First, it uses both context-specific data and representative data. Context-specific data refers to online content related to particular events, people, or topics. This enables us to measure the popularity of online hate in specific domains where we theoretically expect hate speech (or other online content) to be most prevalent. By contrast, representative data refers to posts produced by a random sample of users in a population of interest. This enables us to measure the prevalence of online hate speech among users of a specific platform or residents of a particular country, for example, regardless of the topics they discuss. Second, our approach requires the use of at least two methods of text-classification to ensure that any trends we observe in either content-specific or representative data are not driven by the biases introduced by a particular method. Third, our framework introduces multiple outcomes of interest when assessing trends in the use of online speech over time. These include 1) the overall volume of posts containing the language, a measure of absolute prevalence; 2) the number of unique users or accounts producing such language, a measure of absolute popularity; 3) the proportion of posts containing the language, a measure of relative prevalence; and 4) the proportion of unique users producing such language, a measure of relative popularity. When measuring increases in the use of particular types of language on online platforms it is possible to observe changes in overall volume without changes in relative volume; alternatively, we could observe changes in the numbers of posts without changes in the number of users producing them. To assess changes in the use of online speech systematically, therefore, it is important to assess each of these distinct outcome measures to avoid missing key trends.

Demonstrating the utility of this framework, we apply it in the American Twittersphere

during the 2016 presidential election campaign and its aftermath. Twitter is an ideal platform on which to study changes in the prevalence and popularity of online hate speech over time as it is widely used by journalists and political elites, helps shapes conventional media reporting, and is used by approximately a quarter of Americans. In particular, we use our approach to test the widely-held proposition that the use of hate speech and white nationalist language increased over the course of Donald Trump's divisive 2016 campaign and following his unexpected election election. We analyze over one billion tweets, including approximately 750 million context-specific political tweets as well as 400 million tweets collected from a random sample of 500,000 American Twitter users. We collected our data across a two year period from the start of the presidential election campaign through the summer of 2017. To measure dynamic changes in the prevalence of hate speech over the course of Trump's campaign and in the aftermath of his election, we develop two distinct text-classification approaches. The first is a a machine learning augmented dictionary-based method and the second is a novel non-dictionary-based method leveraging data from Reddit communities associated with the alt-right movement. Finally, satisfying the third component of our famework, we assess changes not only in the volume and relative volume of online hate speech and white nationalist rhetoric in our datasets, but also changes in the absolute and relative numbers of unique users producing such content.

Our paper therefore offers three novel contributions. First, we lay out a new research framework for measuring the real-time popularity of online hate or other forms of online speech. This framework involves using both context-specific and representative data, employing multiple approaches to text classification, and using several measures to assess trends in online hate speech over time. Second, we present two distinct methods for identifying hate speech in social media data. Third, we provide an empirical application of our framework in one particularly consequential domain: The American Twittersphere during the 2016 US presidential election campaign and its aftermath.

In contrast to the prevailing popular narrative, we find no persistent increase in hate speech or white nationalist language either over the course of Trumps campaign or in the aftermath of his election. Instead, hate speech was bursty: while there were notable spikes in hateful language in response to particular events, these effects quickly dissipated. This application of our framework demonstrates the need to move beyond short term or small-scale datasets when studying politically relevant behavior online.

In the remainder of this article, we present an application of our method to illustrate its utility for measuring systematic changes in online hate—as well as other forms of online speech—over time.

Application: Online hate in the American Twittersphere

The spread of online hate speech on popular social media platforms has received increasing attention in the U.S. media, particularly as a consequence of Donald Trump's political rise. Citing a "massive rise" in online hate speech, media reports suggest that Trump's divisive campaign and subsequent election legitimized extremist ideologies—popularizing hostile messages that were once relegated to the dark corners of the Internet (*8, 9*). Articles like the *USA Today's* "Massive rise in hate speech on Twitter during presidential election," *The New Yorker's* "Hate is on the Rise After Trump's Election," *The Guardian's* "Trump's Election led to Barrage of Hate," and *Vox's* "The Wave of Post-Election Hate Reportedly Sweeping the Nation, Explained," have proliferated. James King's year-in-review column, "The Year in Hate: From Donald Trump to the Rise of the Alt-Right," Salon's "A Short History of Hate," which tracks the alt-right's 2016 ascendance, and the New York Times' hate-speech aggregator, "This Week in Hate," are just a few examples of this trend (*8, 10–15*). Fearing that Trump's election created a new "safe space for hate," academics, journalists, policy makers, and everyday citizens are increasingly voicing concern about the consequences of Trump's actions and rhetoric both on and offline (*12, 16, 17*).

However, despite a wealth of anecdotal and small-scale empirical evidence of this "Trump effect"—and widespread acceptance of this prevailing narrative among academics, journalists, and policy makers—little is known about how the quantity of online hate speech, or the number of individuals producing it, has changed over time. In particular, although headlines such as "Massive Rise in Hate Speech on Twitter during the Presidential Election" (*8*) have proliferated, no studies have systematically analyzed changes in the use of hate speech on Twitter in this period.

Here we apply the framework outlined above to assess whether Trump's campaign and election were associated with an increase in hate speech and white nationalist rhetoric on Twitter. First, we rely on two sources of data. For our context-specific dataset we use a political dataset containing over 600 million tweets referencing Donald Trump and over 150 million referencing Hillary Clinton. These are collections of tweets containing keywords associated with the candidates produced between June 17, 2015 – the day after Trump announced his candidacy – and June 15, 2017. This political Twitter dataset gives us a comprehensive snapshot of political discourse throughout the 2016 election period and its aftermath, a place where we might expect to see an increase in the use of online hate speech connected to the election. For our representative dataset, we use a collection of tweets sent by a random sample of 500,000 American Twitter users. This enables us to assess the popularity of online hate speech among American Twitter users more broadly, beyond

explicitly political discourse. These users were sampled by generating random user IDs and then checking that their accounts were active and located in the United States (details are available in the supplementary materials, section S1.1).

Following our framework, we then develop two distinct methods for classifying text as containing hate speech and white nationalist rhetoric. The goal here is to ensure that any changes we observe in the popularity of online hate speech are not driven by our particular classification approach. We then use these methods to measure changes both in the volume, proportion, and number and proportion of unique users producing this content in both of our datasets.

Seeking to understand changes in the use of this language over time we care not only about the raw count of tweets containing such language, but also their relative popularity, and the number and relative number of unique users producing such content in each dataset. Below we describe our classification methods and present the results first of the dictionary-based method and then of our approach leveraging data from alt-right subreddits. While the results presented in the main body of the paper show the proportion of tweets containing hate speech or white nationalist rhetoric, they look very similar to analysis of both raw counts and unique users, which are provided in the supplementary materials. This increases our confidence that our findings are not driven by one particular measure of popularity or prevalence of online hate, demonstrating the utility of including these measures in our framework.

Drawing on commonly used definitions in the hate speech literature, we define hate speech and white nationalist rhetoric as follows:

- **Hate Speech** is bias-motivated, hostile and malicious language targeting a person or group because of their actual or perceived characteristics, especially when the group or individual are unnecessarily labeled (*18, 19*).
- **White Nationalist Rhetoric** is content that praises known white-nationalist groups, espouses white supremacist or white separatist ideologies, or focuses on the alleged inferiority of nonwhites (*20, 21*).

Past studies of online hate speech have frequently relied on dictionary-based methods, which require a priori knowledge of words and phrases associated with hate speech (*22–24*). Other studies have identified hate speech using human coded content, network analysis of online hate groups, sentiment analysis, natural language processing, neural networks, and other machine learning approaches (*25–29*). We first use a dictionary-based method in which we develop lists of anti-Asian, anti-Black, anti-immigrant, anti-Muslim, anti-Semitic,

homophobic, and misogynistic slurs, as well as a dictionary of white nationalist rhetoric. We select these terms using two pre-existing databases of online hate speech, Hatebase and the Racial Slur Database (*30, 31*), in addition to the Anti-Defamation League's database of white-nationalist language (*32*) (the full list of terms is available in the supplementary materials, data S1).

One of the main challenges in dictionary-based hate-speech detection is distinguishing between hate speech itself and posts employing these terms for other purposes, such as self-referential use of slurs. These methods often have low precision because they identify all messages containing particular slurs as hate speech, failing to account for alternative uses of such words. To confront this problem, we used two binary supervised classifiers, one to identify hate speech tweets and one to identify tweets containing white nationalist rhetoric. These classifiers were trained on a random sample of 25,000 tweets containing hate speech or white nationalist terms from our three datasets. These tweets were labeled by undergraduate volunteers and crowd-sourced coders (details are available in the supplementary materials, section S1.2). Our classifiers allowed us to remove false positives from our dictionary-filtered datasets, significantly improving the accuracy of our method (details are available in the supplementary materials, sections S1.3 and S1.4). We found that fewer than one third of the tweets containing terms from our hate-speech dictionaries actually contained hate speech or white nationalist language, highlighting the need to move beyond a purely dictionary-based approach (examples of these tweets are available in the supplementary materials, Table S1).

Results

The raw data offer important clues about the effect of Trump's political rise and election on the popularity of online hate speech. Figure 1 shows the monthly proportion of hate speech tweets produced in the Clinton, Trump, and random sample datasets between June 17, 2015 and June 15, 2017 (see Figure S2 for hate speech disaggregated by target; see Figure S11 and S12 for plots displaying raw counts of tweets rather than proportions, which yield similar conclusions). We find that in general between 0.1% and 0.2% of tweets contain contain hate speech. Figure 1 shows that Trump's election (in November 2016) does not appear to have increased the proportion of hate speech in the Clinton and random sample datasets. The Clinton dataset contains less hate speech after the election, while the random sample data remains about the same for several months. The largest spike in monthly hate speech in the Trump dataset occurs in late January 2017. Analysis of the Trump data reveals that this spike is largely explained by a large uptick in misogynistic hate speech following the announcement of Trump's "travel ban" executive order. This increased misogynistic

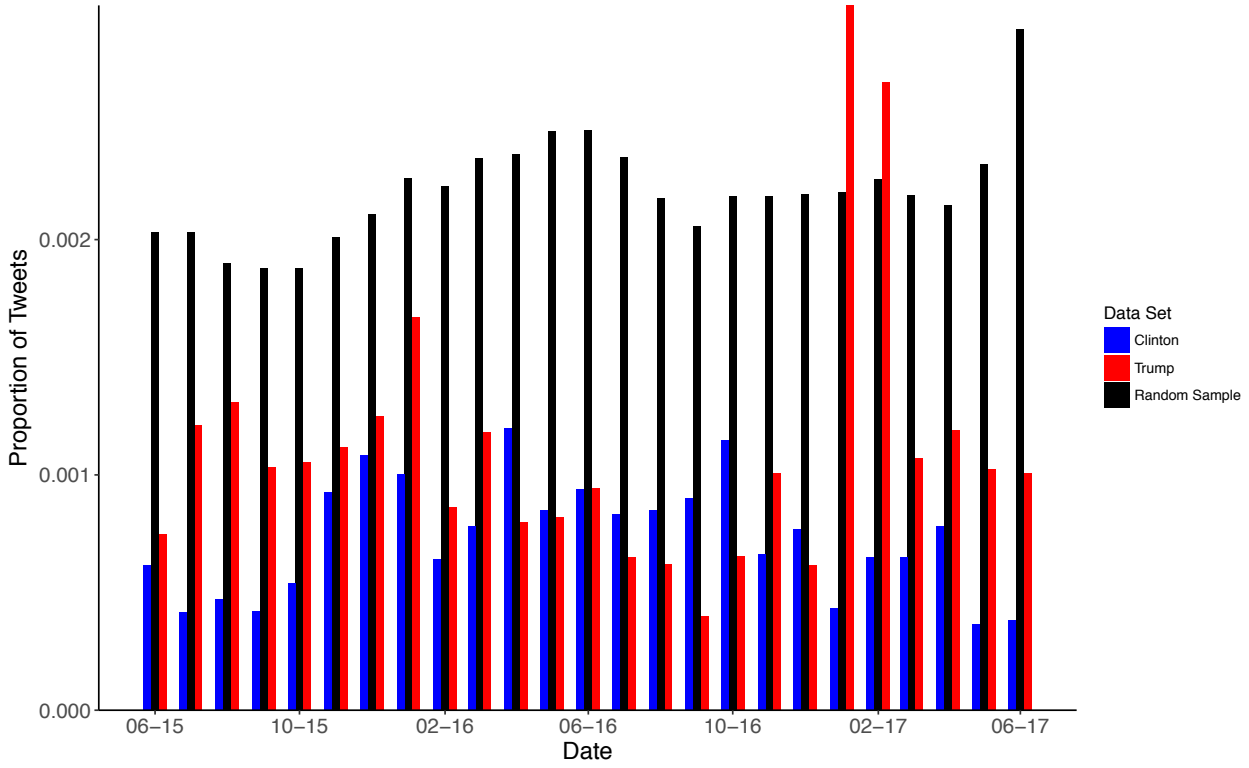

Figure 1. Monthly Proportion of Classified Hate Speech Tweets in the Clinton, Trump, and Random Sample Datasets. This figure shows the monthly proportion of classified hate-speech tweets in the Clinton, Trump, and random sample datasets of tweets containing hate speech dictionary terms. We classified tweets as hate speech (or not) using a Naive Bayes Classifier to remove false positives from our datasets. Similar plots for white nationalist language, and plots displaying raw counts of the data rather than proportions are available in the supplementary materials (Figure S8-S10).

language appears to be a reaction to Clinton's decision to break her post-election silence to criticize the ban, as well as language directed at attorney general Sally Yates, who declined to defend the ban and was then fired by Trump. There are also spikes in anti-Asian, anti-Muslim, and anti-Black language in this period, though their volume is much lower (see Figure S2).

While Figure 1 shows little evidence of a persistent increase in hate speech over the course of the campaign or in the aftermath of the election, perhaps the data aggregating all types of hate speech masks important more granular trends. Additionally, we are concerned not just with the total volume of tweets, but also with changes in the number of unique users producing them (monthly plots of the number of unique users tweeting this content are available in the supplementary materials, Figure S14 and S15, and look quite similar to Figure 1). Testing the extent to which this language became more popular either over the course of the 2016 campaign or following Trumps election more systematically, we rely on Interrupted Time Series Analysis (ITSA)—a powerful quasi-experimental design for assessing the longitudinal impact of an event or intervention (*33, 34*). In particular, ITSA allows us to measure whether online hate speech was increasing over the course of Trump's campaign, as well as whether Trump's unexpected election emboldened people to share more hostile and extreme content (details of the model are provided in Section S.2.1 of the supplementary materials).

Conducting ITSA using our political Twitter and random sample datasets, we find no evidence of a lasting increase in hate speech or white nationalist rhetoric either over the course of the campaign or in the aftermath of Trump's election. We include retweets in our analysis because they play an important role in increasing the visibility of online hate speech, and the majority of hateful tweets in our datasets are, in fact, retweets. When we remove retweets we still observe no persistent increases and many of the short-term bursts of hate speech and white nationalist language disappear (Figure S8, S9, and S10 show the volume of tweets vs. retweets in our Clinton, Trump, and random sample datasets). Additionally, while bot activity is always a concern when studying Twitter data, we do not make an effort to exclude bots from our analysis as any hate speech tweets produced by bots would increase the visibility of hateful language in the American Twittersphere and are therefore relevant to our analysis. It is also unlikely that many bots are present in our random sample of 500,000 American Twitter users due to the manner in which this collection was created (see supplementary materials for details).

In Figure 2 we plot the pre and post-election trend lines over the observed daily proportion of hate speech tweets and white nationalist language tweets in our datasets. Beginning with the Trump dataset (Panels A and B)—by far the largest collection—we see no significant increase in total hate speech or white nationalist language in either period. As Panel A demonstrates, the largest spike in hate speech in the Trump dataset occurred in late January and early February, in the period surrounding the aforementioned travel ban. By contrast the largest spike in white nationalist rhetoric occurs following Trump's retweet of a white supremacist account in February 2016. Similarly, there are no persistent increases in the

number of unique users producing this content and these results hold using both linear and quadratic models (plots and regression tables displaying these results for all datasets and outcome variables are available in the supplementary materials, Figure S16-S33 and Tables S8-S25).

Turning to our Clinton hate speech data (Panel C in Figure 2), we again observe no change in the proportion of hate speech over the course of the 2016 campaign or in the aftermath of Trump's election. In fact, we actually observe a statistically significant post-election decrease in the number of unique users producing hate speech. We do, however, observe a statistically significant increase in the number of unique users tweeting hate speech over the pre-election period (these results are provided in the supplementary materials, Figure S21 and Table S13). This effect is primarily driven by the increase in misogynistic rhetoric over the course of the Clinton campaign, particularly a spike following her April 2016 debate against Bernie Sanders and a general uptick as the election approached. While we see no evidence of increasing white nationalist rhetoric over the course of the campaign in the Clinton dataset (Panel D in Figure 2), we do see an increase in the proportion of tweets containing white nationalist rhetoric—and the proportion of unique users tweeting them—after Trump's election (See Figure S16, S17, S19, and S20 and Tables S8, S9, S11, and S12 in the supplementary appendix). However, this increase represents a change of only a fraction of a percentage point and—even on the most prolific days—we never observe more than a few hundred white nationalist tweets in the Clinton dataset.

Following our framework, we replicate the findings from our content-specific political Twitter datasets in a representative sample of data in order to address the possibility that our political Twitter data differ systematically from the U.S. Twittersphere as a whole. As we described earlier, we use a dataset of over 400 million tweets produced by a random sample of 500,000 American Twitter users. Consistent with our results on political Twitter, Panel E in Figure 2 shows no lasting increase in online hate speech over the course of Trump's campaign. Furthermore, there is no statistically significant increase in the proportion of tweets containing hate speech following the election—or the number of unique users tweeting them—although we do see a brief one-day spike in the number of unique users tweeting hate speech on election day. Once again, this spike in hate speech is largely driven by misogynistic language, though we see a one-day spike in anti-black language as well in the random sample data.

Examining trends in white nationalist rhetoric in the random sample dataset, which we plot in Panel F in Figure 2, we see no increase in the proportion of tweets containing white nationalist rhetoric over the course of the campaign. While there is some evidence

of a slight increase in white nationalist rhetoric after Trump's election, this effect is not statistically significant across specifications and the volume of tweets is even lower than it is in the Clinton collection. Taken together, while we do observe slight increases in white nationalist rhetoric following Trump's election in the Clinton and random sample datasets, and a one-day spike in hate speech in the random sample data, these results do not provide support for the conventional wisdom that Trump's election prompted a "mainstreaming" or popularization of online hate.

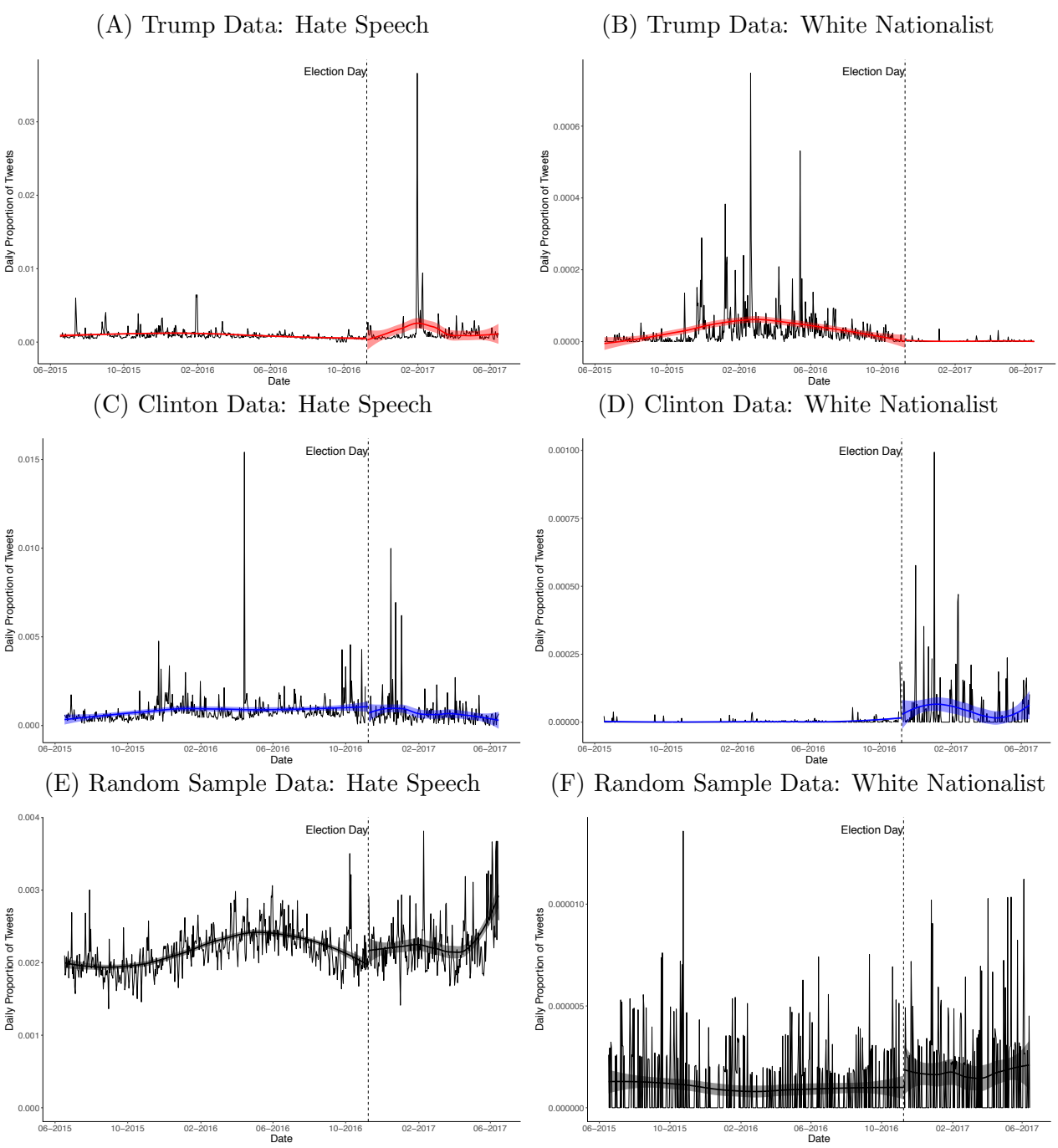

Figure 2. Effect of 2016 Election on Daily Proportion of Hate Speech and White Nationalist Language Tweets. These plots show the pre and post election trends in our ITSA regression models, plotted as local regression lines with loess smoothing and 95% confidence intervals. These trend lines are plotted against the observed daily proportion of hate speech tweets and white nationalist language tweets in our datasets of over 600 million tweets referencing Donald Trump (A and B), 150 million tweets referencing Hillary Clinton (C and D), and 400 million tweets sent by a random sample of American Twitter Users collected using Twitter's Streaming API between June 17, 2015 and June 15, 2017 (E and F). Hate speech and white nationalist language tweets were identified both using dictionaries of slurs and Naive Bayes classifiers trained to remove false positives from our data.

One of the potential pitfalls of using dictionary-based methods for identifying hate speech—no matter how sophisticated the application of these approaches—is that they force the analyst to rely on a corpus of words used in the "past" to code speech in the present. In most contexts, this is unlikely to be problematic, due to the long lifespan of slurs and derogatory language. However, given our surprising finding that hate speech and white nationalist rhetoric did not increase persistently either over the course of the 2016 campaign or in the aftermath of Trump's election, we must seriously consider that we have somehow failed to identify a significant subset of hateful language on Twitter. Because language evolves quickly on social media (*35*), it is entirely possible that our dictionary-based method is underestimating the prevalence of hate speech on this platform.

With this concern in mind, applying the multi-method text classification component of our framework, we repeat our analyses using a non-dictionary-based classification method to measure the prevalence of hateful language. The method is explained and validated in detail in the supplementary materials (sections S2.2, S2.3, and S2.4), but the underlying motivation is to find an example of hate speech in the wild, or a large collection of labeled text that contains the types of hostile rhetoric people actually use online. For this task, we rely on publicly available comments posted on Reddit.com. Reddit is a popular news aggregation and discussion website organized into topics or subreddits, some of which are are infamous for their explicitly racist, hateful and extreme alt-right content. Further, Reddit users can up-vote or down-vote posts. By deleting posts that have net negative votes from the data used to train our classifier, we subject our text to two forms of annotation: whether it is posted in the subreddit in the first place; and whether users of that subreddit think it belong there.

We can thus harness large naturally annotated corpora of text containing hate speech and white nationalist language—Reddit comments posted in alt-right communitiesto develop a non-dictionary based approach to classifying tweets. More specifically, we can train a classifier that outputs the probability that a particular document belongs to a particular corpus (e.g., that a tweet belongs to a collection of alt-right subreddits). This probability measures how semantically similar each document is to this subreddit (or group of subreddits), compared to other subreddits. Put another way, we check whether the words and phrases in the political and random sample tweets produced each day are more similar to words and phrases that are popular on alt-right subreddits than they are to the words and phrases used in discussing other topics. In a manner analogous to the dictionary-based method, we can then model whether this similarity increases over the course of the campaign or in the aftermath of the election. The advantage here is that unlike in our first method, we do not need to explicitly provide a dictionary of alt-right terms and phrases. Instead, our model

can automatically learn relevant terms from the corpus of subreddit comments. This method is particularly useful for measuring the popularity of alt-right language, as there are many well-known communities on Reddit that openly declare their alt-right views.

In line with our dictionary-based analysis, and again contrary to the received wisdom, we do not observe the language in the political or random sample Twitter collections becoming more similar to content produced on alt-right subreddits over the course of the campaign. Trump's election also has no effect on these probabilities. These findings are displayed in Figure 3 (regression tables displaying results are provided in the supplementary materials, Tables S29, S30, and S31). Thus using two different datasets (a random sample of American Twitter users and a large corpus of political tweets), employing both dictionary and non-dictionary based methods, and using the volume and proportion of tweets and unique users as measures of prevalence, we find no evidence that online hate speech systematically increased either over the course of Trump's 2016 campaign or in the aftermath of his election.

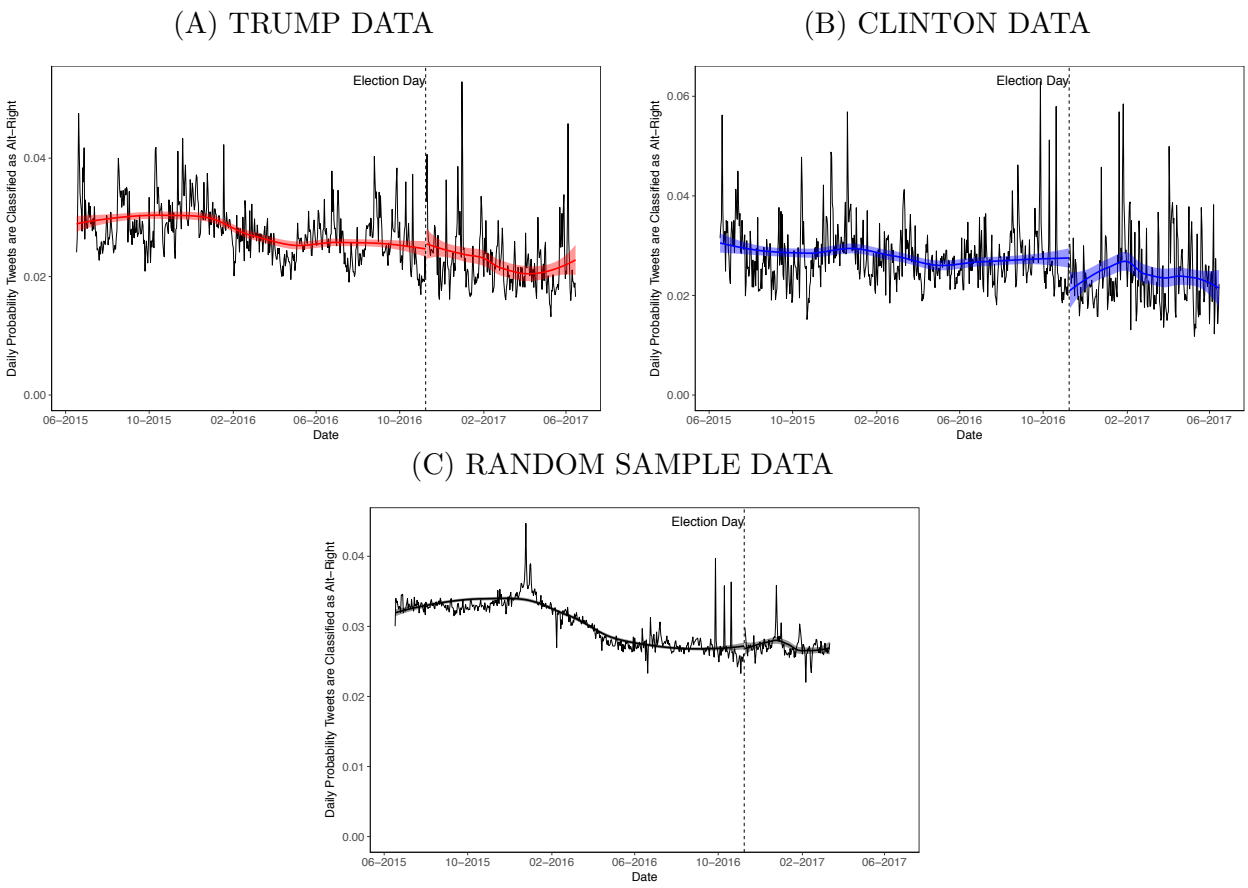

Figure 3. Effect of 2016 Election on Probability that Tweets are Classified as Alt-Right. This Figure shows the pre and post election trends from our ITSA regression models, plotted as local regression lines with loess smoothing and 95% confidence intervals. These trend lines are plotted against the average daily predicted probabilities that tweets in the Trump (A), Clinton (B), and random sample (C) datasets collected using Twitter's Streaming API between June 17, 2015 and June 15, 2017 are classified as belonging to alt-right subreddits.

Discussion

By highlighting the shortcomings of the conventional wisdom regarding the rise of hate speech on Twitter over the course of the 2016 election campaign and its aftermath, this application of our method demonstrates the importance of moving beyond short term or small scale datasets when studying online speech. Precisely because social media platforms like Twitter are so large and diverse, it is easy to find evidence—even relatively large-scale evidence—of almost any conceivable attitude or behavior. However just because a particular kind of discourse in which we may be interested is found to be especially prevalent online in a given moment, this does not necessarily mean the behavior has either increased or changed

over time.

This is particularly true given the bursty nature of online data, where topics trend briefly in response to events and then re-equilibrate shortly afterwards (tables of dates with the highest volume of hate speech and white nationalist rhetoric are displayed in the supplementary materials, Tables S5-S7). While there is certainly evidence of thousands of tweets containing hate speech and white nationalist rhetoric on Twitter over the course of the Trump campaign and in its aftermath, thus making it possible to use snapshots of this data as evidence of a "Trump effect," when we zoom out and examine the relative popularity of this language over time, we see that such content did not become more common in political discussions or among American Twitter users in general.

Our framework therefore offers key innovations for the study of online hate speech and online behavior more broadly that we hope will be adapted by scholars and practitioners alike. Firstly, our analysis employs two different but equally informative datasets: a collection of all tweets referencing the two candidates in the 2016 election and a random sample of 500,000 American Twitter users. This allows us to study online hate speech both in an explicitly Trump-related political context—where we might expect to see a "Trump effect"—and in a representative sample of American Twitter users. Second, by using both a machine-learning augmented dictionary-based analysis and a non-dictionary approach leveraging data from subreddits to classify hate speech, we avoid drawing conclusions that are biased by one particular classification strategy. Finally, by exploring changes in the volume, proportion, and number and proportion of unique users producing hateful content online, we ensure that our results are not driven by our measurement approach.

To be clear about the scope of our findings, this applicaiton of our method is limited to Twitter data. While Twitter is of course not the only platform on which hate speech may have proliferated during the election period, our approach enables us to test whether people on a large, popular social media platform were likely to be incidentally exposed to hate speech, rather than seeking it out on specialized platforms such as Gab, Voat, or particular communities on Reddit. While recent studies have begun to investigate the spread of this language on such alternative platforms (*36, 37*), this is beyond the scope of our research on the mass popularization of online hate. Additionally, although trolling and harassment of journalists on Twitter—particularly anti-Semitic attacks—were frequently reported over the course of the election campaign and may have contributed to the perception of increased widespread online hate in this period, our approach to measuring trends in online hate speech over time does not allow us to capture these specific incidents if they did not include references to Trump or Clinton or were not perpetrated by users in our random sample

of American Twitter users. Thus it is possible that hateful attacks on individuals could have increased over the time period we analyzed, even while hate speech was not increasing generally on Twitter or in discussions of the elections. However, this too would need to be carefully studied, as hateful attacks on individuals on Twitter were taking place before the summer of 2015 as well (*38*). Finally, our analysis of Twitter data tells us nothing about trends in hate crimes, bias incidents, or other offline events that have also contributed to the popular narrative of a "Trump effect" and deserve further study (*39, 40*).

By providing a new framework and empirical tools to study the over-time dynamics of hate speech and other discourse on widely used platforms like Twitter, our work offers a valuable contribution to the study of online hate speech and online behavior more broadly. Further, the research framework we have introduced in this paper—combining dictionary and non-dictionary based methods, marrying analyses of subject specific content with a general, random sample of users, and using multiple outcome measures—could be applied to the study of trends in many types of online discussion beyond hate speech and white nationalist rhetoric. Finding consistent results across two different data sets, employing two different approaches to measuring changes in the popularity of hate speech, substantially increases our confidence that we are drawing meaningful inferences about behavior on Twitter over time. Our hope is that by bringing new tools and data sources to the study of online hate speech, our work will enable academics, policymakers, and everyday citizens alike to better understand and address divisive social and political forces currently at play in the United States and in democracies around the world.

Acknowledgements

We thank Sean Kates for his feedback in designing our coding scheme, NYU Undergraduate SMaPP Research Assistants for their coding work, and Yvan Scher and Leon Yin for programming support.

Data and Materials Availability: The following code and data will be archived in the Harvard Dataverse Network: 1) R analysis code for training classifiers and analyzing our data; 2) Aggregate data for deriving our results and the summary statistics reported in the main body of the paper and the supplementary materials, in accordance with Twitters terms of service.

Funding: The authors gratefully acknowledge financial support for the NYU Social Media and Political Participation (SMaPP) lab, which is Co-Directed by Tucker, Nagler, and Bonneau, from the INSPIRE program of the National Science Foundation

(Award SES-1248077), the John S. and James L. Knight Foundation, the William and Flora Hewlett Foundation, the Bill and Melinda Gates Foundation, the Rita Allen Foundation, the Craig Newmark Foundation, the New York University Global Institute for Advanced Study, and Dean Thomas Carews Research Investment Fund at New York University.

Author Contributions: A.S. and J.T. designed the research plan and outline for the paper. A.S. conducted the statistical analysis and wrote the first draft of the paper. A.S, J.S., and B.P. designed and implemented the dictionary-based coding method. E.N. designed and conducted the non-dictionary based analysis and wrote the corresponding section of the Supplementary Materials. P.B., R.B., and J.N. contributed to the data collection and design of the analytic tools. J.T. revised the original draft, and all of the authors contributed to the editing of the text.

Competing Interests: Authors declare no competing interests.

Supplemantary Materials

- Materials and Methods
- Supplementary Text
- FIgures S1 to S39
- Tables S1 to S31
- Captions for Data S1
- External Databases S1
- References (S1-S7)

References

1. C. Fleishman, A. Smith, Exposed: The secret symbol neo-nazis use to target jews online, *Tech. Mic* (2016).

2. H. K. Vindu Goel, S. Frenkel, Making america hate again? twitter and hate crime under trump, *The New York Times* (2018).

3. P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, *ACM Computing Surveys (CSUR)* **51**, 85 (2018).

4. I. Gagliardone, *et al.*, Mechachal: Online debates and elections in ethiopia-from hate speech to engagement in social media (2016).

5. A. Olteanu, C. Castillo, J. Boy, K. R. Varshney, The effect of extremist violence on hateful speech online, *arXiv preprint arXiv:1804.05704* (2018).

6. L. Rainie, J. Anderson, J. Albright, The future of free speech, trolls, anonymity and fake news online (2017). Available at: http://www.pewinternet.org/2017/03/29/the-future-of-free-speech-trolls-anonymity-and-fake-news-online/.

7. A. Marwick, Are there limits to online free speech? (2017). Available at: https://points.datasociety.net/are-there-limits-to-online-free-speech-14dbb7069aec.

8. J. Guynn, Massive rise in hate speech on twitter during presidential election, *USA Today* (2016). Available at: http://www.usatoday.com/story/tech/news/2016/10/21/massive-rise-in-hate-speech-twitter-during-presidential-election-donald-Trump/92486210/.

9. ADL, Anti-semitic targeting of journalists during the 2016 presidential campaign (2017). Available at: https://www.adl.org/sites/default/files/documents/assets/pdf/press-center/CR_4862_Journalism-Task-Force_v2.pdf.

10. J. Desmond-Harris, The wave of post-election hate reportedly sweeping the nation, explained, *Vox* (2016). Available at: http://www.vox.com/2016/11/17/13639138/Trump-hate-crimes-attacks-racism- xenophobia-islamophobia-schools.

11. J. King, The year in hate (2017). Available at: http://www.vocativ.com/383234/hate-crime-donald-{T}rump-alt-right-2016/.

12. S. Milligan, 'a safe space for hate' (2017).

13. J. Weisberg, Trump's election led to 'barrage of hate' (2016). Available at: http://www.slate.com/articles/podcasts/{T}rumpcast/2016/11/how_the_alt_right_harassed_david_french_on_twitter_and_at_home.html$.

14. A. Okeowo, Hate on the rise after Trumps election (2017). Available at: http://www.newyorker.com/news/news-desk/hate-on-the-rise-after-{T}rumps-election.

15. M. Sidahmed, Trump's election led to 'barrage of hate' (2016). Available at: https://www.theguardian.com/society/2016/nov/29/{T}rump-related-hate-crimes-report-southern-poverty-law-center.

16. B. L. Ott, The age of twitter: Donald j. Trump and the politics of debasement, *Critical Studies in Media Communication* **34**, 59 (2017).

17. M. Barkun, President Trump and the 'fringe', *Terrorism and Political Violence* **29**, 437 (2017).

18. R. Cohen-Almagor, Fighting hate and bigotry on the internet, *Policy & Internet* **3**, 1 (2011).

19. N. D. Gitari, Z. Zuping, H. Damien, J. Long, A Lexicon-based Approach for Hate Speech Detection, *International Journal of Multimedia and Ubiquitous Engineering* **10**, 215 (2015).

20. J. Kaplan, *Encyclopedia of white power: A sourcebook on the radical racist right* (Rowman & Littlefield, 2000).

21. R. C. Fording, The political origins of extremism: Minority descriptive representation and the mobilization of american hate groups (2014). SSRN Scholarly Paper ID 3116303, Social Science Research Network, Rochester, NY.

22. L. Silva, M. Mondal, D. Correa, F. Benevenuto, I. Weber, "analyzing the targets of hate in online social media", *Unpublished Manuscript* (2016). Available at: https://arxiv.org/abs/1603.07709v1.

23. C. Tuckwood, The state of the field: Technology for atrocity response, *Genocide Studies and Prevention: An International Journal* **8**, 9 (2014).

24. N. J. Stroud, J. M. Scacco, A. Muddiman, A. L. Curry, Changing deliberative norms on news organizations' Facebook sites, *Journal of Computer-Mediated Communication* **20**, 188 (2014).

25. K. Coe, K. Kenski, S. A. Rains, Online and uncivil? Patterns and determinants of incivility in newspaper website comments, *Journal of Communication* **64**, 658 (2014).

26. M. Chau, J. Xu, Mining communities and their relationships in blogs: A study of online hate groups, *International Journal of Human-Computer Studies* **65**, 57 (2007).

27. M. Oz, P. Zheng, G. M. Chen, Twitter versus Facebook: Comparing incivility, impoliteness, and deliberative attributes, *New Media & Society* (2017).

28. T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, *Unpublished Manuscript* (2017). Available at: https://arxiv.org/pdf/1703.04009.pdf.

29. Z. Waseem, D. Hovy, Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter, *Proceedings of NAACL-HLT* pp. 88–93 (2016).

30. S. Belnik, Racial slur database (2017). Available at: http://www.rsdb.org/.

31. C. Tuckwood, Hatebase: Online database of hate speech, *The Sentinal Project* (2017). Available at: https://www.hatebase.org/.

32. ADL, Databse of hate symbols and terms, *Anti-Defmation League* (2017). Available at: https://www.adl.org/education/references/hate-symbols.

33. J. A. L. Bernal, A. Gasparrini, C. M. Artundo, M. McKee, The effect of the late 2000s financial crisis on suicides in spain: an interrupted time-series analysis, *The European Journal of Public Health* (2013).

34. J. L. Bernal, S. Cummins, A. Gasparrini, Interrupted time series regression for the evaluation of public health interventions: a tutorial, *International journal of epidemiology* (2016).

35. I. Gagliardone, D. Gal, T. Alves, G. Martinez, *Countering online hate speech* (UNESCO Publishing, 2015).

36. R. Nithyanand, B. Schaffner, P. Gill, Online political discourse in the Trump era, *Unpublished Manuscript* (2017).

37. S. Zannettou, *et al.*, What is gab? a bastion of free speech or an alt-right echo chamber?, *Unpublished Manuscript* (2018).

38. S. Parkin, Gamergate: A scandal erupts in the video-game community, *The New Yorker* **17** (2014).

39. S. Rushin, G. S. Edwards, The effect of president Trump's election on hate crimes (2018). SSRN Scholarly Paper ID 3102652, Social Science Research Network, Rochester, NY.

40. K. Müller, C. Schwarz, Making america hate again? twitter and hate crime under Trump, *Unpublished Manuscript* (2018).