Group Salience, Inflammatory Rhetoric, and the Persistence of Hate Against Religious Minorities

William Hobbs[*] Nazita Lajevardi[†] Xinyi Li[‡] Caleb Lucas[§]

June 30, 2021

Abstract

Recorded hate crimes surged at several points during and after the 2016 U.S. presidential election. Observers argued that hate crimes, especially against Muslims, increased due to inflammatory rhetoric. Did anti-Muslim hate crimes decline after the salience of Muslims and accompanying inflammatory rhetoric receded? Or did emboldening and organizational effects on hate crimes endure? Using data sources on online discussions (4chan, Gab, Reddit), media coverage (newspapers, Google Trends), and hate crime databases (ADL, CAIR, FBI), we show that a sudden mid-2017 decline in media discussion of Muslims and in online communities was associated with a large and sustained drop in anti-Muslim hate. At nearly the same time as these shifts, however, we observe evidence of an increase in violent hate crimes committed against Jews, and Granger causality tests demonstrate that week-to-week changes in online extremist speech targeting one group versus another also predicted subsequent shifts in hate crimes and bias incidents. Platform-level and within-individual analyses of online social media users suggest that increased anti-Jewish speech was partly driven by far-right communities and extremists who previously promoted anti-Muslim speech.

[*] William Hobbs, Cornell University.

[†] Nazita Lajevardi, Michigan State University.

[‡] Xinyi Li, Cornell University.

[§] Caleb Lucas, Michigan State University.

Recent studies in the United States have shown that politicians make overt racial appeals (Reny, Valenzuela, and Collingwood 2020; Valentino, Neuner, and Vandenbroek 2018; Hutchings, Walton Jr, and Benjamin 2010; Parker and Barreto 2014), and the media in turn repeats and reinforces this messaging by disproportionately providing sensationalized coverage about stigmatized groups (Entman and Rojecki 2001; Terman 2017). Further, extremists at times echo the rhetoric of these elected officials (Fording and Schram 2020; Mudde 2019; Siegel et al. 2021; Nithyanand, Schaffner, and Gill 2017). This state of affairs is especially troubling given that public expressions of prejudice online *and* offline are on the rise (Hopkins 2021; Schaffner 2020; Abrajano and Hajnal 2015; Jardina and Piston 2021; Tesler 2013; Müller and Schwarz 2018, 2020; Dancygier et al. 2021). While past work has suggested linkages between group salience, hate speech, and hate crimes (e.g., Van Dijk 1993; Chyzh, Nieman, and Webb 2019; Czymara 2020; Dugan and Chenoweth 2020; Awan and Zempi 2016; Williams et al. 2020), it remains an open question whether these links persist even once group salience and inflammatory rhetoric declines. Specifically, when group salience and inflammatory rhetoric decline, do hate crimes fall as well? And, if not, when might elevated hate and rates of violence against marginalized groups endure?

On the one hand, we may expect hate crimes to drop when specific groups are no longer subject to inflammatory rhetoric and salient in related media coverage. Without sustained coverage and salience, targeted groups and accompanying hate towards them might fade from fringe discourse, until they are reignited again in the national conversation. On the other hand, hate and hate crimes may become endemic to extremist communities and may endure without any further encouragement from mainstream sources. This might especially be the case given that fringe social media sites have the potential to harbor and concentrate extremists (Zannettou et al. 2018a; Grover and Mark 2019), and these cohesive extremist communities could perhaps substitute targets of hate and hate crimes from one group to another.

In considering whether inflammatory rhetoric targeting one group might set the stage for or persist in the targeting of another, we focus our analyses on hate targeted at Jews and Muslims in the U.S. These two groups each historically have experienced rampant discrimination (Cohen 2010; Beydoun 2018) and while the roots, determinants, and expressions of prejudice against each differ from one another (Gottschalk 2013; Meer 2013),[1] both were targeted by online extremists (Zannettou et al. 2020; Ekman 2015), and both experienced an increase in offline hate crimes following the 2016 U.S. presidential election.

We examine the persistence of hate and our target substitution hypothesis in the years following the 2016 U.S. presidential election by using a variety of data sources on hate crimes and bias incidents from both advocacy organizations and the FBI, media attention in the form of news reporting, and online conversations on both fringe and mainstream social media websites. Turning to multiple sources allows us to verify that the patterns we observe are not specific to a single online platform or hate crime reporting system. For one of the fringe platforms, we are able to evaluate within-user shifts in hate speech targets. Our analyses quantify shifts in ratios of group salience and hate across a variety of measures and data sets using a difference-in-difference framework. Specifically, we measure sudden shifts in the ratio of anti-Muslim to anti-Jewish hate crimes and hate speech, assess to what extent these shifts might be driven by a large decline in anti-Muslim hate on its own and/or a shift from anti-Muslim hate to anti-Jewish hate. We further consider whether hate originating from the same individuals might explain shifts in both anti-Muslim and anti-Jewish hate speech, and also whether week-to-week changes in anti-Muslim versus anti-Jewish hate speech predict changes changes in anti-Muslim versus anti-Jewish hate crimes and bias incidents.

Findings from these analyses both validate prior increases and demonstrate their durability. First, the results show that the salience of Muslims in both mainstream and fringe outlets tracked anti-Muslim hate crimes for declines as well as previously documented rises in hate crimes, suggesting that the majority of people might be activated *and* deactivated by prevailing rhetoric. Specifically, our analyses show that hate crimes and harassment declined when in mid-2017 Muslims were covered at lower rates in mainstream media, 'muslim' was less searched for online, Donald Trump more frequently mentioned ISIS rather than 'muslim' or 'Islam', and Muslims and Arabs were less mentioned in online communities. Second, when group salience declined, anti-Muslim hate speech on fringe social media sites was replaced by hate speech targeting Jews; that is, hate did appear to

[1] For example, while anti-Semitism and Islamophobia remain high, public attitudes towards Muslims have been comparably more negative. Muslims have also been visible in the mainstream media in recent years following the attacks on 9/11, with research finding that they were featured prominently and negatively in the news media disproportionate to their group size (Lajevardi 2021), while news coverage of Jews was relatively more positive (Bleich and Veen 2018). But, policy attitudes appear to have improved slightly following Trump's "Muslim Ban" (Oskooii, Lajevardi, and Collingwood 2019), and attacks against them declined once Trump took office. In contrast, while anti-Semitic attacks remained high during this period, they have risen dramatically since 2017.

still persist in a new form, at least among extremists. Finally, a shift in targets of hate appeared to occur not only on the same extremist online platforms, but even among the same users, suggesting that 1) that the same individuals account for a large fraction of extremist hate targeting multiple groups, and 2) extremists whose anti-Muslim expressions of hate are associated with instances and coverage of terrorist attacks quickly shift to other targets when a 'terrorism' justification for hate loses popularity. Together, these findings demonstrate that anti-Muslim violence and hate among non-extremists persists only through continued group salience and negative media coverage. This is also true for extremist communities, though they tend to replace old targets of hate for new ones.

Group Target Substitution and Enduring Emboldening Effects

Inflammatory rhetoric is on the rise in American politics, with politicians more regularly making explicit racial appeals and singling out stigmatized groups (Newman, Shah, and Collingwood 2018; Stephens-Dougan 2020; Tesler 2016; Valentino, Neuner, and Vandenbroek 2018; White 2007). This rhetoric is arguably not without consequence. Some scholarship has shown that voters reject inflammatory rhetoric styles (Costa 2020), but many others have argued that such discourse may have meaningfully shaped the role of prejudice in guiding voters' candidate and policy preferences (Schaffner 2020; Newman et al. 2020; Bursztyn, Egorov, and Fiorin 2020). In fact, there is a great deal of scholarly evidence that xenophobic and inflammatory rhetoric, particularly since the 2016 presidential campaign has emboldened white supremacists and has normalized a type of bigotry in online discourse (Giani and Méon 2019; Siegel et al. 2021; Nithyanand, Schaffner, and Gill 2017; Mathew et al. 2019), and in offline behavior (Müller and Schwarz 2018, 2020; Williams 2018).

While some social media sites have recently begun to crack down on inflammatory rhetoric, others—such as 4chan, 8chan, and Gab—have continued to provide platforms for speech ingrained with racism, sexism, religious bigotry, anti-LGBTQ+ attitudes, and anti-immigrant animus.[2] Some argue that these minimally moderated sites have attracted extremists and would-be extremists—which range from from neo-Nazis across the Western world to ISIS sympathizers (Burris, Smith, and Strahm 2000; Berger and Morgan 2015; Sunstein 2018; Conway et al. 2019)—away from more popular social media sites and brought them into closer contact with one another despite

[2] https://www.splcenter.org/news/2018/10/24/splc-announces-policy-recommendations-social-media-internet-companies-fight-hate-online

geographical distance (Hafez and Mullins 2015; Hawdon, Bernatzky, and Costello 2019; Bloom, Tiflati, and Horgan 2019).

With increasingly concentrated communities of extremists, we test whether group salience through media attention or elite political discourse could easily embolden or *activate* these communities. If so, an increase in group salience could thereby increase hate against all stigmatized groups, as emphasized by prior work (Giani and Méon 2019; Bursztyn, Egorov, and Fiorin 2020; Newman et al. 2020), but could also quickly change *which* groups they most target. This effect could be due to a type of agenda-setting (McCombs and Shaw 1972; Garimella et al. 2018) for extremists where stand-alone, far-right online communities might expose individuals to a more cohesive group of leaders and peers than they would otherwise encounter. In this case, hate might become endemic to a fringe online community, such that targets of hate can be sustained within a small, active group of extremists and without further mainstream encouragement.

Data and Methods

To test these possibilities, we examine the relationship between the salience of minority groups and the extent to which hate crimes and bias incidents target those groups (e.g. Stacey, Carbone-López, and Rosenfeld 2011; Byers and Jones 2007). Our analyses primarily measure long-term declines in these events along with sustained shifts in hate speech targeting one group to another, although we also consider week-to-week changes in both online hate speech and hate crimes/bias incidents. In our long-term analyses, we focus specifically on differences before and after 1) a decline in mentions of Muslims in news media, and 2) the Unite the Right rally in Charlottesville, Virginia (2017), a major White supremacist rally that could have injected anti-Semitic rhetoric into the national discourse. Our analyses consider temporary and week-to-week shifts in outcomes to assist in interpreting the long-term patterns we observe. Specifically, the analyses test whether a decline in mainstream salience is associated with a subsequent decline in extremist hate towards Muslims along with a shift in *extremist* attention from anti-Muslim to anti-Jewish hate. For clarity, Table 1 previews the research questions, answers, and tests employed in the analyses that follow.

Methodological Considerations

Although studying hate crimes is normatively and substantively important, interpreting associations between hate crimes and other phenomena is especially difficult. As Green and Spry (2014) note,

Table 1: Empirical Tests

Question	Answer	Test
1. Is there an association between group salience and inflammatory rhetoric?	Yes	Google searches, newspaper searches, and group mentions on mainstream social media (compared to changes in group mentions on fringe social media)
2. Does terrorism receive more media coverage if attributed to Muslims/Islam or if terror attacks occur during a presidential election campaign?	Yes	Timeline of global/U.S. terrorist attacks and media coverage, and President Trump's Twitter
3. Does a decline in the salience of Muslims correspond with a decline in inflammatory rhetoric in far-right and extremist online communities?	Yes	Analyses of posts on Gab and 4chan, and within-user analyses on Gab
4. Following the Unite the Right rally, did users on fringe social networking sites substitute anti-Muslim speech with anti-Jewish speech?	Yes	Analyses of posts on Gab and 4chan, and within-user analyses on Gab
5. Does the target substitution of Jews for Muslims across extremist online platforms mirror offline changes in hate crimes and bias incidents?	Yes	CAIR, ADL, and FBI bias incident and hate crime data (compared to changes in group mentions on 4chan and Gab), controlling for terror attacks and hate crime media coverage in Granger causality tests

however, literature in this area can still make progress by being in step with methodological advances in social science. Once hate crime statistics became more readily available in the 1990s, social science research examined over-time aggregate relationships and found that racially motivated crimes coincide with patterns of macroeconomic conditions (e.g. Green, Glaser, and Rich 1998) and demographic change (e.g. Green, Strolovitch, and Wong 1998; Esses, Jackson, and Armstrong 1998). And, related to the group salience argument considered here, recent work has shown that immigration and resulting change in minoritized group rank (i.e. by relative group population size) from 1990 through 2010 was associated with changing rates of hate crimes (Cikara, Fouka, and Tabellini 2020).

This scholarly work on hate should continue to keep pace not only with newly available data, but also shifts in the extremist ecosystem. Beyond their increasing importance for extremists, online extremist communities have provided for academics a new means to study the dynamics of extremist communities and their associations with hate and hate crimes. This data has already been fruitfully leveraged in other research (e.g. Zannettou et al. 2018a; Zannettou et al. 2020). We add to this body of work by (1) studying variation across multiple platforms, (2) documenting large, relatively quick shifts in attention paid by extremists that appear to track changing rates of hate crimes in ways not seen in mainstream social media and news outlets, and (3) assessing whether changing rates of hate speech and hate crimes against one group can be, at times, inversely linked to the targeting of another, perhaps through extremist coordination and 'target substitution', where extremist groups might substitute one target of hate for another.

But, even with new data and a broadened analytic lens, investigating the relationship between an increase in inflammatory rhetoric and an increase in hate remains challenging. Governmental hate crime data in particular relies on voluntary reporting (Freilich and Chermak 2013), and increases in hate crimes might be attributed to reporting artifacts. Declines in mainstream attention and inflammatory rhetoric are less studied than increases (but see Alrababah et al. 2021), yet declines bring new evidence to the study of hate speech and hate crimes. Specifically, studying declines allows us to more fully understand the association between group salience and hate crimes if declines are subject to some different reporting artifacts than increases. For example, if civil rights organizations and social movements lead to more complete reporting of hate crimes (McVeigh, Welch, and Bjarnason 2003), we would not expect mobilized actors to suddenly demobilize and for reporting to quickly become more incomplete again. Although extremists might quickly change targets, we might not expect members of *targeted* groups to suddenly forget what constitutes a hate crime or to quickly discount the importance of reporting such crimes, as anxieties about increased hate crimes and discrimination often have serious effects on lives (Samari 2016; Samari et al. 2020; Sediqe 2020) that are unlikely to suddenly disappear. Instead, we might expect chilling effects on reporting, where targeted groups are suddenly afraid to report, though we can address this particular artifact by assessing reports from *both* advocacy organizations and government sources concurrently.

Separately, the case of Unite the Right further provides leverage for studying the influence of rhetoric on hate crimes because the event was not associated with a terrorist attack or the action of an individual who happens to come from a particular ethnic or religious group. For example, rhetoric after a terrorist attack *might* generalize individual to group associations with the attack, thereby stigmatizing an entire racial, ethnic or religious group.[3] But, the absence of a such an attack would suggest that some form of social influence—either inflammatory rhetoric or White supremacist organizing—motivated shifts in hate crimes. In particular, White supremacist chants at Unite the Right reintroduced into mainstream conversation anti-Semitic rhetoric and slogans, which are thought to originate from a conspiracy theory that is also anti-Muslim.[4] At the same time, it is possible that this rhetoric preceded the Unite the Rally in extremist communities, and so we evaluate whether the rally was an abrupt turning point or a manifestation of shifting extremist targets in our analyses.

Data

Measuring Group Salience: News Media, Google Searches, and (Mainstream) Social Media

To measure changes in group salience of Muslims and Jews, we turn to three databases: an original collection of U.S. newspaper articles, Google Search Trends, and Reddit. First, our media analysis relies on news articles published by newspapers in the United States. We collected this data with the LexisNexis API and searched the entirety of the database's global news repository for articles published between 2014-2018 that included terms that identify these groups. We use 'muslim' for the Muslim corpus (N = 74,131), 'jew' for the Jewish corpus (N = 57,776), 'white supremacy' for the White supremacy corpus (N = 4,078), 'terrorism'/'terrorist' for the terrorism corpus (N = 97,996), and 'hate crime' for the hate crime corpus (N = 16,200) which was later subset to articles that specifically mentioned 'muslim' or 'islam' (N = 3,410) or 'jew', 'judaism', or 'semit' (to capture 'anti-Semitic'/'anti-Semitism') (N = 3,646). We removed articles published by foreign outlets by filtering newspapers that are not recorded in the Library of Congress' Chronicling America project's U.S. Newspaper Directory, which maintains a list of over 20,000 U.S. newspapers. This also ensures non-newspaper documents are not represented in the data. The most common newspapers represented in the data are *The New York Times*, *The Washington Post*, *The Boston Globe*, *The Los Angeles Times*, and *The Chicago Tribune*.

Next, for our online search analyses, we use Google Trends, which displays aggregated numbers

[3] There is evidence to suggest that the occurrence of inflammatory rhetoric after a terrorist attack instead depends on prevailing social norms (Álvarez-Benjumea and Winter 2020).

[4] https://www.washingtonpost.com/graphics/2017/local/charlottesville-timeline/

of private searches by Google users. These data are provided pre-standardized so that the largest value in a period is set to 100, and *each* of the trends we analyze will be standardized to have their largest value equal to 100. In doing so, we consider relative changes in attention rather than absolute ones, and these results do not represent the overall popularity of a search. While we cannot infer the number of unique searches or users from this data, these data do indicate shifts in group interest and salience over time.

Finally, we also assess group salience using posting activity on Reddit, which is a social media website comprised of subforums ('subreddits'). Each of these communities focus on specific topics, such as politics or traveling. We elaborate on the site's form and moderating policy in the SI. Relative to the extremist platforms we also analyze, Reddit represents the "mainstream" social media website in our analysis. We assemble data from the site using `pushshift.io`. Pushshift ingests the entirety of Reddit in real-time and makes it available to the public through an API. Our analyses are conducted using all posts between 2014–2020 made by a random sample of 5 million users. Analyses in the main text focus on posts from late 2016 through August 2018 – all of our social media (Reddit, Gab, 4chan) and hate crime/bias incident databases (ADL, CAIR, FBI) contain data for this period. Further information regarding these data collection processes and the resulting data is available in the SI.

Measuring Disproportionate and Sustained Coverage of Muslims/Islam after Terrorist Attacks

Beyond studying news coverage alone, we also evaluate to what extent rates of U.S. news coverage mentioning Muslims and Islam merely tracked U.S. and global terrorist attacks, or if news coverage was 1) more prolonged if descriptions of a terror attack mentioned 'muslim' and/or 'islam', and 2) more extensive during the U.S. 2016 presidential campaign. As a conservative test of 1), we exclude mentions of the 'Islamic State.'

We also consider whether the large shift in the salience of Muslims that is the focus of our paper *followed* or *preceded* shifts in terrorist attacks. It is important to note here that in studying whether hate speech and hate crimes can shift from Muslims to another group is itself a test of the role of terrorist attacks in inciting hate. Hate that is solely motivated by terrorist attacks should presumably not shift to a group that has not been associated with terrorism in the news media or in public discourse.

For this analysis, we use the Global Terrorism Database (GTD) (LaFree and Dugan 2011), one of the most complete sources of data on terrorism. The GTD classifies an event as "terrorism" if it is violent, intentional, and conducted by a sub-national actor. At least two of the three following requirements must also be met: conducted outside of legitimate warfare, conducted to attain social or religious goals, and be used to intimidate individuals beyond the immediate victims. Because mainstream attention to terrorism in the U.S. is high even when events take place outside the country's borders (Nacos 2016), we study both terrorist attacks in the United States and attacks globally using the GTD.

In our analyses comparing week-to-week changes in extremist speech online with hate crimes and bias incidents (see 'Time Series Models' section below), we report estimates with and without controls for terror attacks. The Granger causality analysis described below covers the period December 2015 through December 2018 (the U.S. presidential election starting from the month of Donald Trump's 'travel ban'/'Muslim ban' press release through the end of our data) and analyses of contemporaneous media coverage of terror attacks compares the presidential campaign period (December 2015 through November 2016) to January 2014 through December 2018.

Measuring Inflammatory Rhetoric: Extremist Social Media and (soon-to-be) President Trump's Tweets

Next, we measure inflammatory rhetoric employed by extremists online. We focus on two extremist social media sites: Gab, which we source from `pushshift.io`, and 4chan's notorious 'Politically Incorrect' subforum ('/pol/'), which we source from `4plebs.org` as archived on the Internet Archive. Gab is a website modeled on Twitter that emphasizes its commitment to 'free speech' and lack of moderation. The site is known for attracting alt-right users and the widespread use of hate speech (Zannettou et al. 2018a). Data for Gab was available for August 2016 (the beginning of the site) through August 2018, though the site gained more popularity among the far-right after the 2016 election.[5] 4chan is a website that hosts a collection of anonymous message boards. The 'Politically Incorrect' board ('/pol/') is characterized by rampant hate speech and xenophobia. Data for 4chan was available from December 2013 through December 2019. We identify group mentions using

[5] See, for example, Figures B.5 and B.6 showing large end-of-2016 increases in numbers of posts, users, and numbers of posts mentioning Muslims or Arabs and Jews.

a random sample of Gab and 4chan posts that we tasked workers on the crowdsourcing website Amazon Mechanical Turk to hand label.[6] Because we do not expect extremists—or even mainstream social media users—to distinguish between Muslims and Arabs, we asked coders to label 'Muslim and/or Arab' mentions. As a comparison for this hand labeling, we also measure group mentions using keywords 'muslim', 'islam', and/or ' arab ' and 'jew' and/or 'judaism'.

For our extremist social media analyses, we consider mentions of racial, ethnic, or religious groups on *fringe* platforms (Gab and 4chan) as a measure of the salience of that group among extremists, as well as possible hate speech. We focus on salience given the difficulty of labeling any specific post as hate speech without considering its broader context (a user's self-description or full post history, for example), and because some users may falsely claim to be a member of a group when making negative statements about it. Nonetheless, to assess whether the vast majority of posts were negative mentions about groups, we coded the crowd-sourced hand labeled group mentions for whether the text contained an "unambiguously negative" statement about the group. This labeling did *not* count comments with seemingly negative content that might also be neutral or positive. In that coding, 90% of the posts about Jews, Muslims, and/or Arabs were considered unambiguously negative by at least one of the two coders, and 62% of the posts by both coders. We show in SI Figure B.7 that neutral or positive statements about Jews tended to occur prior to Unite the Right.

Hand labels were sufficient for the aggregate, monthly analyses, but labeled a small number of posts per week and labeled posts for only a small fraction of users. To power all of our analyses, especially weekly and within-user analyses, and to avoid relying solely on keyword analyses, which could conceivably miss many references to Muslims/Arabs and/or Jews, we predicted group mention labels for the full corpus using fine-tuned BERT (Bidirectional Encoder Representations from Transformers) models (Devlin et al. 2018). BERT is a pre-trained machine learning model that conditions on context in both directions around a token in textual data. This results in more contextualized representations. We provide the full model specification details in the SI.[7] In the main text Figure 4, we compare these predicted labels (means of predicted probabilities by month) to an approach using only keywords. Hand labels identify more mentions (see Figures B.5 and B.8),

[6] Figure B.2 in the SI details the full instructions given to workers.

[7] See Section B.3.

but trends in the ratio of ‘Muslims or Arabs’ to ‘Jews’ mentions over time are essentially unchanged.

As a further robustness check for the analyses of extremists websites, we repeat the analyses for Gab after adding predicted hate speech labels trained on a dataset of hand labeled hate speech – the Gab Hate Corpus (Kennedy et al. 2018). In this data, even highly trained coders disagreed on hate speech labels (Fleiss’ kappa inter-rater reliability 0.2 to 0.3 for the hate speech corpus compared to 0.6 to 0.7 in our group labels), and supervised models trained on it were also noisy (see Section B.4). However, we nonetheless replicate our aggregate analyses using it. As in group mentions, we again predicted hate speech labels using a BERT model (see SI section B.4). Predicted probabilities from these models were multiplied by the group mentions predicted probabilities to designate group mentions that contained hate speech.

One lingering question may be the expressed identities of those driving discussions on Gab. Gab is widely considered an alt-right and White nationalist platform, but we coded user descriptions and timelines to confirm this. Coders evaluated whether the user’s content/social media had an alt-right or related (e.g. White nationalist) branding. We find that 82% of the accounts appearing in our group mention sample were labeled alt-right by two of two coders, and 93% by at least one of two coders. We are unable to repeat this user-level analysis on 4chan /pol/ because posts on the site are anonymous.

Finally, we downloaded tweets from Donald Trump’s @realDonaldTrump Twitter account using the GetOldTweets3 Python library. Using these tweets, we counted the number of times that @realDonaldTrump mentioned: 1) ‘muslim’ and/or ‘islam’ and 2) ‘isis’ and/or ‘isil’ (i.e. by acronym rather than the ‘*Islamic* State’). This analysis assesses whether Donald Trump associated Muslims and Islam with terrorist attacks by ISIS more during the time periods we consider in other analyses, especially the news media analysis linking terrorist attacks to unusually elevated mentions of Muslims and terrorism during the U.S. presidential campaign.

Measuring Experienced Hate: Hate Crime and Bias Incident Databases

Finally, we rely on three unique databases to measure hate crimes: the FBI’s Uniform Crime Reports (UCR), as well as databases assembled by the Anti-Defamation League (ADL) and the Council on American-Islamic Relations (CAIR). The ADL and CAIR record anti-Jewish and anti-Muslim hate crimes respectively. Both ADL and CAIR emphasize a victim’s report that bias existed in an incident to code whether the event constituted a hate crime or bias incident. Both groups actively solicit the public to report such events to their organizations, which is a primary way in which new observations are added to the datasets. ADL also examines police reports for evidence of these events. We obtained the CAIR dataset directly from the organization and use publicly available ADL data.[8] We downloaded the FBI UCR data from their public website.[9]

In our analyses of the advocacy organization data, we consider the hate crime and bias incidents from each organization separately and we also consider abrupt changes in *relative* incidents across the two organizations over time. In studying these two sources, we limit our analyses of the ADL and CAIR data to types of bias incidents and hate crimes appearing in both data sets.[10] At the time of this analysis, ADL does not release complete hate and bias incident data for years prior to 2016, and so the analyses begin on January 2016.

We use sources beyond the FBI data because the FBI data depend on voluntary police reporting, yielding both under- and over-reporting concerns (Freilich and Chermak 2013). In SI Figure B.10 and Table B.23, we compare these sources, and demonstrate abrupt drops in reporting to the FBI compared to both advocacy sources after the 2016 presidential election.

To be included in the FBI database, a crime must be filed with a local law enforcement agency. Beyond this, sub-state jurisdictions do not always follow mandates set forth by federal law in their reporting to the FBI (Barth et al. 2019; McVeigh, Welch, and Bjarnason 2003; Freilich and Chermak 2013). Police also sometimes fail to record incidents as hate crimes even when there is clear evidence of bias (Iannelli 2018). In studies on the general public, there is strong evidence that factors much more common than violent extremism, such as agreement with White and Christian nationalist views, predict greater uncertainty in attributing extremist attacks to hate (Leander et al. 2020).

However, the FBI data contains more detailed information on the types of crimes committed than the ADL and CAIR,[11] and is potentially more reliable for particularly violent crimes than other crimes. As shown in SI Figure B.10, government data records more instances of physical violence than advocacy organizations. While the ADL and CAIR data are likely to be more complete sources of hate crime and especially bias incident reporting over time, and – critical for many of the

[8] https://www.adl.org/education-and-resources/resource-knowledge-base/adl-heat-map.

[9] https://crime-data-explorer.fr.cloud.gov/downloads-and-docs.

[10] See SI Tables A.5 and A.6 for lists of bias incidents and hate crimes recorded by both organizations.

[11] Both ADL and CAIR record physical violence, but, for example, do not distinguish types of assault, such as aggravated assaults that cause serious bodily harm or involve the use of a deadly weapon.

analyses here – less subject to chilling effects on reporting given their advocacy role, incidents that lead to deaths, hospitalizations, and/or that require calls to emergency services (e.g. paramedic or firefighter response) would appear very likely to be investigated by police. That is, the crimes are likely to be reported, even if not necessarily classified as a crime for which there is evidence of bias against a protected class.

Given this additional information and reporting concerns (affected by trust in local police and government, as well as voluntary local reporting to the FBI), we subset the FBI data to likely to be reported and extreme crimes: aggravated assault, manslaughter, murder, arson, and kidnapping. SI Table A.7 reveals that aggravated assault and arson make up the vast majority of crimes in this subset.[12] This analysis is also substantively relevant since the severity of hate crimes is an outcome of interest beyond the occurrence of hate crimes. If we believe that increases in hate crimes—especially anti-Jewish hate crimes—originate from extremist communities, then we might expect these crimes to be more violent as well.

Table B.10 and Figure B.23 in the SI further compare advocacy records with hate crimes reported to the FBI. Although CAIR and ADL record fewer assaults than the governmental data, the advocacy organizations maintain more complete records of vandalism and harassment, especially harassment not rising to the level of criminal intimidation. We further note that the data sources are largely correlated, with the exception of an abrupt increase (beyond the increase in governmental data) in reports to advocacy organizations after the 2016 election and the 2017 change in presidential administration.

Methods

Each of our analyses study the group salience of Muslims, anti-Muslim rhetoric, and anti-Muslim hate crimes. As mentioned in the data section above, we further include anti-Arab rhetoric and, where available, anti-Arab hate crimes in our social media analyses because we do not expect anti-Muslim individuals to distinguish between Arabs and Muslims (Beydoun 2013). We compare these measures to the group salience of Jews, anti-Jewish rhetoric, and anti-Jewish hate crimes. In most of our analyses, we estimate these changing rates of anti-Muslim or Arab and anti-Jewish hate before and after Unite the Right using a difference-in-difference design. We plot trends for each

[12] Analyses of ratios of hate crimes with this data used pseudo-logs (adding 1 to each weekly or monthly count before calculating a ratio or logging) to account for zeros in some periods.

primary analysis. In more formal models, we use methods that account for outcomes that are count variables, which assess changes in relative ratios of attacks.

However, a difference-in-difference alone provides little information on the target substitution hypothesis, since multiple patterns of shift hate could result in the same estimates from difference-in-difference. Because of this, we assess evidence for and against two distinct patterns of hate speech and hate crimes, the second of which is more consistent with target substitution: 1) a decline in hate targeting Muslims or Arabs and no change in hate targeting Jews, and 2) a decline in hate targeting Muslims and a corresponding *increase* in hate targeting Jews.

Beyond this, we use individual-level data from one of the extremist sources (Gab) to evaluate whether the same *individuals* might have switched from targeting Muslims or Arabs to Jews. Here, our analysis is slightly more complex (although it is made simpler by the nearly panel nature of the social media data): we use an errors-in-variables approach described below to estimate changes in the associations between (negative) mentions of Muslims and (negative) mentions of Jews.

Finally, we consider whether weekly shifts in online hate speech predicts future hate crimes using Granger causality tests, including tests controlling for terror attacks and media coverage of hate crimes.

Difference-in-Difference Models and Visualizations

Given that a majority of our analyses use the mechanics of a difference-in-differences design, we initially imagine that anti-Muslim or Arab and anti-Jewish extremists form two separate groups. Put simply, these analyses compare hate targeting Muslims compared to Jews before and after Unite the Right. We further display month-by-month changes in outcomes of interest, which constitutes an event history analysis.

We qualify the description of our analyses above by saying the "mechanics of a difference-in-differences" design because we expect that both of these groups were "treated" at nearly the same time. That is, both groups were affected by major shifts in the political context in mid-2017. When using difference-in-differences in this context, these estimates then capture the combined differences in movements for both groups. We assume parallel trends for hate crimes targeting Muslims and Jews, a dynamic we assess below by graphing the trends and the relative changes in the trends. The models include indicators for before or after December 2015 and before or after August 2017.

Consequently, they estimate shifts relative to the December 2015 through July 2017 period.

Given that these estimates capture the relative change in combined shifts in hate across groups, we further display separate time series for both anti-Muslim or Arab and anti-Jewish hate to allow readers themselves to visually assess changes in counts of news coverage, mentions on social media, and hate crimes. We are able to display monthly counts of outcomes with relatively high precision because these counts represent the activities of tens to hundreds of thousands of people. We also replicate results for Gab and Reddit using counts of unique *users* mentioning a group in the SI (see Figure B.6), as users on social media sites tend to display heavily skewed levels of activity. A small fraction of users can consequently account for a very large fraction of overall activity.[13]

Models on aggregate data use count models, specifically quasi-Poisson regression. For social media analyses, dependent variables in these regression models use the monthly means of predicted group mentions, though similar results are achieved with hand labels, supervised predictions, and keywords. For Reddit, the mainstream social media site, we did not collect hand labels, and we estimate changes in mentions of groups on Reddit using keywords only.

From count models, we report exponentiated coefficients: incidence rate ratios. All confidence intervals are 95% confidence intervals and are estimated by profiling. We also provide models fit to the FBI UCR hate crime data in the SI[14] that control for overall racial/ethnic hate crimes and month of year fixed effects. For models combining the shifts across both groups, we report estimates from linear regressions on logged ratios. For the data we analyze here, these estimates are not meaningfully different from a quasi-Poisson regression with a difference-in-difference specification, but have the advantage of clearly stating that the model estimates the difference between two groups of equal interest, rather than a change for only one group compared to reference group.

Within-User Model

The data from the extremist website Gab contains usernames that individually identify users on the site. We are therefore able to assess within-user changes in mentions of Muslims and Jews using fixed effects or, as we will, first-differences. However, for this analysis, we also expect that whether an individual mentions Muslims and Arabs in a particular period on social media is a noisy measure of their long-run propensity to mention Muslims and Arabs. Including rates of posts

[13] https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/

[14] See Section B.9.2.

mentioning Muslims and Arabs without adjustment as an independent variable in a regression can underestimate the extent to which the same extremists drive *both* declines in anti-Muslim or Arab speech and increases in anti-Jewish speech.

Our target substitution estimates should then account for how consistently we can expect a user to talk about even the *same* group over time, so that we can assess the magnitude of shifts from mentions of one group to another relative to that baseline. To make this comparison, our within-user estimates of target substitution draw on errors-in-variables techniques for panel data (Griliches and Hausman 1986) to estimate an adjusted effect and calculate appropriate standard errors. In this technique, two-stage least squares regression allows us to compare the changes in cross group coefficients to same group coefficients over the same time periods. For example, this approximates the coefficients we observe when the coefficient for changes in waves 2 and 3 with waves 1 and 4 *for the same group* is 0.10 and the coefficient for changes of mentions *one group* in waves 2 and 3 *with another* in waves 1 and 4 is -0.04, then the final target substitution coefficient will be -0.40.

Because we are interested in estimating a unique effect for a specific period and because our target substitution hypothesis anticipates that the analyzed extremists tend to express hate targeting all minority groups, we further allow for the possibility that the shifts in mentions of groups are correlated within individual over time absent any major event. We estimate the target substitution effect around Unite the Right using an interaction term for June to August (covering both a decline in anti-Muslim salience in the news media and the Unite the Right) – instrumented using May to September – compared to equally spaced periods from December 2016 through October 2018. Like the rest of our analyses, we log the fraction (or number – we display both models) of posts about a specific group to estimate a change in the ratio of group mentions over time. We are able to log these fractions and counts by using predicted probabilities from our supervised label prediction model – all of which are non-zero. We report robust standard errors clustered at the user level.

Time Series Tests

To supplement our analyses on relative shifts in anti-Muslim or Arab and anti-Jewish rhetoric in mid-2017, we further conduct Granger causality tests to: 1) assess whether the U.S. was more likely to sustain mentions of Muslim and Islam (and mentions of both Muslims/Islam and terrorism) after terrorist attacks during our study period, and 2) whether hate speech online predicted subsequent offline hate crimes and bias incidents. In these Granger causality tests, as standard, we control for a lag (or, as appropriate, multiple lags) of the dependent variables and test whether the addition of our variables of interest significantly improve predictions of future outcomes. For example, for the hate speech and hate crimes models, we test whether the previous two weeks (see below) of hate speech on fringe social media sites predict the following week's recorded hate crimes and bias incidents significantly better than a model that only includes the previous two weeks' recorded hate crimes and bias incidents.

Our Granger causality test uses an approach by Toda and Yamamoto (1995) which provides valid results whether or not the studied time series are cointegrated. These tests can be used for difference stationary time series (all series considered here were order 1 – see SI section B.8), and 'differencing' is accomplished using additional (untested) 'lags' for both levels of the dependent variable and levels of the independent variable(s) in the null model. These models further include linear trends and intercepts; that is, the added variables must significantly contribute to predictions beyond a model that includes lag (or multiple lags) of the dependent variable, a linear trend, and an intercept. As usual with these methods, the number of considered lags for each model (the same number of lags for both the dependent variable and the added independent variable) was selected using the Akaike information criterion (AIC). This lag was one week in the terrorism coverage models, and two weeks in the hate speech and hate crimes models.[15]

In models that test the association between online hate speech and hate crimes, we further display findings from tests that included logged counts of terror attacks and media coverage of hate crimes (logged ratio of anti-Jewish to anti-Muslim hate crime articles) in the null model. These models evaluate whether lagged online hate speech significantly predicted hate crimes and bias incidents above and beyond a model that included lagged hate crimes, lagged terror attacks, and lagged hate crime news coverage.

Because the ADL records many bias incidents only at the monthly level and lists those incidents as occurring on the first day of the month, we necessarily limit our analyses to ADL and CAIR hate crimes and bias incidents that were recorded on days *other than* the first day of the month.

[15] The two week lag for hate speech and hate crimes tests is drawn from the largest model which included controls for terror attacks and for the ratio of media coverage of hate crimes committed against Jews to media coverage of hate crimes committed against Muslims.

The ADL began recording some incidents at a daily level in 2017, and so our advocacy organization database analysis begins in January 2017 and continues through the end of the Gab data in August 2018. The FBI hate crime data covers a longer period, but we have social media data for all of 2015, 2016, and 2019 only for 4chan. Repeating the time frames in our other news and social media analyses, we test associations between 4chan group mentions and FBI recorded hate crimes for December 2015 through July 2017 and August 2017 through December 2018.

As in other analyses of social media data, these models use means based on predicted probabilities of group mentions on Gab and 4chan. Logged ratios of mentions were combined into a single variable for both Gab and 4chan by adding the logged ratios from each site.

Results

Below, we first turn to newspaper coverage and web searches to evaluate 1) *when* we might expect Jews and Muslims (and Arabs) to be made more (and less) salient to extremists, and 2) whether the salience of (anti-Jewish) White supremacists after 'Unite the Right' might have *displaced* terrorism coverage and mentions of Muslims in the media, or if a decline in the salience of Muslims might have *preceded* Unite the Right (leading to a possible vacuum effect).

We consider news coverage and web searches for mentions of 'white supremacy' rather than 'Jews' alone because,[16] even if mainstream attention does not directly drive extremists' conversations about specific groups, it could still direct extremists to the same far-right communities or leaders. This could lead to shifts in Islamophobic versus anti-Semitic rhetoric through extremists' identification with newly discovered far-right movements and neo-Nazi groups that have long targeted Jews. This could also resemble the amplification of fear and anger in anti-Muslim organizations by the mainstream media after the September 11th attacks (Bail 2012).

Next, we document trends in extremist speech referencing Jews and Muslims or Arabs. While we focus on sudden shifts around "Unite the Right," similar to work on anti-Jewish and anti-Black content on 4chan /pol/ and Gab (Zannettou et al. 2020), we also demonstrate that Islamophobic rhetoric increased in the early period of the 2016 presidential campaign.

Finally, we analyze within-user and over-time communication on social media to test whether the same individuals shifted from speech targeting one group to speech targeting another group.

[16] See news coverage mentioning 'Jews' in SI figure B.1.

For the last test, we argue that *some* of the shift in targeted groups can be attributed to shifts in attention originating from the same extremists, and we estimate a within-user substitution ratio for this effect.[17] This test goes beyond establishing that when mentions of Muslims or Arabs declined mentions of Jews increased; it also evaluates how much of this shift might be due to shifts *from one group to another* by the same individuals.

After these hate speech analyses, we conclude by testing whether the shifts in hate speech were associated with shifts in hate crimes and bias incidents. We conduct tests for these effects in both the advocacy organization and, for more extreme and likely to be investigated hate crimes, government data.

Group Salience: Mainstream Media and Web Searches

The news media is perhaps the primary lens into how groups are portrayed to the mass public, and so tracing the media coverage of groups can serve as an objective indicator of group salience (Lajevardi 2021). Thus, we begin by first documenting shifts in media coverage of Muslims and their associations with Google Trends. This analysis then compares the shifts in media coverage of Muslims to coverage of White supremacists. We display searches and media coverage for 'jew' and 'terrorism' in SI Figure B.1.

Figure 1 displays counts of news articles containing Muslim and White supremacy keywords (see SI), as well as Google Trends for keywords 'Muslim' and 'White supremacy'. The news articles are standardized to the same scale as Google Trends; that is, we divide by the maximum value and multiply by 100. The figure shows that mentions of Muslims in the news drop in the month leading up to the Unite the Right rally in Charlottesville, Virginia. For both Muslims and White supremacy, news coverage and Google Trends are highly correlated.

The decline in mentions of Muslims was large and sustained (quasi-Poisson IRR: 0.50%, 95% CI: 0.41 - 0.61, see SI Table B.1 for all model coefficients). In contrast, mentions of Jews did not increase in the media (see SI Figure B.1, mentions of Jews declined after the 2016 election) and the increases for White supremacy were short-lived, implying that large, long-term shifts in the salience of White supremacy and anti-Semitism would have to be sustained by activity in fringe

[17] Of the three sites we study, only Gab has usernames and is minimally moderated. We cannot conduct the same test for 4chan (anonymous) or Reddit (explicit hate speech is regularly deleted by community moderators, or the community will be banned).

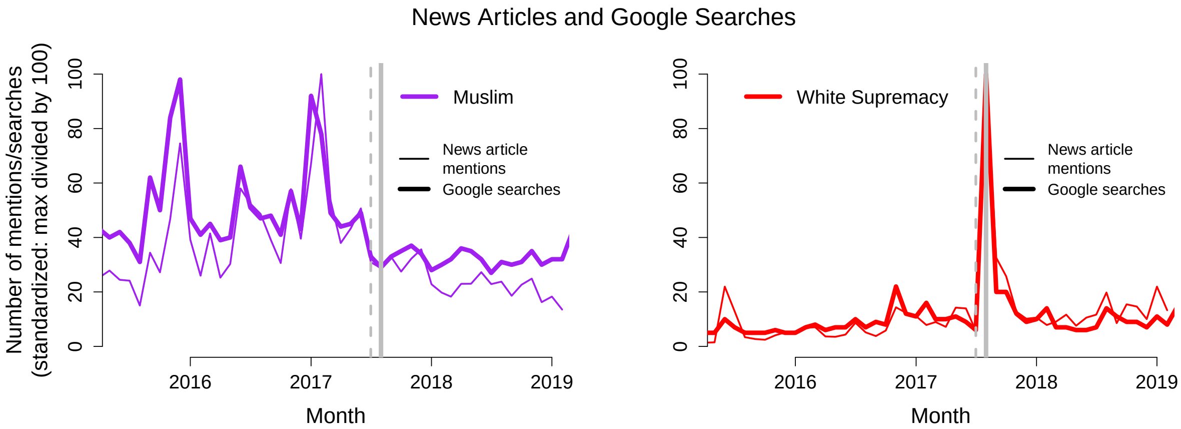


Figure 1: In news data and in Google Trends, we see declines in Muslim searches about a month prior to Unite the Right. This data suggests that the *mainstream* salience of Muslims declined before the far-right event – and, perhaps, that extremists might also have been less likely to target Muslims during the event than they might have been only months earlier. This figure shows a large (relative) increase in the salience of white supremacy (and, presumably, white supremacist groups) at a time when news reports of Muslims, as well as terrorism (see SI Figure B.1), were declining. From 2014 through 2018, 74,131 articles contained text matches for 'muslim' and 4,078 for 'white supremacy'.

communities instead.

Overall, these tests suggest that the mainstream salience of Muslims declined *one month prior* to Unite the Right, while the salience of White supremacists increased the same month as Unite the Right. That is, the decline in salience of Muslims and the increase in salience of White supremacists are not contemporaneous in the mainstream media, and increased attention given to White supremacists probably did not displace terrorism coverage and mentions of Muslims in the media – instead, a decline in the salience of Muslim preceded Unite the Right and, as we will argue, may have left a space later filled by other forms of inflammatory rhetoric. Concretely with this analysis, we identify the time frame—July to August 2017 for initial effects, with perhaps some continued shifts later in 2017 and into 2018—where we would expect shifts among extremists.

Mentions of Muslims in the U.S. Media and Globally Recorded Terrorist Attacks

We continue our analysis of media coverage and group salience by next considering associations between terrorist attacks and media mentions of Muslims and Islam. In Figure 2, we display logged number of terrorist attacks linked to keywords 'isis', 'isil', 'muslim', and/or 'islam', comparing global attacks with attacks occurring in the United States. Rates of global attacks in this figure largely reflect elevated rates of terrorist attacks by ISIS in Syria and Iraq in 2014 through 2017. These terror attack data include 12 attacks occurring in the United States.

Notably, global records of terror attacks did decline in late mid to late 2017. However, not only do ‘Muslim’ Google searches and media mentions decline before (global) terrorist attacks decline in 2017/2018, but media mentions of Muslims were more closely associated with terrorist attacks during the 2016 presidential campaign period.

Table 2 shows the results of a Granger causality test, which tests whether weekly shifts in coverage of terrorism depended on whether the attack was associated with Muslims or Islam. Global deadly terror attacks did not in general predict future coverage of terrorism. However, deadly terror attacks in the GTD whose attack descriptions specifically contained mentions of ‘muslim’ or ‘islam’ significantly predicted a future rise in numbers of articles about terrorism and Muslims/Islam (keywords ‘muslim’ and ‘islam’, excluding ‘Islamic State’).

We did not find the same association for terrorist attacks occurring in the United States. However, terror attacks in the United States are sufficiently rare that this particular test might be under-powered. Terror attacks in the U.S. are a very small fraction of the total number of attacks worldwide. Instead, as shown in SI table B.12, contemporaneous terror attacks in the United States were associated with increased contemporaneous coverage, at least during the U.S. 2016 presidential campaign. This association was no different when accounting for the number killed by terror attacks (SI table B.13).

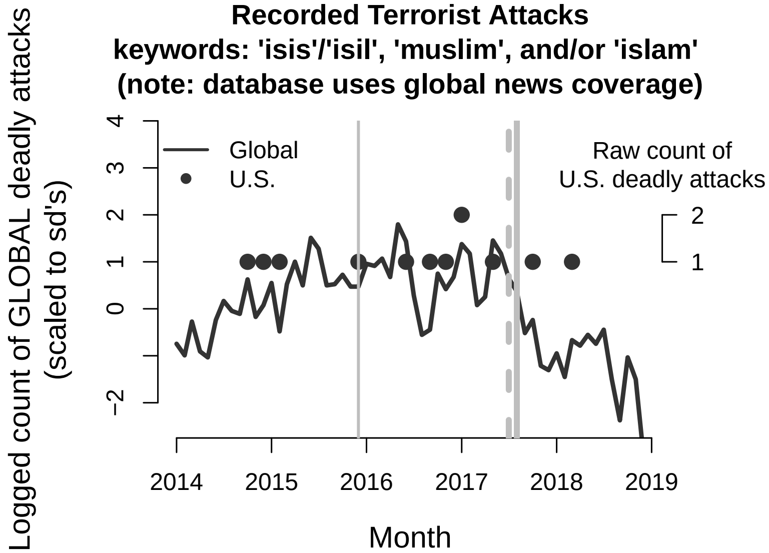


Figure 2: Global and U.S. terrorist attacks from GTD database (keywords: ‘isis’/‘isil’, ‘muslim’, and/or ‘islam’)

In Figure 3, we display the number of tweets from Donald Trump’s Twitter account (@realDonaldTrump) that mentioned ‘muslim’ or ‘islam’ compared to ‘isis’ or ‘isil’. This analysis reinforces the previous

global deadly terror attacks → subsequent articles mentioning terrorism in U.S. news

‘muslim’/‘islam’ terrorism	all terrorism
$\chi^2 = 6.6$, p = 0.01	$\chi^2 = 0.26$, p = 0.61
number killed by ‘muslim’/‘islam’ terrorism	number killed by all terrorism
$\chi^2 = 4.1$, p = 0.04	$\chi^2 = 0.09$, p = 0.76

Toda-Yamamoto Granger test on logged time series
Order of integration/number of surplus lags=1
Number of weeks: 161 (December 2015 through December 2018)
Number of lags tested: 1 (selected by AIC)

Table 2: *Terrorist attacks and subsequent coverage of ‘muslim’/‘islam’ terrorism in the news.* This table shows that for December 2015 through December 2018 week-to-week changes in global deadly terror attacks did not in general predict future coverage of terrorism, whether counts of attacks or number killed by attacks. However, deadly terror attacks in the GTD whose attack descriptions specifically contained mentions of ‘muslim’ or ‘islam’ significantly predicted a future rise in numbers of articles about terrorism and Muslims/Islam (excluding ‘Islamic State’), perhaps through coverage of political rhetoric following attacks.

finding: Donald Trump promoted messages about both Muslims/Islam and ISIS during the same period that the media was simultaneously covering Muslims and terrorism after terrorist attacks – perhaps *due to* Trump’s campaign messaging about Muslims.

On the right side of this panel, we show that this shift away from co-occurring mentions of Muslims and ISIS (when mentions of ISIS increased without mentions of Muslims/Islam) corresponded to a shift to a focus on messaging related to promoting the U.S. military and supporting service members and veterans. This messaging shift occurred a month after an escalation of coalition airstrikes in Syria, which further increased in August 2017.[18] That is, intentionally or not, Donald Trump in his Twitter account appears to have reduced inflammatory rhetoric targeting Muslims and Islam, even when ISIS was salient.

As a further and related note here, ‘terrorist’ attacks in the Global Terrorism Database may have partly declined during this period due to the coalition successfully degrading the capacity of ISIS and its affiliates to conduct attacks. The group might have also shifted to waging traditional warfare against the coalition and its local partners while defending their strongholds instead of employing terrorist tactics. Repeating a part from our data section for convenience here, the Global Terrorism Database requires that at least two of the three following requirements be met for an intentional act of violence by a subnational actor be considered “terrorism”: conducted outside of legitimate warfare, conducted to attain social or religious goals, and be used to intimidate individuals beyond

[18] https://www.bbc.com/news/world-middle-east-48473979

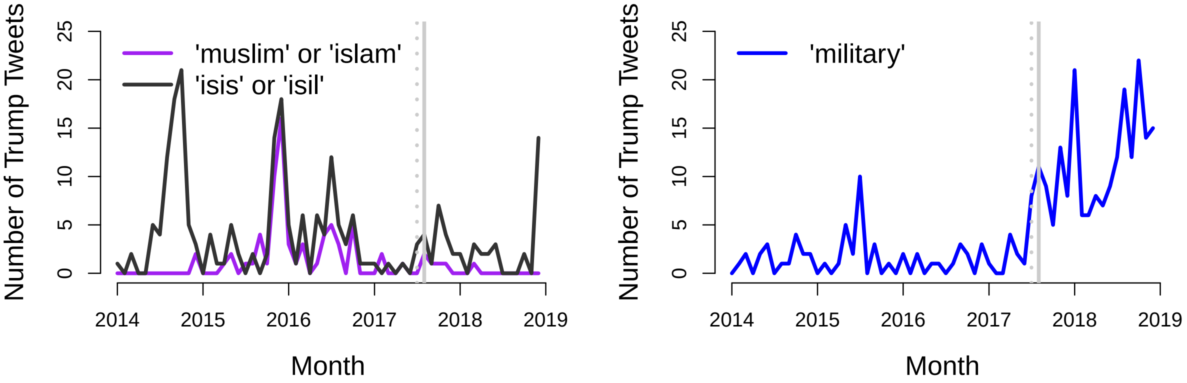


Figure 3: Associations between Muslims/Islam and ISIS/ISIL in Donald Trump's tweets. This figure shows that Donald Trump mentioned both Muslims and Islam during the U.S. presidential campaign period, but switched to mentioning only ISIS or ISIL (and not 'Islamic State') in 2017. The increase in mentions of ISIS without also referencing Muslims or Islam occurred at the same time as a large rise in coalition airstrikes against ISIS in Syria. The Trump 'travel ban'/'Muslim ban' was also being assessed in the U.S. courts during this 2017 period, during which time the Trump administration argued that the ban did not specifically target members of a religious group.

the immediate victims.

Inflammatory Rhetoric: Extremist Social Media Communities

We next consider whether the decline in the salience of Muslims was associated with a corresponding decline in far-right and extremist communities. Here, it is possible that mentions of Muslims would stay the same or increase in these communities, given the increased salience of White supremacy in the mainstream media. At the same time, we might see a drop in interest due to 1) a lack of mainstream coverage of Muslims or 2) limited interest in specifically Islamophobic rhetoric compared to resurgent anti-Semitic rhetoric.

To test group salience on mainstream versus extremist online platforms, we consider several websites ordered by their extremity, as measured by the proportion of posts including racial or ethnic slurs and, given that we can collect them, that are not quickly removed by moderators.[19] Of the websites, 4chan /pol/ is the most extreme (use of a Black slur is more frequent than the word 'Black', for example), slurs appear in less than 0.1% of posts on the mainstream site Reddit (that have not been deleted by moderators or moderation bots prior to archiving), and Gab, a website advertised for "free speech" lies between the two.

In these analyses, the main text figures show the ratio of 'Muslims or Arabs' to 'Jews' mentions

[19] See Table A.3 in the SI.

using both keywords[20] and supervised predictions from crowd-sourced hand labels.[21]

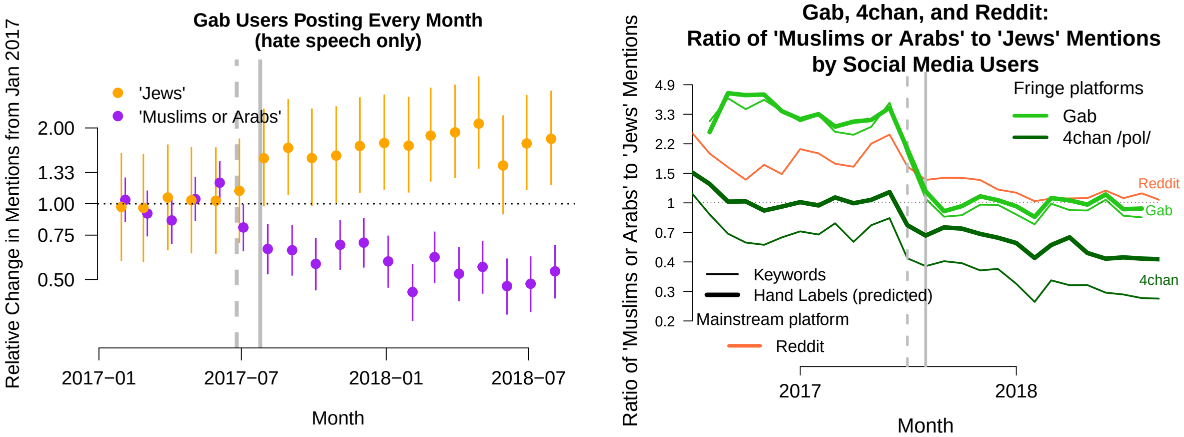


Figure 4: The left panel of this figure displays the change in mentions of Jews and Muslims or Arabs that contained predicted hate speech compared to January 2017 among Gab users who posted every month January 2017 through August 2018. The right panel displays ratio of ‘Muslims or Arabs’ to ‘Jews’ mentions on Gab, 4chan /pol/ and Reddit (our mainstream comparison) by month. Displaying the ratio of ‘Muslims or Arabs’ to ‘Jews’ mentions here treats these mentions as drawn from comparison groups. Our analyses do not consider the universe of factors contributing to hate speech and hate crimes. Note that shifts in mentions of Jews occurred only on the fringe social media platforms (see point estimates and confidence intervals in the main text), especially on Gab, and not on the mainstream platform Reddit. The solid gray line in these figures indicates Unite the Right (August 2017) and the dotted gray line the decline in the salience of Muslims in the news (July 2017).

First, patterns on Reddit, a *mainstream* social site, closely those in mainstream news media. We see a decline in mentions of Muslims or Arabs (IRR 0.58, 95% CI: 0.48 - 0.71), and no change in mentions of mentions of Jews (IRR 0.96, 95% CI: 0.90 - 1.03). See SI Figure B.3 for separate analyses of ‘subreddits’ focused on politics and religion/atheism which demonstrate the same finding.

On the extremist site Gab, however, we find an abrupt shift from mentions of Muslims or Arabs to mentions of Jews. The left panel displays changes in the numbers of mentions of Muslims or Arabs and Jews that also contained hate speech (compared to rates of mentions in January 2017) for Gab users that posted every month from January 2017 until the end of the data coverage in August 2018. Hate speech labels shown in the left panel were assigned using supervised models trained on data from the Gab Hate Corpus (Kennedy et al. 2018).[22] This result establishes that the shift from anti-Muslim or Arab to anti-Jewish speech on Gab was not solely driven by an influx of new users.

The right panel of Figure 4 shows that the ratio of these mentions declined abruptly and substantially. Note that Unite the Right is marked by a solid gray line, and that we observe a shift

[20] ‘Muslim’ and ‘Jew’ here are coded using keywords only (‘muslim’, ‘jew’, ‘islam’, ‘judaism’, ‘ arab ’)
[21] See SI sections B.2 and B.3 for labeling and model training details.
[22] See SI Section B.4 for details.

starting in the month prior, similar to the news analysis. Modeling the mean of predicted labels by month (see Tables B.8 and B.9 for full summary tables), mentions of Muslims or Arabs declined to 75% of their previous level (95% CI: 0.68 - 0.82), while mentions of Jews increased to 228% of their prior level (95% CI: 2.01 - 2.60). At the user level, we see the fraction of users mentioning Muslims or Arabs in a month decline to 60% of their previous level (95% CI: 0.51 - 0.71), and mentions of Jews increase to 180% of their prior level (95% CI: 1.54 - 2.12). In SI Table B.10, we present a summary table for the ratio of Muslims or Arabs to Jews mentions, which merely combines the above group-by-group estimates.

On 4chan, mentions of Muslims or Arabs drop at approximately the same time that mentions of Jews are increasing. Mentions of Muslims or Arabs decline to 74% of their previous level (IRR 95% CI: 0.63 - 0.88), while mentions of Jews rose to 172% of their prior level (IRR 95% CI: 1.54 - 1.93). See Table B.8 for a full summary table.[23]

We further find suggestive evidence that this shift occurred among individual users. Given our interest in shifts from one group to another, we cannot simply model one group's mentions with user fixed effects, or model the difference in group mentions. Both of these approaches can identify shifts whether or not they represent within-user shifts from one group to another. Instead, Table 3 estimates the within-user shift from discussion about Muslims or Arabs to Jews using the errors-in-variables regression from (Griliches and Hausman 1986) which we described in the methods section. The table displays robust standard errors clustered at the user level.

These within-user results show that when Gab users mentioned one group more in a month they tended to also mention the other more. These within individual shifts usually cancel out in aggregate. Around Unite the Right, however, this effect is mostly reversed. These results suggest that a significant fraction of the within-user shifts in the ratios of group mentions (for those users active on Gab from June through September 2017) were from one group to another rather than away from or toward mentioning these groups at all.

Observing similar shifts on both minimally moderated sites (4chan and Gab) increases our certainty that the shift from mentions of Muslims to mentions of Jews generalizes across social

[23] Unlike the shift on Gab, however, the increase in 'Jews' mentions on 4chan /pol/ was gradual and continued into 2018. Further, as we show in Figure B.8, the increase in mentions of Jews at Unite the Right is sudden only if we think of Muslims or Arabs mentions as a reference, and we expect the groups to both continue seeing the same increases in mentions as they saw in early 2017.

Before v After Unite the Right

Gab: within-user	Δln(‘Jews’$_{\mathrm{t}}$)	
	fraction	count
Δln(‘Muslims or Arabs’$_{\mathrm{t}}$)	0.44	0.23
	(0.39, 0.49)	(0.14, 0.31)
	$p < 0.01$	$p < 0.01$
June to Aug 2017	0.18	−0.13
	(0.03, 0.33)	(−0.34, 0.08)
	$p = 0.02$	$p = 0.22$
Δln(‘Number of Posts’$_{\mathrm{t}}$)		1.09
		(0.94, 1.24)
		$p < 0.01$
Δln(‘Muslims or Arabs’$_{\mathrm{t}}$):June to Aug 2017	−0.36	−0.45
	(−0.51, −0.20)	(−0.69, −0.22)
	$p < 0.01$	$p < 0.01$
Δln(‘Number of Posts’$_{\mathrm{t}}$):June to Aug 2017		0.75
		(0.32, 1.18)
		$p < 0.01$
Intercept	0.03	0.06
	(0.02, 0.04)	(0.05, 0.07)
	$p < 0.01$	$p < 0.01$
Number of Users	24,106	24,106

Table 3: *Gab: within-user shifts in ‘Jews’ and ‘Muslims or Arabs’ mentions.* This table displays an errors-in-variables regression that compares the within-user consistency of shifts in same group mentions over time to shifts from one group to another (see text for more details). Variables ‘Jews’ and ‘Muslims or Arabs’ are non-zero average predicted probabilities for the ‘fraction’ model and sums of predicted probabilities for the ‘count’ model. Overall, when users shift toward mentioning one group they tend to also mention the other. Around Unite the Right, we estimate a separate effect where users who mentioned Muslims and/or Arabs less mentioned Jews substantially more than they otherwise would.

media platforms that harbor extremists. In particular, the shift should not be attributed to the vagaries of a single platform, and so may be more likely to extend more generally offline as well. At the same time, the decline in anti-Muslim or Arab rhetoric appears to have been particularly abrupt on both extreme and more moderate platforms.

Persistence of Hate: Declines and Target Substitution in Hate Crimes

Having established that a sizable decline in mentions of Muslims across all platforms was followed by a shift to mentions of Jews on extremist sites, we next examine whether these changes in hate speech mirror offline changes in hate crimes and bias incidents. Our analysis of hate crimes uses two types of data: 1) reports collected by advocacy organizations, and 2) reports of violent and 'likely to be reported' hate crime sent to the FBI.

The left panel of Figure 5 displays the number of hate crimes and bias incidents reported to CAIR and ADL between 2016 and 2019. The figure demonstrates that violence, assault, harassment, and destruction of property directed toward Muslims declined in August 2017 (quasi-Poisson IRR: 0.49, 95% CI: 0.38 - 0.62), while aggression directed toward Jews stayed the same or increased (IRR: 1.18, 95% CI: 0.88 - 1.57). These anti-Jewish bias incidents were higher than levels prior to the 2016 election, and largely lower than incidents just after the election and during presidential transition. The right panel of Figure 5 shows the ratio of anti-Semitic hate crimes to anti-Muslim hate crimes, and shows a relative increase in anti-Semitic hate crimes in August 2017 (OLS log-linear change in ratio, CAIR to ADL: 0.40, 95% CI: 0.31 - 0.52).

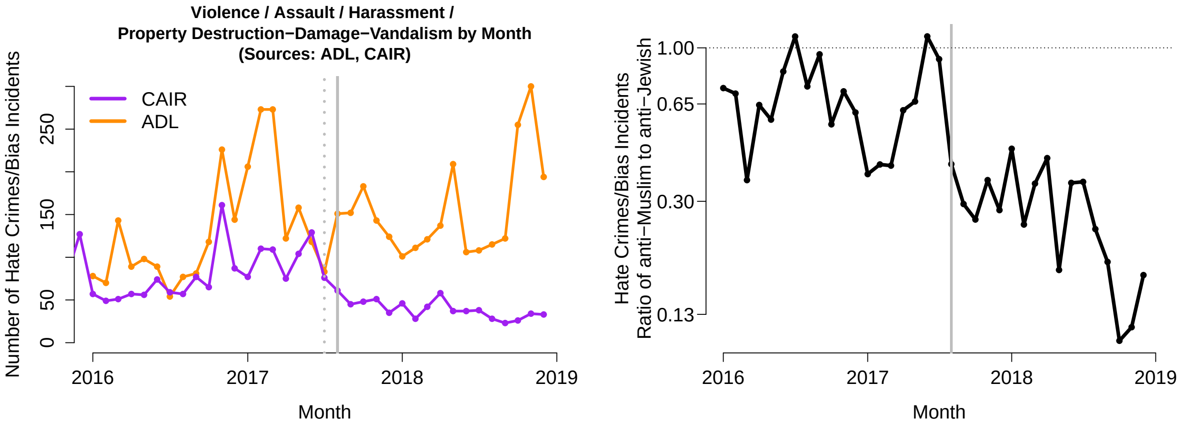


Figure 5: The left panel above displays hate crimes and bias incidents recorded by ADL and CAIR by month. The right panel compares these two sources, and displays the ratio of anti-Muslim (CAIR) to anti-Jewish (ADL) events on a log scale (using unlogged labels). This comparison isolates shifts where the groups saw diverging rates of hate crimes and bias incidents. Note that, in contrast with the CAIR data, the FBI data analyzed below contains both anti-Muslim and anti-Arab hate crime incidents.

In addition to considering level changes in hate crimes committed against Jews and Muslims, we also test whether week-to-week changes in mentions of these groups on the fringe social media sites Gab and 4chan predict *future* hate crimes and bias incidents. Table 4 reports the findings from a Toda-Yamamoto Granger causality test. This table shows that the log ratio of mentions Jews versus Muslims/Arabs significantly predicts future hate crimes and bias incidents against Jews versus Muslims, including in models that include the number of terror attacks with 'muslim' or 'islam' in attack descriptions from the Global Terrorism Database and the ratio of news articles (as in other analyses, mentions of Jews to Muslims) about hate crimes.

log ratio of group mentions on **fringe/extremist social media** (Gab, 4chan)
↓
subsequent log ratio of **hate crimes, bias incidents** (ADL, CAIR)

	controlling for terror attacks (‘muslim’, ‘islam’, ‘isis‘/isil’)	controlling for terror attacks (‘muslim’, ‘islam’, ‘isis‘/isil’) and hate crime media coverage
$\chi^2 = 8.7$, p = 0.01	$\chi^2 = 7.6$, p = 0.02	$\chi^2 = 10.7$, p < 0.01
Using the same data to test for the reverse: **hate crimes/bias incidents → group mentions**		
$\chi^2 = 0.8$, p = 0.67	$\chi^2 = 0.2$, p = 0.9	$\chi^2 = 0.9$, p = 0.65

Toda-Yamamoto Granger causality test on logged time series
Order of integration/number of surplus lags=1
Number of weeks: 87
Number of lags tested: 2 (selected by AIC)

Table 4: This table demonstrates that mentions of Muslims or Arabs or Jews on Gab and 4chan significantly predicted future hate crimes and bias incidents in the CAIR and ADL data. Counts of deadly terror attacks include all attacks in the Global Terrorism Database which contained the keywords ‘muslim’, ‘islam’ (including ‘Islamic State’), or ‘isis’/‘isil’ in their attack descriptions. We replicate these results using controls for the number of people killed in terror attacks in SI Table B.14.

Next, we consider violent and ‘likely to be reported’ hate crimes in government records. Anti-Muslim aggravated assault, manslaughter, murder, arson, and kidnapping increased during the U.S. presidential campaign (after December 2015 – IRR: 2.69, 95% CI 1.82 - 4.01, p-value: <0.01), but declined in mid-2017 (IRR: 0.63, 95% CI 0.43 - 0.91, p-value: 0.02), while anti-Semitic violent crimes did not substantially change during the anti-Muslim crime spike and then appeared to dramatically increase after the drop in anti-Muslim hate crimes in mid-2017 (IRR: 1.82, 95% CI 0.99 - 3.42, p-value: 0.07). This effect in violent and likely to be reported crimes is much larger than the estimated effect for all bias incidents reported to ADL, though it is also estimated less precisely than that effect. Figure B.10 in the SI displays this result visually. Finally, in Table B.20, we control for other hate crimes related to race/ethnicity and include month of year fixed effects (i.e. January, February, etc, since hate crimes go down in the winter), and find that the effect declines only slightly (IRR: 1.60, p-value: 0.13). However, this estimate may be biased downwards if White supremacists increasingly targeted other groups as well.

In the FBI data, Granger causality tests find that the ratio of ‘Muslims or Arabs’ to ‘Jews’ mentions on 4chan significantly predicted the future ratio of violent and ‘likely to be reported’ hate

anti-Muslim or anti-Arab Hate Crimes (FBI Uniform Crime Reports)
aggravated assault, arson, kidnapping manslaughter, murder

2016 Presidential Campaign / 'Muslim ban' statement / Increase in anti-Muslim Rhetoric	0.99 (0.60, 1.41) p <0.01
Decline in anti-Muslim Rhetoric / 'Unite the Right'	−0.46 (−0.84, −0.09) p = 0.02

anti-Jewish Hate Crimes (FBI Uniform Crime Reports)
aggravated assault, arson, kidnapping manslaughter, murder

2016 Presidential Campaign	−0.07 (−0.72, 0.58) p = 0.85
Increase in anti-Jewish Rhetoric / 'Unite the Right'	0.60 (−0.01, 1.23) p = 0.07
N	60

Table 5: This table displays changes in rate of violent and 'likely to be reported' hate crimes in the FBI UCR data by month. Coefficients are logged incidence rate ratios – the anti-Muslim or Arab hate crimes declined to 63% of their December 2015 through July 2017 level, while anti-Jewish hate crimes increased to 182% of their Dec. 2015 through July 2017 level. See Figure B.10 for visualizations of the raw count data, along with a comparison to advocacy organization reports and models with controls for other race and ethnicity related hate crimes. In the FBI data, Granger causality tests (see Tables B.15 and B.16 in the SI) find that the ratio of 'Muslims or Arabs' to 'Jews' mentions on fringe social media significantly predicted the future ratio of violent and 'likely to be reported' hate crimes after Unite the Right ($p < 0.01$) and did not predict this ratio before it ($p = 0.91$).

crimes after Unite the Right ($p < 0.01$) and did not predict this ratio before it ($p = 0.91$).[24] Unlike in the ADL and CAIR data, violent hate crimes did appear to predict the subsequent ratio group mentions on 4chan ($p = 0.06$), though this was less statistically significant than the reverse. As a reminder, we use 4chan data for this test because, unlike the Gab data, it covers the longer time frame in the FBI data. As for the ADL and CAIR data, this analysis controlled for both terror attacks and media coverage of hate crimes, and results were not different when controlling for the log number killed in terror attacks, rather than the log number of attacks.

[24] See Section B.8, Tables B.15 and B.16.

Discussion

Since the early days of the 2016 presidential election, the U.S. has been witnessing an uptick in far-right extremism and violent hate crimes. This era saw a rise in White supremacist events, such as rallies and demonstrations, growth in extremist social media forums, and a nearly tripling in White supremacist propaganda from 2017 to 2018.[25] But despite a rise in recent years of anecdotal evidence and scholarly work on online extremism, there is still much to learn about its over-time dynamics and consequences, and the factors that threaten the safety of socially marginalized groups in the United States.

In this paper, we examine two open and pressing questions: is group salience and inflammatory rhetoric associated with hate crimes? And does the effect depend on continued salience and incitement? Our analyses here provide strong suggestive evidence on both of these questions. When group salience and inflammatory rhetoric go down, hate crimes go down. The existing literature linking group salience to hate crimes is partly hindered by the threat that increases in group salience might lead to more complete reporting without shifts in crimes or bias incidents themselves, and shifts in anti-Muslim rhetoric may have occurred around the same time as terrorist attacks. Our study of these questions instead evaluates these questions by examining declines in group salience and declines in hate crimes. Here, specifically, after a decline in the salience of Muslims in the mainstream media and online communities, declines in hate crimes occurred in a matter of weeks to months.

Notable for scholars of far-right extremism, our results also show that increased hate does not evaporate following a decline in group salience or inflammatory rhetoric; although expression of hate might decline in general, such as on mainstream platforms, related expressions of hate can increase in extremist communities where they require no further propagation to persist. We observe some of the same social media communities and users who previously promoted anti-Muslim speech begin to promote anti-Jewish speech. This online behavior was accompanied by an increase in anti-Jewish bias incidents and hate crimes. That hate crimes could decline and shift in this way suggests that perpetrators of hate can be traced to the same communities and that they shift their targets and pretexts for hate over time.

[25] https://www.adl.org/resources/reports/white-supremacists-step-up-off-campus-propaganda-efforts-in-2018.

Of course, our study does not explain the universe of potential drivers of hate crimes, nor can we say whether those publicly expressing hate on these fringe social media sites are the very same individuals as those committing hate crimes and bias incidents offline. At the same time, our results point to July and August 2017, the months before and after the Unite the Right rally, as a pivotal period in which the mainstream salience of Muslims declined and was followed by an increase in the salience of groups promoting White supremacy outright. This event saw hundreds of far right protesters chanting racist and anti-Semitic slogans descend on Charlottesville, Virginia while carrying weapons, neo-Nazi symbols, Confederate battle flags, and Deus Vult crosses. This was an unusually visible and widely reported display by White supremacists. Yet, the number of protesters there was small, much smaller than those later active online, and likely to be only a fraction of those now supporting its content.

Finally, while we cannot provide a causal answer to all questions raised by this analysis, we hope that the evidence put forth here will inform future research – especially work to better understand the range of mechanisms driving persistent and *shifting* hate among the same communities and individuals. For now, the large, real-world, online and offline associations shown here underscore the importance of scholarly attention to extremist discussion on online social media and how it might serve as an important indicator, predictor, and propellant of group conflict.

References

Abrajano, Marisa, and Zoltan L Hajnal. 2015. *White Backlash.* Princeton University Press.

Almerekhi, Hind, Haewoon Kwak, Bernard J Jansen, and Joni Salminen. 2019. "Detecting Toxicity Triggers in Online Discussions". In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, 291–292. HT '19. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3342220.3344933.

Alrababah, Ala, William Marble, Salma Mousa, and Alexandra Siegel. 2021. "Can Exposure to Celebrities Reduce Prejudice? The Effect of Mohamed Salah on Islamophobic Behaviors and Attitudes". *American Political Science Review.*

Álvarez-Benjumea, Amalia, and Fabian Winter. 2020. "The breakdown of antiracist norms: A natural experiment on hate speech after terrorist attacks". *Proceedings of the National Academy of Sciences* 117 (37): 22800–22804. https://www.pnas.org/content/117/37/22800.short.

Awan, Imran, and Irene Zempi. 2016. "The affinity between online and offline anti-Muslim hate crime: Dynamics and impacts". *Aggression and violent behavior* 27:1–8.

Bail, Christopher A. 2012. "The Fringe Effect". *American Sociological Review* 77 (6): 855–879.

Barth, Whittney, Kathy Bruce, Beth Daviess, Madeline Hall, Gabriel Lazarus, and Caroline Sabatier. 2019. *Researching Hate: Challenges to Tracking Hate Crimes & Practices for Collecting Better Data.* SSRN.

Baumgartner, Jason. 2019. *The Gab Pushshift Dataset.* \url{https://files.pushshift.io/gab/}.

Baumgartner, Jason, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. "The Pushshift Reddit Dataset". *Proceedings of the International AAAI Conference on Web and Social Media* 14, no. 1 (): 830–839. https://www.aaai.org/ojs/index.php/ICWSM/article/view/7347.

Berger, Jonathon M, and Jonathon Morgan. 2015. "The ISIS Twitter Census: Defining and describing the population of ISIS supporters on Twitter". *The Brookings Project on US Relations with the Islamic World* 3 (20): 1–4.

Beydoun, Khaled A. 2018. *American Islamophobia: Understanding the roots and rise of fear.* Univ of California Press.

— . 2013. "Between Muslim and White: The Legal Construction of Arab American Identity". *NYU Ann. Surv. Am. L.* 69:29.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* O'Reilly Media, Inc.

Bleich, Erik, and A Maurits van der Veen. 2018. "Media portrayals of Muslims: a comparative sentiment analysis of American newspapers, 1996–2015". *Politics, Groups, and Identities*: 1–20.

Bloom, Mia, Hicham Tiflati, and John Horgan. 2019. "Navigating ISIS's Preferred Platform: Telegram1". *Terrorism and Political Violence* 31 (6): 1242–1254. https://doi.org/10.1080/09546553.2017.1339695.

Burris, Val, Emery Smith, and Ann Strahm. 2000. "White supremacist networks on the Internet". *Sociological Focus* 33 (2): 215–235.

Bursztyn, Leonardo, Georgy Egorov, and Stefano Fiorin. 2020. "From extreme to mainstream: The erosion of social norms". *American Economic Review* 110 (11): 3522–3548.

Byers, Bryan D, and James A Jones. 2007. “The impact of the terrorist attacks of 9/11 on anti-Islamic hate crime”. *Journal of Ethnicity in Criminal Justice* 5 (1): 43–56.

Chyzh, Olga, Mark David Nieman, and Clayton Webb. 2019. “The effects of dog-whistle politics on political violence”. Unpublished Manuscript. https://lib.dr.iastate.edu/pols_pubs/59/.

Cikara, Mina, Vasiliki Fouka, and Marco Tabellini. 2020. “Hate crime increases with minoritized group rank”. Unpublished Manuscript. https://files.osf.io/v1/resources/2z3kw/providers/osfstorage/5fdaad350694b7049ef3a124?format=pdf&action=download&direct&version=3.

Cohen, Jeffrey E. 2010. “Perceptions of Anti-Semitism among American Jews, 2000–05, A Survey Analysis”. *Political Psychology* 31 (1): 85–107.

Conway, Maura, Moign Khawaja, Suraj Lakhani, Jeremy Reffin, Andrew Robertson, and David Weir. 2019. “Disrupting Daesh: Measuring Takedown of Online Terrorist Material and Its Impacts”. *Studies in Conflict & Terrorism* 42 (1-2): 141–160. https://doi.org/10.1080/1057610X.2018.1513984.

Costa, Mia. 2020. “Ideology, Not Affect: What Americans Want from Political Representation”. *American Journal of Political Science* 65 (2): 342–358.

Czymara, Christian S. 2020. “Propagated preferences? Political elite discourses and Europeans’ openness toward Muslim immigrants”. *International Migration Review* 54 (4): 1212–1237.

Dancygier, Rafaela, Naoki Egami, Amaney Jamal, and Ramona Rischke. 2021. “Hate crimes and gender imbalances: fears over mate competition and violence against refugees”. *American Journal of Political Science.*

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. “Bert: Pre-training of deep bidirectional transformers for language understanding”. *arXiv Preprint* arXiv:1810.04805:1–16.

Dugan, Laura, and Erica Chenoweth. 2020. “Threat, emboldenment, or both? The effects of political power on violent hate crimes”. *Criminology* 58 (4): 714–746.

Ekman, Mattias. 2015. “Online Islamophobia and the politics of fear: Manufacturing the green scare”. *Ethnic and Racial Studies* 38 (11): 1986–2002.

Entman, Robert M, and Andrew Rojecki. 2001. *The Black image in the White mind: Media and race in America.* University of Chicago Press.

Esses, Victoria M, Lynne M Jackson, and Tamara L Armstrong. 1998. “Intergroup competition and attitudes toward immigrants and immigration: An instrumental model of group conflict”. *Journal of Social Issues* 54 (4): 699–724.

Farrell, Tracie, Miriam Fernandez, Jakub Novotny, and Harith Alani. 2019. “Exploring Misogyny across the Manosphere in Reddit”. In *Proceedings of the 10th ACM Conference on Web Science*, 87–96. WebSci ’19. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3292522.3326045.

Fleiss, Joseph L. 1971. “Measuring nominal scale agreement among many raters.” *Psychological bulletin* 76 (5): 378.

Fording, Richard C, and Sanford F Schram. 2020. *Hard white: the mainstreaming of racism in American Politics.* Oxford University Press.

Freilich, Joshua D, and Steven M Chermak. 2013. *Hate crimes.* US Department of Justice, Office of Community Oriented Policing Services. http://www.justiceacademy.org/iShare/Library-COPS/cops-p268-pub.pdf.

Garimella, Kiran, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. "Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship". In *Proceedings of the 2018 World Wide Web Conference*, 913–922.

Giani, Marco, and Pierre-Guillaume Méon. 2019. "Global racist contagion following Donald Trump's election". *British Journal of Political Science* 1:1–8.

Gottschalk, Peter. 2013. *American Heretics: Catholics, Jews, Muslims, and the History of Religious Intolerance.* St. Martin's Press.

Green, Donald P, Jack Glaser, and Andrew Rich. 1998. "From lynching to gay bashing: the elusive connection between economic conditions and hate crime." *Journal of Personality and Social Psychology* 75 (1): 82.

Green, Donald P, and Amber D Spry. 2014. "Hate crime research: Design and measurement strategies for improving causal inference". *Journal of Contemporary Criminal Justice* 30 (3): 228–246.

Green, Donald P, Dara Z Strolovitch, and Janelle S Wong. 1998. "Defended neighborhoods, integration, and racially motivated crime". *American Journal of Sociology* 104 (2): 372–403.

Griliches, Zvi, and Jerry A Hausman. 1986. "Errors in Variables in Panel Data". *Journal of Econometrics* 31:93–118. https://www.sciencedirect.com/science/article/pii/0304407686900588.

Grover, Ted, and Gloria Mark. 2019. "Detecting potential warning behaviors of ideological radicalization in an alt-right subreddit". In *Proceedings of the International AAAI Conference on Web and Social Media*, 13:193–204.

Hafez, Mohammed, and Creighton Mullins. 2015. "The radicalization puzzle: A theoretical synthesis of empirical approaches to homegrown extremism". *Studies in Conflict & Terrorism* 38 (11): 958–975.

Hawdon, James, Colin Bernatzky, and Matthew Costello. 2019. "Cyber-routines, political attitudes, and exposure to violence-advocating online extremism". *Social Forces* 98 (1): 329–354.

Hine, Gabriel Emile, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. 2017. "Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web". In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*, 92–101.

Hobbs, William, and Nazita Lajevardi. 2019. "Effects of divisive political campaigns on the day-to-day segregation of Arab and Muslim Americans". *American Political Science Review* 113 (1): 270–276.

Hopkins, Daniel J. 2021. "The Activation of Prejudice and Presidential Voting: Panel Evidence from the 2016 US Election". *Political Behavior* 43:663–686.

Hutchings, Vincent L, Hanes Walton Jr, and Andrea Benjamin. 2010. "The impact of explicit racial cues on gender differences in support for confederate symbols and partisanship". *The Journal of Politics* 72 (4): 1175–1188.

Iannelli, Jerru. 2018. *One Year Later, MDPD Likely Not Reporting Hate-Crime Data Despite Promises.*

Jardina, Ashley, and Spencer Piston. 2021. "The Effects of Dehumanizing Attitudes about Black People on Whites' Voting Decisions". *The British Journal of Political Science.*

Jhaver, Shagun, Amy Bruckman, and Eric Gilbert. 2019. "Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit". *Proc. ACM Hum.-Comput. Interact.* (New York, NY, USA) 3, no. CSCW (). https://doi.org/10.1145/3359252.

Kennedy, Brendan, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2018. "The Gab Hate Corpus: A collection of 27k posts annotated for hate speech".

Krafft, P M, and Joan Donovan. 2020. "Disinformation by Design: The Use of Evidence Collages and Platform Filtering in a Media Manipulation Campaign". *Political Communication* 37 (2): 194–214.

LaFree, Gary, and Laura Dugan. 2011. *Global Terrorism Database.*

Lajevardi, Nazita. 2021. "The media matters: Muslim american portrayals and the effects on mass attitudes". *The Journal of Politics* 83 (3).

Leander, N Pontus, Jannis Kreienkamp, Maximilian Agostini, Wolfgang Stroebe, Ernestine H Gordijn, and Arie W Kruglanski. 2020. "Biased hate crime perceptions can reveal supremacist sympathies". *Proceedings of the National Academy of Sciences* 117 (32): 19072–19079.

Mathew, Binny, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. "Spread of hate speech in online social media". In *Proceedings of the 10th ACM conference on web science*, 173–182.

McCombs, Maxwell E, and Donald L Shaw. 1972. "The agenda-setting function of mass media". *Public Opinion Quarterly* 36 (2): 176–187.

McVeigh, Rory, Michael R Welch, and Thoroddur Bjarnason. 2003. "Hate crime reporting as a successful social movement outcome". *American Sociological Review* 68 (6): 843–867.

Meer, Nasar. 2013. "Racialization and religion: race, culture and difference in the study of anti-semitism and Islamophobia". *Ethnic and Racial Studies* 36 (3): 385–398.

Mudde, Cas. 2019. *The Far Right Today.* John Wiley & Sons.

Müller, Karsten, and Carlo Schwarz. 2020. "Fanning the flames of hate: Social media and hate crime". *Journal of the European Economic Association.*

— . 2018. "Making America hate again? Twitter and hate crime under Trump". *Twitter and Hate Crime Under Trump (March 30, 2018).*

Nacos, Brigitte. 2016. *Mass-mediated terrorism: Mainstream and digital media in terrorism and counterterrorism.* Rowman & Littlefield.

Newman, Benjamin J, Sono Shah, and Loren Collingwood. 2018. "Race, place, and building a base: Latino population growth and the nascent Trump campaign for president". *Public Opinion Quarterly* 82 (1): 122–134.

Newman, Benjamin, Jennifer L Merolla, Sono Shah, Danielle Casarez Lemi, Loren Collingwood, and S Karthick Ramakrishnan. 2020. "The Trump Effect: An Experimental Investigation of the Emboldening Effect of Racially Inflammatory Elite Communication". *British Journal of Political Science* 0:1–22.

Nithyanand, Rishab, Brian Schaffner, and Phillipa Gill. 2017. "Online political discourse in the Trump era". *arXiv preprint arXiv:1711.05303.*

Oskooii, Kassra A R, Nazita Lajevardi, and Loren Collingwood. 2019. "Opinion Shift and Stability: The Information Environment and Long-Lasting Opposition to Trump's Muslim Ban". *Political Behavior*: 1–37.

Parker, Christopher S, and Matt A Barreto. 2014. *Change They Can't Believe In: The Tea Party and Reactionary Politics in America-Updated Edition.* Princeton University Press.

Pitofsky, Marina. 2018. "Hate crimes are up in America's 10 largest cities. Here's why". *USA Today* (). https://www.usatoday.com/story/news/2018/07/17/hate-crimes-up-america-10-largest-cities/776721002/.

Reny, Tyler T, Ali A Valenzuela, and Loren Collingwood. 2020. ""No, you're playing the race card": Testing the effects of anti-black, anti-Latino, and anti-immigrant appeals in the post-Obama era". *Political Psychology* 41 (2): 283–302.

Samari, Goleen. 2016. "Islamophobia and public health in the United States". *American Journal of Public Health* 106 (11): 1920–1925.

Samari, Goleen, Ralph Catalano, Héctor E Alcalá, and Alison Gemmill. 2020. "The Muslim Ban and preterm birth: Analysis of US vital statistics data from 2009 to 2018". *Social Science & Medicine* 265:113544.

Schaffner, Brian F. 2020. *The acceptance and expression of prejudice during the Trump era.* Cambridge University Press.

Sediqe, Nura A. 2020. "Stigma Consciousness and American Identity: The Case of Muslims in the United States". *PS: Political Science & Politics* 53 (4): 674–678.

Siegel, Alexandra A, Evgenii Nikitin, Pablo Barberá, Joanna Sterling, Bethany Pullen, Richard Bonneau, Jonathan Nagler, Joshua A Tucker, et al. 2021. "Trumping Hate on Twitter? Online Hate Speech in the 2016 US Election Campaign and its Aftermath". *Quarterly Journal of Political Science* 16 (1): 71–104.

Stacey, Michele, Kristin Carbone-López, and Richard Rosenfeld. 2011. "Demographic change and ethnically motivated crime: The impact of immigration on anti-Hispanic hate crime in the United States". *Journal of Contemporary Criminal Justice* 27 (3): 278–298.

Stephens-Dougan, LaFleur. 2020. *Race to the Bottom: How Racial Appeals Work in American Politics.* University of Chicago Press.

Sunstein, Cass R. 2018. *# Republic: Divided democracy in the age of social media.* Princeton University Press.

Terman, Rochelle. 2017. "Islamophobia and media portrayals of Muslim women: A computational text analysis of US news coverage". *International Studies Quarterly* 61 (3): 489–502.

Tesler, Michael. 2016. *Post-racial or most-racial?: Race and politics in the Obama era.* University of Chicago Press.

— . 2013. "The return of old-fashioned racism to White Americans' partisan preferences in the early Obama era". *The Journal of Politics* 75 (1): 110–123.

Toda, Hiro Y, and Taku Yamamoto. 1995. "Statistical inference in vector autoregressions with possibly integrated processes". *Journal of Econometrics* 66:225–250.

Valentino, Nicholas A, Fabian G Neuner, and L Matthew Vandenbroek. 2018. "The changing norms of racial political rhetoric and the end of racial priming". *The Journal of Politics* 80 (3): 757–771.

Van Dijk, Teun A. 1993. *Elite discourse and racism.* Vol. 6. Sage.

White, Ismail K. 2007. “When race matters and when it doesn’t: Racial group differences in response to racial cues”. *American Political Science Review* 101 (2): 339–354.

Williams, Aaron. 2018. “Hate crimes rose the day after Trump was elected, FBI data show”. *The Washington Post* 23.

Williams, Matthew L, Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. 2020. “Hate in the machine: Anti-Black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime”. *The British Journal of Criminology* 60 (1): 93–117.

Zannettou, Savvas, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. 2018a. “What is Gab: A Bastion of Free Speech or an Alt-Right Echo Chamber”. In *Companion Proceedings of the The Web Conference 2018*, 1007–1014.

Zannettou, Savvas, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018b. “On the Origins of Memes by Means of Fringe Web Communities”. In *IMC*, 188–202. New York, New York, USA: ACM Press. http://dl.acm.org/citation.cfm?doid=3278532.3278550.

— . 2018c. “On the Origins of Memes by Means of Fringe Web Communities”. In *Proceedings of the Internet Measurement Conference 2018*, 188–202. IMC ’18. Boston, MA, USA: Association for Computing Machinery. https://doi-org.proxy1.cl.msu.edu/10.1145/3278532.3278550.

Zannettou, Savvas, Joel Finkelstein, Barry Bradlyn, and Jeremy Blackburn. 2020. “A Quantitative Approach to Understanding Online Antisemitism”. In *Proceedings of the International AAAI Conference on Web and Social Media*, 14:786–797.

Supplementary Information

Table of Contents

List of Tables

List of Figures

A Methodological Background

A.1 Measuring Group Salience: Media Data and Elite Discourse

We rely on a corpus of news articles to document the salience of minority groups and White supremacy and to drive our expectations of shifts in the discussion of different groups. We compiled these corpora using a web application called Text Assembler developed by a large midwestern public university in the United States. The application relies on the university's subscription to the LexisNexis API to search the entirety of LexisNexis' global news sources for user-supplied keywords. We conducted searches across all available news sources for years 2014-2018 for key terms related to the groups that we are interested in (e.g., similar to Hobbs and Lajevardi 2019; Lajevardi 2021). Specifically, we searched for the following terms: 1) "muslim" for the Muslim corpus (N = 74,131), 2) "jew" for the Jewish corpus (N = 57,776), 3) "white supremacy" for the White supremacy corpus (N = 4,078), 4) "terrorism"/"terrorist" for the terrorism corpus (N = 97,996), and 5) 'hate crime' for the hate crime corpus (N = 16,200) which was later subset to articles that specifically mentioned 'muslim' or 'islam' (N = 3,410) or 'jew', 'judaism', or 'semit' (to capture anti-Semitic) (N = 3,646). Based on our analysis of returned data, LexisNexis expanded these keywords to include articles not directly using the search terms (e.g. returning articles only including the word 'islam' but not 'muslim' in the text). We removed articles without the search terms listed above, however, retaining expanded search results does not alter our results.

Since our project focuses on the United States, we removed articles published by outlets based outside of the country. We did this by matching their names with outlets in the Library of Congress' Chronicling America project's U.S. Newspaper Directory,[26] which maintains a list of over 20,000 newspapers based in the United States. This processing step also removes items from our corpora that are from non-newspaper sources that LexisNexis indexes, such as broadcast news transcripts and legal publications. Approximately 70% of the keyword mentions in our data come from 25 sources. We list these sources in the table below.

The New York Times	The Seattle Times
The Washington Post	The Boston Herald
The Boston Globe	The Philadelphia Inquirer
Los Angeles Times	The Baltimore Sun
Chicago Tribune	The Arizona Republic (Phoenix)
The New York Post	The Courier-Journal (Louisville, Kentucky)
Daily News (New York)	THE DALLAS MORNING NEWS
St. Louis Post-Dispatch (Missouri)	Detroit Free Press (Michigan)
Pittsburgh Post-Gazette	Star Tribune (Minneapolis, MN)
Orlando Sentinel (Florida)	Plain Dealer (Cleveland, OH)
The Christian Science Monitor	Tampa Bay Times
USA TODAY	The San Francisco Chronicle (California)
The Houston Chronicle	

Table A.1: 25 newspapers with most mentions of keywords – accounting for 70% of mentions in our analyses.

A.2 Measuring Inflammatory Rhetoric: Social Networking Sites

Table A.2 lists the platforms that we analyze in the main paper, along with their respective moderation policies and the source of the archival data used in our analysis. The next sections explain each of these sites and data archives in more detail.

[26] https://chroniclingamerica.loc.gov/

Platform	Data Source	Format	Moderation Level
Reddit	Pushshift	Collection of Forums	Moderate
Gab	Pushshift	Social Network	Minimal
4chan /pol/	4plebs.org	Bulletin Board	Minimal

Table A.2: Social media data sources and their moderation levels. See dataset by dataset explanations below for more details.

Table A.3 displays the fraction of posts containing slurs by social media site to establish that users on fringe sites 4chan /pol/ and Gab are more likely to use slurs (and not have these slurs removed by moderators) than users on the mainstream site Reddit.

	Gab	**4chan**	**Reddit**	
Fraction of **posts** containing __				
anti-Black slur	0.4%	3.0%	<0.1%	
anti-Jewish slur	0.3%	1.0%	<0.1%	
‘Black’	1.5%	2.1%	0.8%	
‘Jew’	2.2%	3.4%	0.1%	
Fraction of:	**users**	**threads**	**users**	**threads**
anti-Black slur	2.9%	29.0%	0.9%	<0.1%
anti-Jewish slur	1.9%	16.2%	0.2%	<0.1%
‘Black’	13.8%	23.0%	20.0%	3.3%
‘Jew’	10.4%	31.2%	5.8%	0.5%

Table A.3: Slurs by social media site. This table studies use of slurs (and their non-removal) as a measure of social media site extremity. We use anti-Black slurs rather than anti-Muslim slurs because 1) we believe that extremists might use ‘Muslim’ itself as if it is a slur, and 2) anti-Arab slurs use the same words as anti-Black slurs.

A.2.1 Reddit

We rely on Pushshift for the data we analyze from Reddit (Baumgartner et al. 2020). Pushshift ingests massive amounts of data in real-time from various social media websites, stores it in a database, and provides access to it through an API and in file dumps. It has been widely used in academic research, particularly in studies that examine hate speech and the moderation of online content (Zannettou et al. 2018b; Almerekhi et al. 2019; Farrell et al. 2019; Jhaver, Bruckman, and Gilbert 2019). Specifically, we downloaded a list of 78 million Reddit users from the Pushshift website (`https://files.pushshift.io/reddit/authors/RA_78M.csv.zst`) and randomly sampled 5 million users from the list. We then used the Pushshift API (`https://api.pushshift.io/reddit/search/`) to fetch all the posts made by those users since 2010.

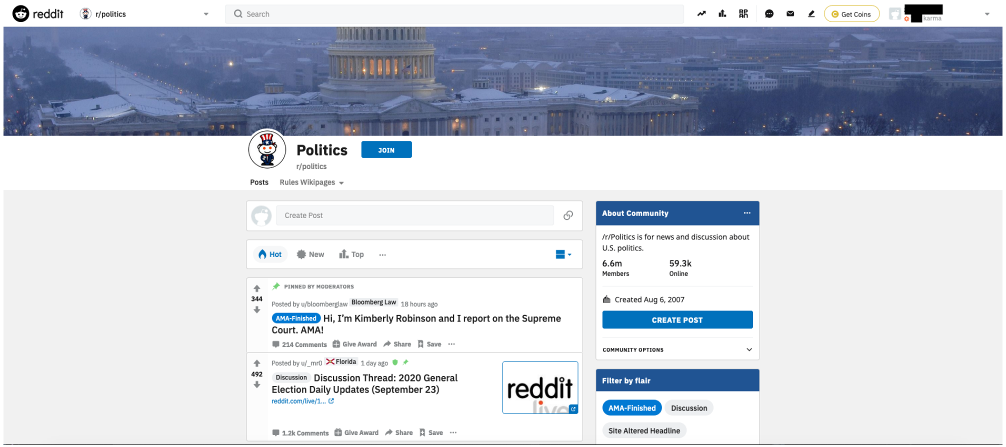
Reddit is a social media website built around subforums (“subreddits”) that are dedicated to a specific topic, such as traveling or pets. Alexa ranks it as being one of the top 20 most popular websites in the world.[27] The site, which started in 2005, employs a point system where users upvote or downvote content. This increases or decreases the visibility of content respectively. Reddit receives more than 20 billion screenviews every month and hosts over 400 million active users.[28]

Each subreddit is primarily governed by moderators specific to that community. These moderators can create local rules that are customized to their community, such as limits on what type of content

[27] `https://www.alexa.com/topsites`

[28] `https://www.redditinc.com/`

can be posted (e.g. r/Politics only allows content related to US politics and r/aww disallows ‘sad’ content) or a restriction on the sort of media that is permitted (e.g. posts in r/AskReddit can only include text). This subreddit-driven moderation strategy is partly credited for driving the site’s success, but also for producing subreddits that are amenable to or that explicitly encourage hate speech.[29] Global rules also govern the site. These restrict content that is illegal, spam, etc., but no global rule on the platform moderated hate speech until June 2020.[30] However, many individual subreddits imposed and enforced rules that restricted hate speech before Reddit updated the global rules. These rules were changed in response to a letter posted on the site signed by hundreds of Reddit moderators[31] in June 2020 that called for a global content policy because certain communities, such as r/The_Donald, became notorious for allowing hate speech. The updated rules state that “[c]ommunities and users that incite violence or that promote hate based on identity or vulnerability will be banned.[32] We include a screenshot of the front page of the Politics subreddit in figure A.1 below.



Posted by u/bloomberglaw Bloomberg Law 18 hours ago

Hi, I’m Kimberly Robinson and I report on the Supreme Court. AMA!

Posted by u/_mr0 Florida 1 day ago

Discussion Thread: 2020 General Election Daily Updates (September 23)

reddit.com/live/1...

/r/Politics is for news and discussion about U.S. politics.

Created Aug 6, 2007

Figure A.1: ‘Hot’ posts (the default) from the front page of Reddit’s r/Politics subreddit on September 24, 2020.

A.2.2 Gab

We use data collected from Gab by Pushshift, the same source we use for Reddit, to analyze the platform (Baumgartner 2019). In our study, we use the Gab dataset available as a single download from Pushshift: https://files.pushshift.io/misc/GAB_posts.ndjson.xz.

Gab, which went online in August 2016, is a social media website modeled on Twitter that markets itself as a free speech alternative to similar platforms. Andrew Torba, the site’s founder, argues that “[b]ig [t]ech social networks have subjective policies and rules that are enforced unfairly, unequally, and with a clear and persistent bias.”[33] Gab initially attracted users that were banned for hate speech from sites such as Twitter.[34] Immediately following the Capitol siege on January 6, 2021, Gab saw a mass migration of extremists and conspiracy theorists to fringe social media

[29] https://cyber.fsi.stanford.edu/news/reddit-hate-speech

[30] The global rules prior to June 2020 specifically noted that Reddit “generally provides a lot of leeway in what content is acceptable.” The site’s global rules in May 2020 are available here: https://web.archive.org/web/20200501101408/https://www.redditinc.com/policies/content-policyhere

[31] The letter is available here: https://www.reddit.com/r/AgainstHateSubreddits/comments/gyyqem/open_letter_to_steve_huffman_and_the_board_of/

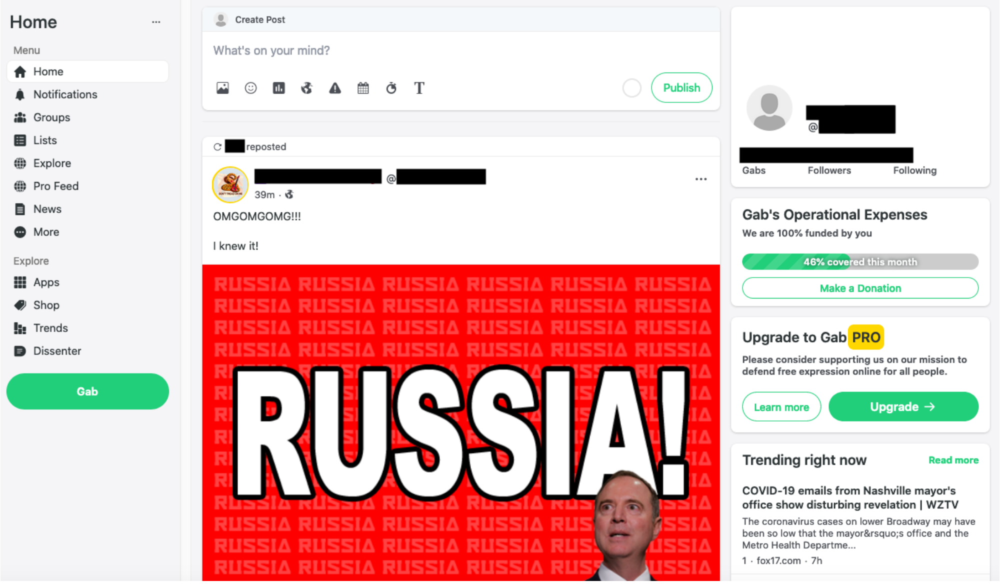
[32] https://www.redditinc.com/policies/content-policy

[33] https://news.gab.com/2019/08/23/gabs-policies-positions-and-procedures-for-unlawful-content-and-activity-on-our-social-network/

[34] See e.g. https://www.theguardian.com/media/2016/nov/17/gab-alt-right-social-media-twitter

platforms, and claimed to have gained over half a million users in the days after Donald Trump's permanent suspension from Twitter.

The site operates with very little content moderation, which is its stated reason for existence. Its Terms of Service[35] only prohibit using the site for activities like illegal activity, such a sharing child pornography, exploitation of minors, and spamming. There are no restrictions that limit hate speech, racism, sexism, etc. Figure A.2 displays the front page of a user account from Gab.



[redacted] reposted

[redacted] @[redacted]

39m ·

OMGOMGOMG!!!

I knew it!

[redacted]

@[redacted]

[redacted]

Gab's Operational Expenses

We are 100% funded by you

46% covered this month

Upgrade to Gab PRO

Please consider supporting us on our mission to defend free expression online for all people.

Trending right now

COVID-19 emails from Nashville mayor's office show disturbing revelation | WZTV

The coronavirus cases on lower Broadway may have been so low that the mayor&rsquo;s office and the Metro Health Departme...

1 · fox17.com · 7h

Figure A.2: The front page of a user account on Gab on September 21, 2020.

A.2.3 4chan

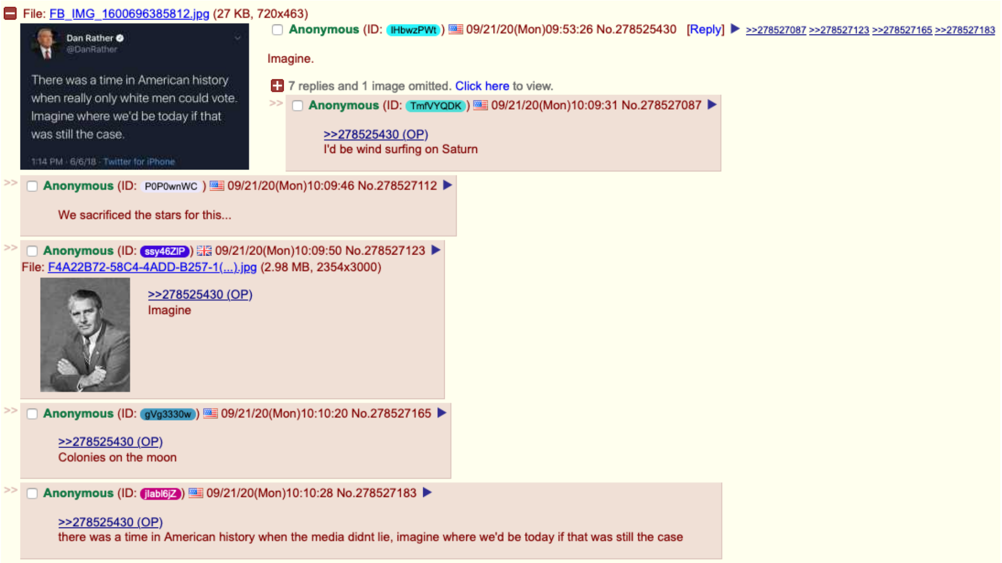
Our analyses include a dataset of 4chan posts collected by 4plebs.org. 4plebs.org maintains an archive of 4chan /pol/ posts and periodically adds all archived text and image to the Internet Archive. The textual data set we use is available here: https://archive.org/details/4plebs-org-data-dump-2020-01. The 4plebs data has been analyzed in previous academic studies (Zannettou et al. 2018c; Krafft and Donovan 2020).

4chan is a website of anonymous bulletin boards where users can post text and images. Individual boards focus on a range of topics, such as politics, fitness, and animals. Active threads, or those receiving replies, remain at the top of their boards while inactive threads are deleted as new content is posted. The website is famous for its toxic culture and 'trolling'.

4chan uses a team of volunteers, comprised of moderators and less-powerful 'janitors,' to enforce its rules. The site's global rules[36] that apply across all boards state that any content that violates United States law, such as posts promoting terrorism or engaging in sex trafficking, will be removed. Doxxing, or maliciously attempting to reveal/request someone's identity online, is also prohibited along with advertising, spamming, posting as a minor, etc. The site-wide rules also state that racism is allowed within the Random board (/b/), but not on other boards. However, moderators across 4chan boards are notoriously lenient regarding hate speech. This is particularly the case in the Politically Incorrect (/pol/) board, as racism, hate speech, sexism, and xenophobia are prevalent on the board and are not moderated. Scholars surmise that 4chan allows this type of content in /pol/ to reduce its presence on other boards (Hine et al. 2017). The Politically Incorrect board's

[35] https://gab.com/about/tos
[36] https://www.4channel.org/rules

only community-specific rules state that pornography is prohibited and that users are free to speak their mind, but are not allowed to attack other users. Other boards generally do not allow political discussion and direct posters to use /pol/ for that purpose. We provide an example of a thread from the /pol/ board in figure A.3.



File: FB_IMG_1600696385812.jpg (27 KB, 720x463)

Dan Rather
@DanRather

There was a time in American history when really only white men could vote. Imagine where we'd be today if that was still the case.

1:14 PM · 6/6/18 · Twitter for iPhone

□ Anonymous (ID: IHbwzPWt) 09/21/20(Mon)09:53:26 No.278525430 [Reply] ▶ >>278527087 >>278527123 >>278527165 >>278527183

Imagine.

7 replies and 1 image omitted. Click here to view.

>> □ Anonymous (ID: TmfVYQDK) 09/21/20(Mon)10:09:31 No.278527087 ▶

>>278525430 (OP)
I'd be wind surfing on Saturn

>> □ Anonymous (ID: P0P0wnWC) 09/21/20(Mon)10:09:46 No.278527112 ▶

We sacrificed the stars for this...

>> □ Anonymous (ID: ssy46ZIP) 09/21/20(Mon)10:09:50 No.278527123 ▶
File: F4A22B72-58C4-4ADD-B257-1(...).jpg (2.98 MB, 2354x3000)

>>278525430 (OP)
Imagine

>> □ Anonymous (ID: gVg3330w) 09/21/20(Mon)10:10:20 No.278527165 ▶

>>278525430 (OP)
Colonies on the moon

>> □ Anonymous (ID: jlabl6jZ) 09/21/20(Mon)10:10:28 No.278527183 ▶

>>278525430 (OP)
there was a time in American history when the media didnt lie, imagine where we'd be today if that was still the case

Figure A.3: The front page of 4chan's /pol/ board on September 21, 2020 (this is the first thread with no racial slurs or expletives).

A.3 Measuring Hate: Hate Crime Databases

We describe the various datasets regarding hate crimes and bias incidents analyzed in the main paper in this section. We summarize how each source defines these events and what is required for an incident to be included in them in table A.4. More information about each dataset is available in their respective sections below.

Dataset	Primary Definition	Primary Inclusion Rule
ADL	Victim reports bias in a criminal/non-criminal incident	Victim reports event to ADL or files with police
CAIR	Victim reports bias in a criminal/non-criminal incident	Victim reports event to CAIR
FBI	Criminal offense motivated [partly] by victim's identity	Victim files with local police

Table A.4: Definitions and primary inclusion rules of hate crimes/bias incidents across the datasets we analyze.

A.3.1 Anti-Defamation League (ADL) Anti-Semitic Incidents Data

We use data collected by the Anti-Defamation League (ADL), which exists to fight both anti-Semitism and bigotry in all forms and is one of the world's largest Jewish non-governmental organizations, to measure hate crimes/bias incidents that target Jews. Specifically, we use part of ADL's Hate, Extremism, Antisemitism, Terrorism (HEAT) dataset. The HEAT data is comprised of six different individual datasets[37] regarding hate crimes, bias incidents, and hate speech in the United States. We use the anti-Semitic incidents data because testing our theory of a turning point led by anti-Semitic rhetoric at Unite the Right requires tracking changes in events that target Jews.

The anti-Semitic incident data includes both criminal and non-criminal events involving harassment, vandalism, and assault/violence that "1) include circumstances indicating anti-Jewish animus

[37] Extremist-related murders, Terrorist plots and attacks, Extremist shootouts with police, White supremacist propaganda, White supremacist rallies, and anti-Semitic incidents

on the part of the perpetrator; or 2) result in Jewish individuals or organizations being victimized due to their Jewish or perceived Jewish identity."[38] ADL's research arm uses a combination of traditional news sources, social media, law enforcement reports, extremist websites, and reports by victims to identify these events. We provide the number of the different types of events recorded in the dataset in table A.5 below.

	Type	N	Proportion
1	Harassment	2802	0.55
2	Vandalism	2236	0.44
3	Assault	94	0.02

Table A.5: Anti-Semitic incidents in the ADL data

A.3.2 Council on American-Islamic Relations (CAIR) Bias Incident Data

The Council on American Islamic Relations (CAIR) is the nation's largest Muslim civil rights and advocacy organization. As a non-profit advocacy organization, one of CAIR's missions is to advocate on behalf of Muslims and others who have experienced religious discrimination, defamation, or hate crimes. CAIR recognizes that federal law enforcement institutions, such as the FBI, have illegally surveilled and harassed American Muslims,[39] and they provide guidance and steps to help Muslims navigate those situations.

One of the tools CAIR employs in its advocacy function is its "Report an Incident page."[40] This page invites individuals who are looking for legal representation to report incidents if the person is a victim of a hate crime or several other types of incidents, such as cases that fall into the following categories: freedom of religion, equal protection, due process, freedom from government misconduct, and freedom of speech.

CAIR has published numerous reports using the data it collects from individuals who report an incident to their website.[41] The benefit of this data is that Muslims are much more likely to report incidents to CAIR rather than to law enforcement agencies, given the Muslim community's long distrust for the FBI, which has grown significantly since 9/11 given numerous surveillance programs and prolonged detentions.

CAIR provided us with their data on all hate crimes and bias incidents that were reported to them from 2014 through 2018. We then subset this data to incidents that broadly involved violence, vandalism, or harassment to match the incidence types recorded in the ADL data. This removes reported harassment listing governmental agencies as offenders. The specific categories we use, which are grouped into buckets that are similar to ADL's, and their proportions in the data are available in table A.6 below. Hate Crime here includes acts of violence against individuals, acts of property destruction/damage/vandalism, and unspecified incidents as well.

A.3.3 FBI Uniform Crime Reports Data

To further evaluate results obtained by ADL and CAIR, as well as to construct a baseline measure for race/ethnicity related hate crimes, we also use the FBI Uniform Crime Reports (UCR) database, which provides a national yearly database on hate crime statistics targeting individuals by race, ethnicity, ancestry; religion; sexual orientation; disability; gender; and gender identity.

[38] https://www.adl.org/education-and-resources/resource-knowledge-base/adl-heat-map

[39] E.g., https://www.cair.com/success_stories/protecting-you-from-illegal-surveillance/.

[40] See https://www.cair.com/report/.

[41] See, for example, http://www.islamophobia.org/images/2019/Bias_Brief/BB_2_-_FINAL.pdf, http://www.islamophobia.org/reports/228-2018-civil-rights-report-bias-incident-data.html, and https://ca.cair.com/sacval/updates/new-cair-report-shows-more-than-50-percent-spike-in-anti-muslim-incidents/.

	Type	N	Proportion
1	Harassment/Bullying/Intimidation	1506	0.68
2	Hate Crime (All Types)	535	0.24
3	Destruction/Damage/Vandalism	87	0.04
4	Physical Violence	72	0.03

Table A.6: Hate crimes and bias incidents in the CAIR data

These data depend on voluntary police reporting from sub-state jurisdictions, yielding both under- and over-reporting concerns (Freilich and Chermak 2013). To be considered as part of the FBI database, a crime must be filed with a local law enforcement agency. One limitation of these data, however, is that there is large variation by jurisdiction on what is considered to be an offense punishable by law. One example is that offenses rooted in anti-Muslim animus are punishable has hate crimes under the law in Washington D.C., though not in North Dakota (Barth et al. 2019). Although the UCR data records incidents for which there is evidence of bias (whether or not they are prosecuted as hate crimes), the FBI's data-collection efforts does still often exclude hate-motivated acts (Freilich and Chermak 2013). As such critics are quick to point that the FBI data "is so poor that its hard to know what hate crimes are happening in the U.S ... it's almost like a silent wave of crimes" (Pitofsky 2018).

Below, we report counts of offenses in the UCR data from 2014 through the end of 2018 to describe the crimes that end up in the UCR database in general. We split these data into two categories for our analyses: 1) less likely or 2) more likely to require emergency assistance and be reported as a crime. Specifically, the category 2 crimes here are: murder, aggravated assault, kidnapping, and arson. Crimes that involve multiple offenses are coded as category 2 if they involve any of the category 2 crimes (e.g. robberies are included in category 2 when they also involve aggravated assault).

Given data quality concerns – including the possibility of chilling effects on crime reporting after initial, ostensibly supportive statements toward White supremacists by the US president after Unite the Right – our analyses focus on category 2. Note that category 2 here includes a distinction between aggravated and simple assault not recorded in the advocacy organization data. In the supplementary analyses below, we compare these data to the advocacy organization reports to show that CAIR and UCR reports do not meaningfully differ except that the drop in the CAIR data is more abrupt at Unite the Right and the UCR reports see declines in anti-Jewish vandalism and intimidation from extremely high levels after the 2016 election (while ADL reports remain high), except for a spike at Unite the Right and up until another large increase after the Pittsburgh synagogue attack. We further show that there were increases in reports to advocacy organizations compared to the FBI after the 2016 election and 2017 inauguration. Overall, the UCR data is more likely to record assaults and the advocacy organization data contains many more reports of harassment (not necessarily criminal).

Analyses of ratios of hate crimes with this data used pseudo-logs (adding 1 to each weekly or monthly count before calculating a ratio or logging) to account for zeros in some periods.

Less likely to require emergency assistance and be reported as crime	More likely to require emergency assistance and be reported as crime
Destruction/Damage/Vandalism of Property (9628)	Aggravated Assault (3618)
Intimidation (8411)	Arson (190)
Simple Assault (7394)	Murder and Nonnegligent Manslaughter (43)
Burglary/Breaking & Entering (739)	Kidnapping/Abduction (40)
Robbery (671)	
All Other Larceny (605)	**anti-Jewish**
Drug/Narcotic Violations (396)	Aggravated Assault (58)
Shoplifting (227)	Arson (24)
Theft From Motor Vehicle (215)	Murder and Nonnegligent Manslaughter (1)
Not Specified (200)	
Drug Equipment Violations (162)	**anti-Muslim**
Motor Vehicle Theft (157)	Aggravated Assault (149)
Theft From Building (143)	Arson (11)
Weapon Law Violations (137)	Murder and Nonnegligent Manslaughter (2)
False Pretenses/Swindle/Confidence Game (88)	Kidnapping/Abduction (1)
Rape (64)	
Counterfeiting/Forgery (59)	
Impersonation (56)	
Credit Card/Automated Teller Machine Fraud (54)	
Fondling (49)	
Stolen Property Offenses (45)	
Theft of Motor Vehicle Parts or Accessories (32)	
Pornography/Obscene Material (20)	
Embezzlement (19)	
Identity Theft (19)	
Sexual Assault With An Object (17)	
Sodomy (16)	
Extortion/Blackmail (10)	
Wire Fraud (10)	

Table A.7: Counts of hate crimes in Uniform Crime Reports from 2014 through 2018. We classify aggravated assault, arson, murder, and kidnapping as more likely to both require emergency services and be reported as crimes. We use advocacy organization data to analyze all crimes and FBI (UCR) data to analyze the most violent crimes – ADL and CAIR do not distinguish between simple and aggravated assault. Although we do not include sexual assault as more likely to be reported, its inclusion does not affect later estimates.

B Supplementary Analyses, Validations, and Robustness Checks

B.1 News Models and Figures

This section displays counts of news articles and Google search trend data for the keywords ‘muslim’, ‘white supremacy’, ‘jew’, and ‘terrorism’/‘terrorist’. We see that mentions of Muslims and Jews increased in late 2015 and declined in 2017. Mentions of terrorism in the news begin declining in July and August 2017, and continue declining into 2018. Terrorism searches declined during the spike in ‘white supremacy’ searches and news coverage after Unite the Right, increased again in late 2018, and then declined further in 2018.

	Number of ‘Muslim’ mentions in news		’Muslim’ Google searches (US)	
quasi-Poisson	(1)	(2)	(3)	(4)
Dec. 2015 to July 2017 compared to Jan. 2014 to Dec. 2015	0.55 (0.36, 0.74) $p < 0.01$	0.55 (0.37, 0.73) $p < 0.01$	0.20 (0.04, 0.37) $p = 0.02$	0.20 (0.04, 0.37) $p = 0.02$
Aug. 2017 to Dec. 2018 compared to Dec. 2015 to July 2017	−0.65 (−0.86, −0.44) $p < 0.01$		−0.50 (−0.69, −0.30) $p < 0.01$	
Aug. 2017 to Dec. 2017 compared to Dec. 2015 to July 2017		−0.41 (−0.71, −0.11) $p = 0.02$		−0.45 (−0.75, −0.15) $p = 0.01$
Jan. 2018 to Dec. 2018 compared to Aug. 2017 to Dec. 2017		−0.36 (−0.71, −0.003) $p = 0.06$		−0.07 (−0.41, 0.27) $p = 0.70$
Constant	6.92 (6.78, 7.07) $p < 0.01$	6.92 (6.78, 7.06) $p < 0.01$	3.76 (3.64, 3.87) $p < 0.01$	3.76 (3.64, 3.88) $p < 0.01$
Months	60	60	60	60

Table B.1: Shifts in mentions of Muslims in the news and Muslim searches on Google – quasi-Poisson count model by month. Due to data collection restraints, unlike in other analyses, searches for ‘muslim’ were limited to ‘muslim’ and ‘jewish’ to ‘jew’.

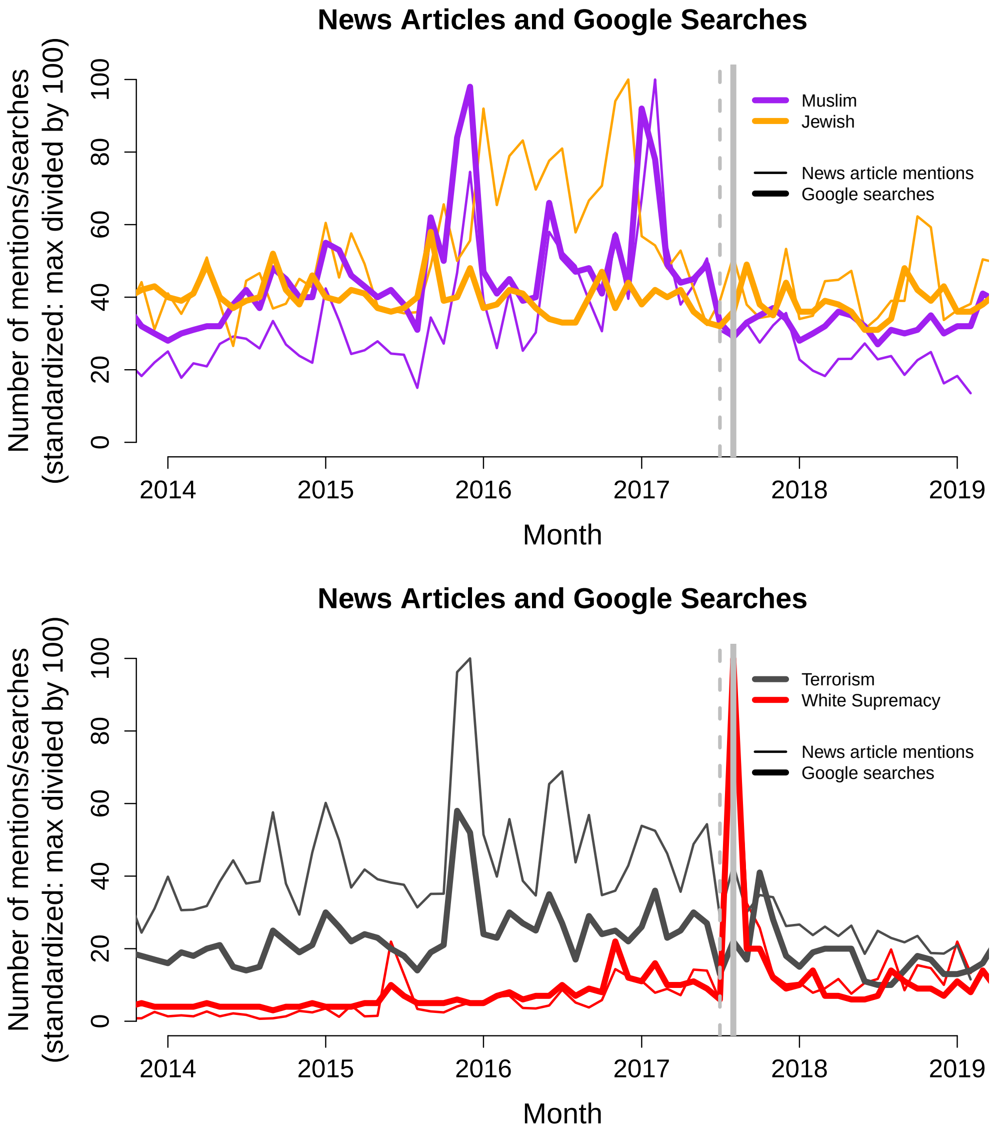


Figure B.1: In news data and in Google search trends, we see declines in Muslim searches about a month prior to Unite the Right. Each of the trends is standardized with 100 as its maximum value.

B.2 Hand-labeled Social Media Posts

We collected hand labels of the social media comments via Amazon Mechanical Turk (MTurk, www.mturk.com). MTurk is a crowdsourcing marketplace where researchers can post their tasks for a distributed workforce to complete virtually. Because 4chan has existed longer than Gab we randomly sampled different numbers of texts according to the size of each corpus. This ensured a similar data density over time. Specifically, we sampled 25,000 4chan comments from 2014 to 2019 and 10,000 Gab comments from 2016 to 2018. We did not include the Reddit corpus in this task due to its comparatively high levels of content moderation. Note that we did not observe meaningful differences (see figures below) in the 4chan and Gab data in hand labeling versus keyword based analyses when considering the *ratio* of 'Muslims or Arabs' to 'Jews' mentions – hand labeling identified more posts without altering long-term trends.

All workers (N=708) were recruited from Human Intelligence Tasks (HITs) posted to MTurk. Before deciding whether to participate in this study, participants could see the title of the task, a description, a preview of the task, and the payment they would earn for completing one HIT. In each HIT, workers read and labeled 10 comment texts. Each HIT started with an instruction page containing a consent message stating that the comment texts are potentially disturbing and participants may terminate the task at any time, followed by 10 text coding pages. For each text, participants were asked to choose all the following categories the text references (see Fig. S3 for the full instructions): Asians, Blacks, Latinos/Hispanics, Whites, Christians, Jews, Mormons, Muslims or Arabs, Immigrants, Refugees, No mentions. We asked more categories than necessary for the validation so that the codes were comprehensive (e.g. covering major religious groups and race/ethnicity groups collected by the U.S. census) and could be used in follow-up studies.

Participants earned 0.2 USD for each HIT as a compensation for their time and each participant could complete a maximum number of 30 HITs (300 texts). The median completion time for a HIT of 10 texts was 94 seconds, converting to an estimated wage of 7.66 USD per hour. To ensure the quality of the participants, they were screened to make sure that they were in the U.S., had previously completed at least 5000 HITs, and had an MTurk HIT approval rate of greater than 95%. Inter-rater reliability was calculated by Fleiss' Kappa (Fleiss 1971) using Python's Natural Language Toolkit package (Bird, Klein, and Loper 2009). See Table B.2 for the inter-rater reliability scores of all categories.

Table B.2: Inter-rater Reliability (Fleiss' Kappa)

Categories	**Gab**	**4chan /pol/**
Jews	0.74	0.70
Muslims or Arabs	0.67	0.60

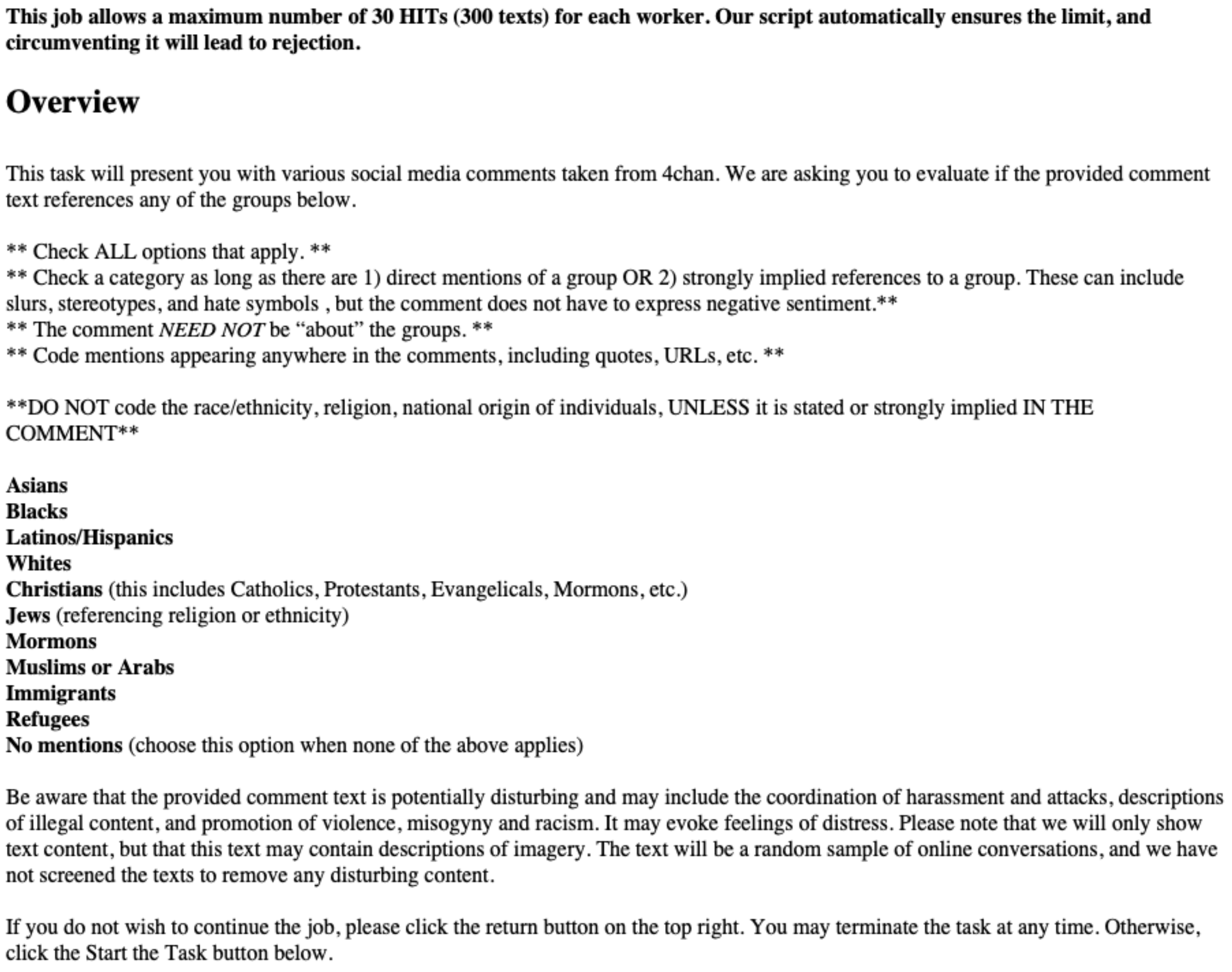
B.3 Group Label Prediction

We used fine-tuned BERT (Bidirectional Encoder Representations from Transformers) models to predict group labels of the Gab and 4chan corpora (Devlin et al. 2018). BERT is a pre-trained transformer language model that has the advantage of conditioning on context from both directions in the text to generate deeper context representations. It has been shown to outperform previous models in various natural language processing tasks (Devlin et al. 2018).

We used the 12 layer bert-base-uncased model for our prediction. After tokenization by the BERT tokenizer, texts were truncated or padded to length 128 (approximately the 99% quantile of all text lengths). A training test split of 0.9-0.1 was applied to the labelled data. We ran a separate fine-tuning process for each category using a similar procedure. Following the recommendation of the authors of the original BERT paper for model fine-tuning (Devlin et al. 2018), we searched through learning rates of 2e-5, 3e-5, 5e-5 and 1e-4; batch sizes of 16 and 32; and number of epochs 2, 3 and 4. During each epoch, the training data was further randomly split into training and validation sets of proportion 0.9 and 0.1. We evaluated the f1 score of all trained models on the hold out test set, and the best model from each category was used for predictions. See Table B.3 for model parameters and performance scores on different categories.

Table B.3: BERT Model Parameters and Performance

Categories	**Parameters**			**Performance**		
	Learning Rate	Batch Size	n Epochs	Precision	Recall	F1 Score
Jews	3e-5	32	4	0.96	0.83	0.89
Muslims	5e-5	32	4	0.85	0.85	0.85



This job allows a maximum number of 30 HITs (300 texts) for each worker. Our script automatically ensures the limit, and circumventing it will lead to rejection.

Overview

This task will present you with various social media comments taken from 4chan. We are asking you to evaluate if the provided comment text references any of the groups below.

** Check ALL options that apply. **
** Check a category as long as there are 1) direct mentions of a group OR 2) strongly implied references to a group. These can include slurs, stereotypes, and hate symbols , but the comment does not have to express negative sentiment.**
** The comment *NEED NOT* be “about” the groups. **
** Code mentions appearing anywhere in the comments, including quotes, URLs, etc. **

DO NOT code the race/ethnicity, religion, national origin of individuals, UNLESS it is stated or strongly implied IN THE COMMENT

Asians
Blacks
Latinos/Hispanics
Whites
Christians (this includes Catholics, Protestants, Evangelicals, Mormons, etc.)
Jews (referencing religion or ethnicity)
Mormons
Muslims or Arabs
Immigrants
Refugees
No mentions (choose this option when none of the above applies)

Be aware that the provided comment text is potentially disturbing and may include the coordination of harassment and attacks, descriptions of illegal content, and promotion of violence, misogyny and racism. It may evoke feelings of distress. Please note that we will only show text content, but that this text may contain descriptions of imagery. The text will be a random sample of online conversations, and we have not screened the texts to remove any disturbing content.

If you do not wish to continue the job, please click the return button on the top right. You may terminate the task at any time. Otherwise, click the Start the Task button below.

Figure B.2: MTurk instructions.

B.4 Hate Speech Label Prediction

For the prediction of hate speech labels, we used a publicly available, annotated dataset, the Gab Hate Corpus (Kennedy et al. 2018). The Gab Hate Corpus contains 27,665 Gab posts with at least 3 non-hyperlink tokens. These posts were randomly sampled from the public Pushshift.io dump. Each post was annotated by at least 3 trained annotators on multiple categories. Specifically, we were interested in the Hate category, which is the conjunction of two subcategories: CV (a 'Call for Violence') and HD (an 'Assault on Human Dignity'). CV marks posts which advocate aggression towards a given group, and HD marks posts that assault the dignity of a given group. See Table B.4 (Kennedy et al. 2018, p. 15) for the inter-rater reliability scores of CV and HD categories.

Table B.4: Inter-rater reliability Gab Hate Corpus (Kennedy et al., 2018)

Categories	**Fleiss**
Call for violence	0.28
Assault on human dignity	0.23

The final label of each post was computed using a majority vote of the annotators, with ties broken towards positive, as in the original article (Kennedy et al. 2018). Hate speech label prediction was performed with BERT model, using a similar procedure as described in section B.3. See Table B.5 for model parameters and prediction performance.

Table B.5: BERT Model Parameters and Performance

Categories	**Parameters**			**Performance**		
	Learning Rate	Batch Size	n Epochs	Precision	Recall	F1 Score
Hate	2e-5	16	2	0.56	0.48	0.50

B.5 Supplemental Social Media Analysis Figures

B.5.1 Reddit by Subreddit Category – Number of Users and Number of Posts

Reddit contains a large number of 'subreddits' on specific topics. Here, we compare shifts in keyword mentions of Muslims and Jews across all of Reddit and for the most common subreddits dedicated to discussions of politics and religion/atheism. These most common subreddits are drawn from the top 500 subreddits by post count. Table B.6 displays the subreddits categorized as being focused on politics or religion.

Politics	**Religion**
politics	atheism
The_Donald	exmormon
neoliberal	Christianity
SandersForPresident	Catholicism
Libertarian	DebateReligion
PoliticalDiscussion	
Conservative	
worldpolitics	
WayOfTheBern	
POLITIC	

Table B.6: Politics and religion subreddits in top 500 most posted subreddits, ordered by number of posts in sample.

Results displayed in the main text show over-time patterns in the ratio of 'Muslims or Arabs' to 'Jews' mentions. Below, we also display the fraction of posts about each group and the fraction of users posting about each group. Y-axes are scaled to the same levels across each data source (i.e. same y-axes for 4chan /pol/, Gab, Reddit).

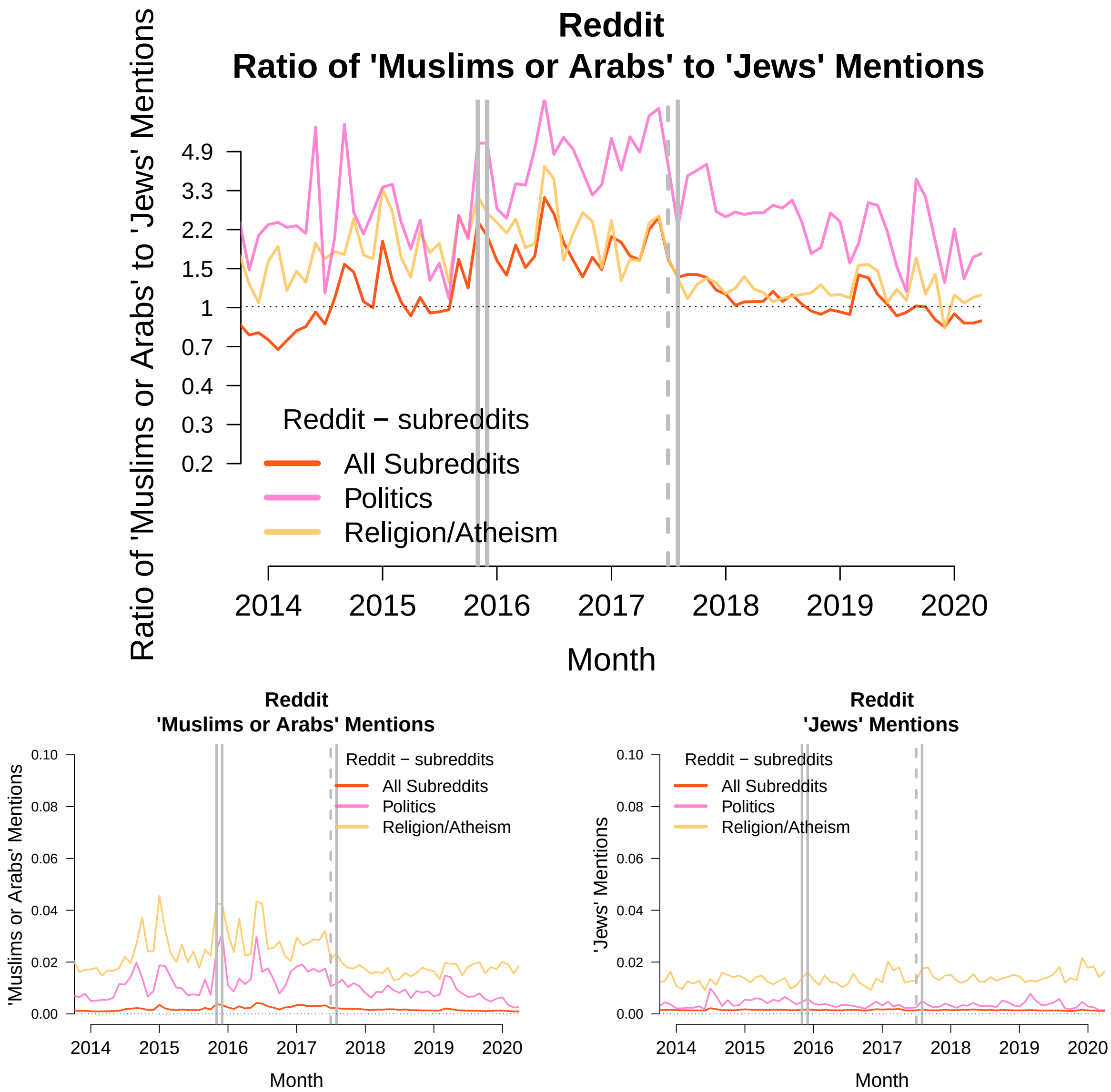


Figure B.3: *Reddit post mentions: (aggregate) ratio of 'Muslims or Arabs' to 'Jews', 'Muslims or Arabs' only, 'Jews' only.* This figure displays the fractions of posts by month that mention either 'Muslims or Arabs' or 'Jews' for all of Reddit, politics subreddits, and religion/atheism subreddits. Keywords for 'Muslims or Arabs' were 'muslim'/'islam'/' arab ' and for 'Jews' 'jew'/'judaism'.

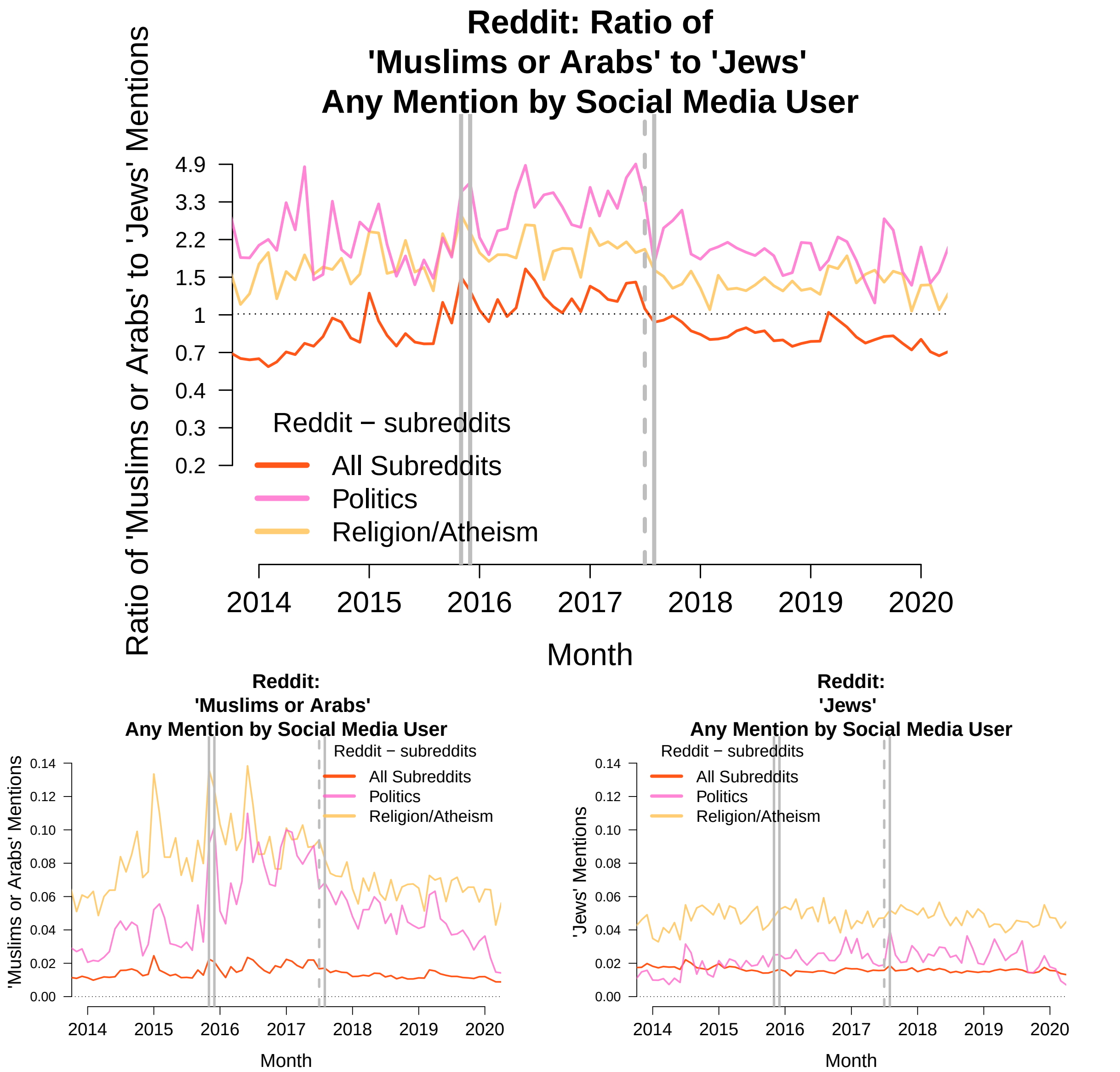


Figure B.4: *Reddit user mentions: (aggregate) ratio of 'Muslims or Arabs' to 'Jews', 'Muslims or Arabs' only, 'Jews' only.* This figure displays the fractions of users by month that mention either 'Muslims or Arabs' or 'Jews' for all of Reddit, politics subreddits, and religion/atheism subreddits. Keywords for 'Muslims or Arabs' were 'muslim'/'islam'/' arab ' and for 'Jews' 'jew'/'judaism'.

B.5.2 Gab by Group – Number of Users and Number of Posts

Results displayed in the main text show over-time patterns in the ratio of 'Muslims or Arabs' to 'Jews' mentions. Below, we also display the fraction of posts about each group, the fraction of users posting about each group, and the number of posts on Gab over time. We further display the raw hand labels compared to predicted labels. Here, there was an increase in mentions of Jews in early 2017 that was not reflected in the predicted labels or in keyword-based labels. This does not alter our interpretation of long-term shifts in mentions.

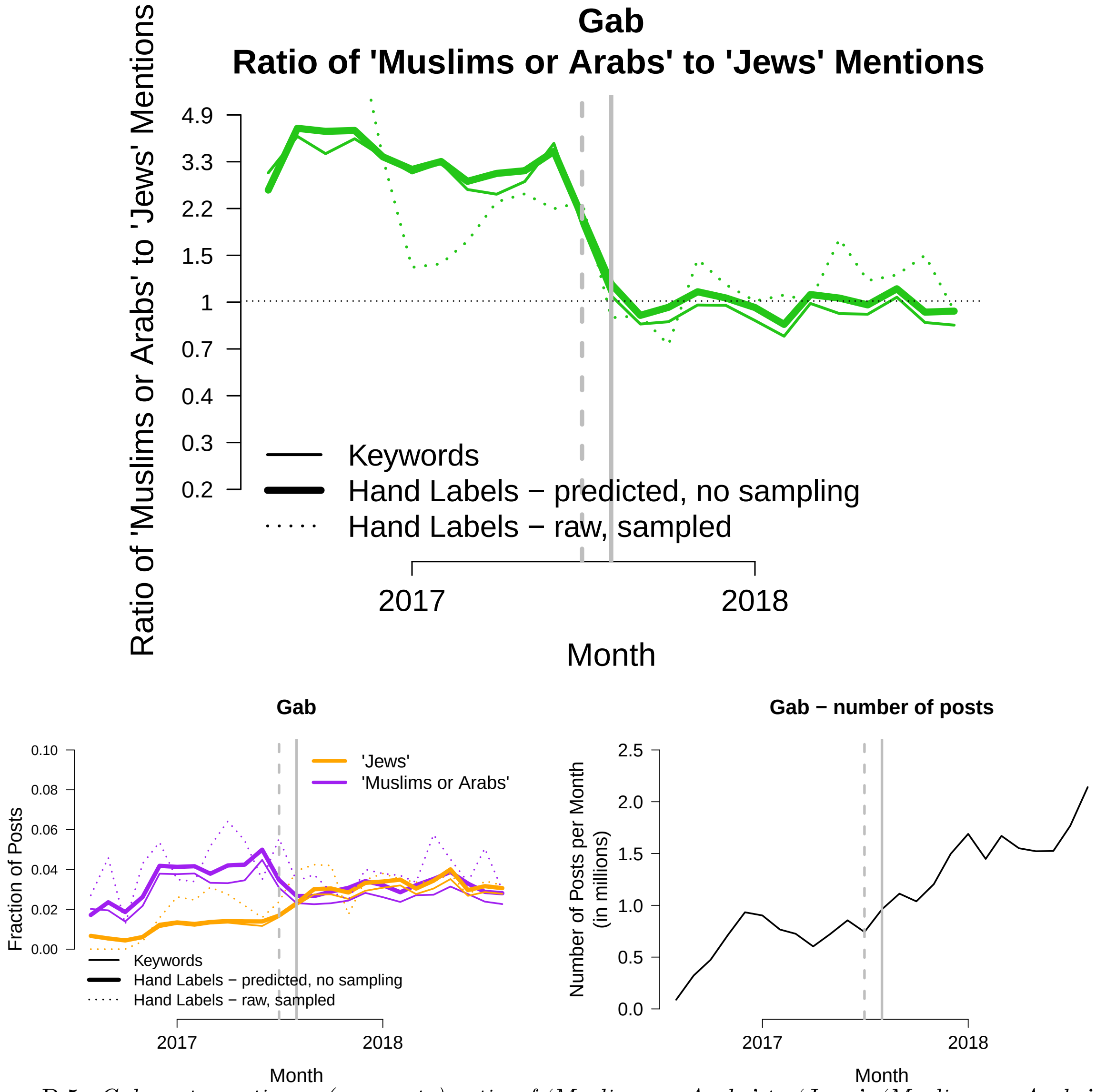


Figure B.5: *Gab post mentions: (aggregate) ratio of 'Muslims or Arabs' to 'Jews', 'Muslims or Arabs' only, 'Jews' only.* This figure displays the fractions of posts by month that mention either 'Muslims or Arabs' or 'Jews' on Gab. Hand labels were collected through crowd-sourcing.

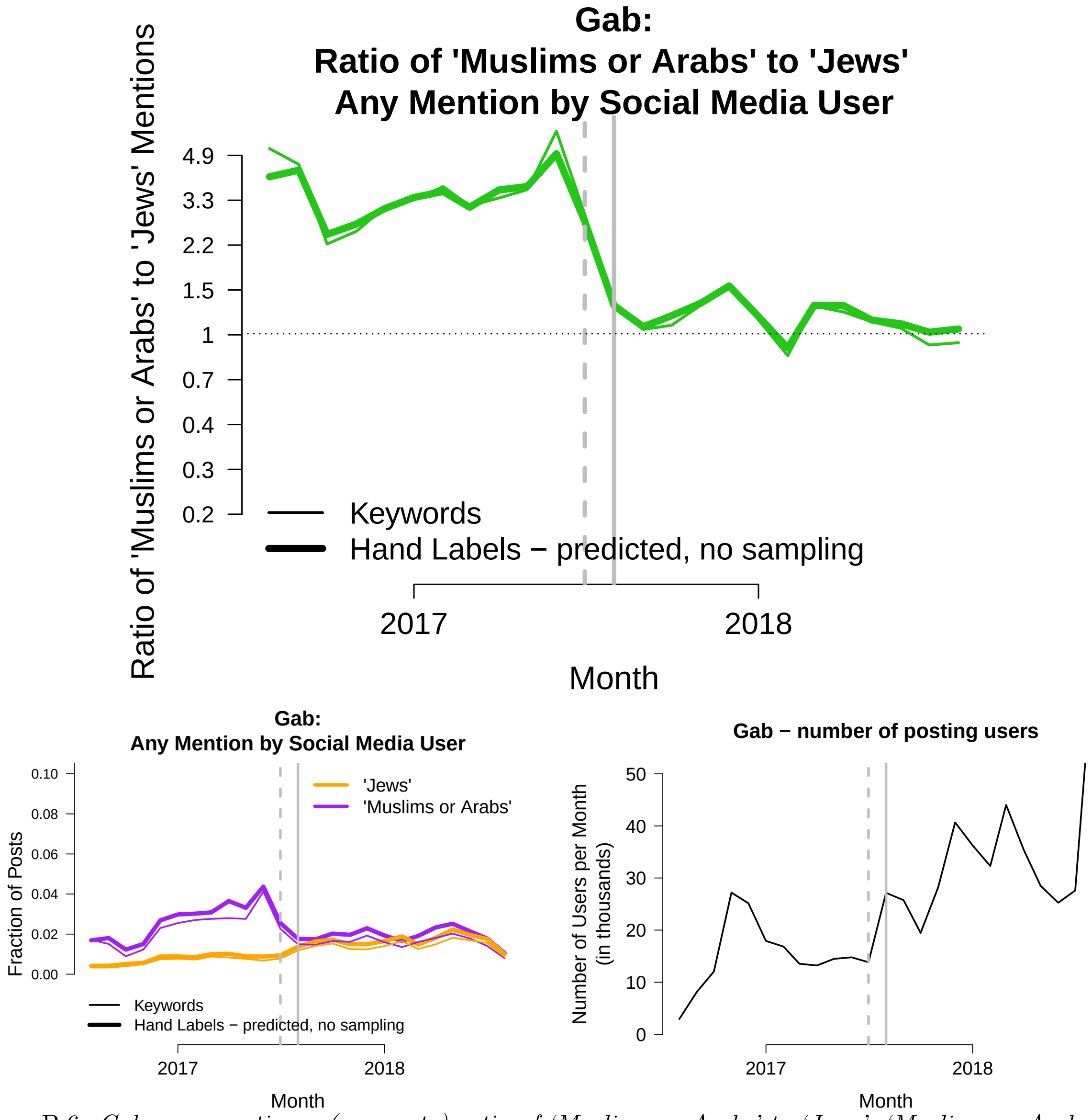


Figure B.6: *Gab user mentions: (aggregate) ratio of 'Muslims or Arabs' to 'Jews', 'Muslims or Arabs' only, 'Jews' only.* This figure displays the fractions of users by month that mention either 'Muslims or Arabs' or 'Jews' on Gab.

B.5.3 Gab – Unambiguously Negative Posts

For our analyses, we consider any mention of a racial, ethnic, or religious group on a fringe platform as a measure of the salience of that group among extremists, as well as possible hate speech. We focus on salience given the difficulty of labeling any specific post as hate speech without considering its broader context (a user's self-description or full post history, for example), and because some users may falsely claim to be a member of a group when making negative statements about it.

Nonetheless, to assess whether the vast majority of posts were negative mentions about groups, we hired undergraduate research assistants to code the crowd-sourced hand labeled group mentions for whether the text contained an "unambiguously negative" statement about the group. This labeling did *not* count comments with seemingly negative content that might also be neutral or positive.

In this coding, 90% of the posts about Jews, Muslims, and/or Arabs were considered unambiguously negative by at least one of the two coders, and 62% of the posts by both coders. We show below that neutral, positive, or possibly negative statements about Jews tended to occur prior to Unite the Right.

		Coder1	
		Neutral, Positive, or Possibly Negative	Unambiguously Negative
Coder2	Neutral, Positive, or Possibly Negative	67	151
	Unambiguously Negative	41	428

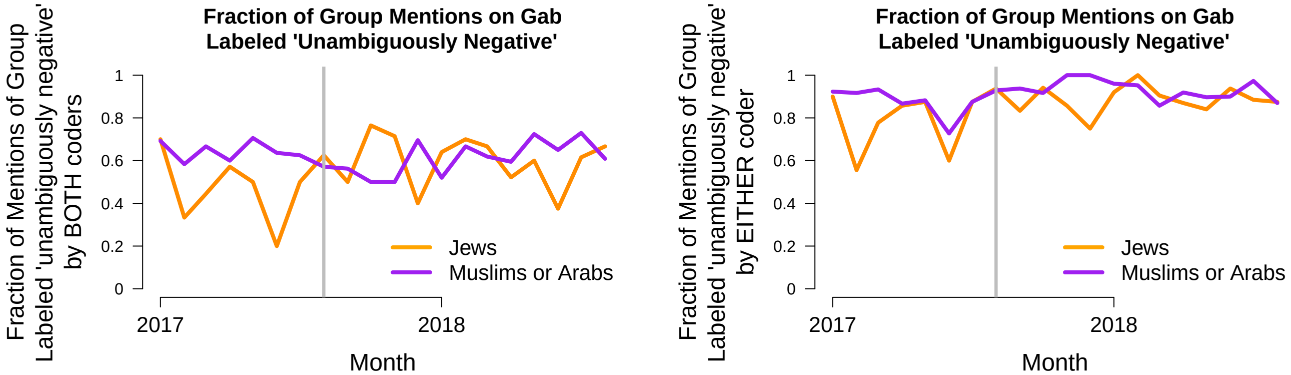


Figure B.7: Fractions of mentions of 'Jews' and 'Muslims or Arabs' that were labeled 'unambiguously negative'.

Coders were in higher agreement when attempting to label whether a Gab *profile* contained alt-right, White nationalist, or White supremacist "branding" – for example in a self-description or in posted content. Specifically, the undergraduate research assistant coders were instructed: "If you can see something on their Gab profile, (e.g., about or posts, etc), please indicate whether you see any far-right/neo-nazi/alt-right branding on their site in the "Gab profile/Altright branding" column. Please mark 1 if you see it and 0 if you do not. We are trying to understand if this person advertises belonging to that group or agreeing with the ideology explicitly in their posts (e.g. we want to see if this is how they advertise themselves)". This labeling used the same crowd-sourced hand labeled group mentions as above.

There, 82% of the accounts appearing in our group mention sample were labeled alt-right by both coders, and 93% by at least one coder.

As an objective though noisy measure of these users' relative extremity, we pasted the usernames directly into Twitter to check whether the associated username was suspended. Of those usernames

		Coder 1	
		No Alt-Right Branding	Contains Alt-Right Branding
Coder 2	No Alt-Right Branding	28	38
	Alt-Right Branding	12	350

Table B.7: Whether user profiles on Gab were branded ‘alt-right’ (according to hand labeling)

with a matching account on Twitter (we did not attempt to verify that it was the same person), 24% were suspended on Twitter.

B.5.4 4chan /pol/ by Group – Number of Posts

Results displayed in the main text show over-time patterns in the ratio of ‘Muslims or Arabs’ to ‘Jews’ mentions. Below, we also display the fraction of posts about each group and the number of posts on 4chan /pol/ over time. We further display the raw hand labels compared to predicted labels.

4chan /pol/ is a largely anonymous imageboard and so we cannot analyze the fractions of users who mentioned a specific group.

Note that, here, both mentions of ‘Jews’ and ‘Muslims or Arabs’ increased in early 2017, and only mentions of Jews continued to increase in July and August 2017. That is, we see an increase in mentions of Jews on 4chan /pol/ at or just prior to Unite the Right is large *only relative* to mentions of Muslims. Further, much of the shift from mentions of Muslims to mentions of Jews occurs from mid 2017 and continues to the end of 2019.

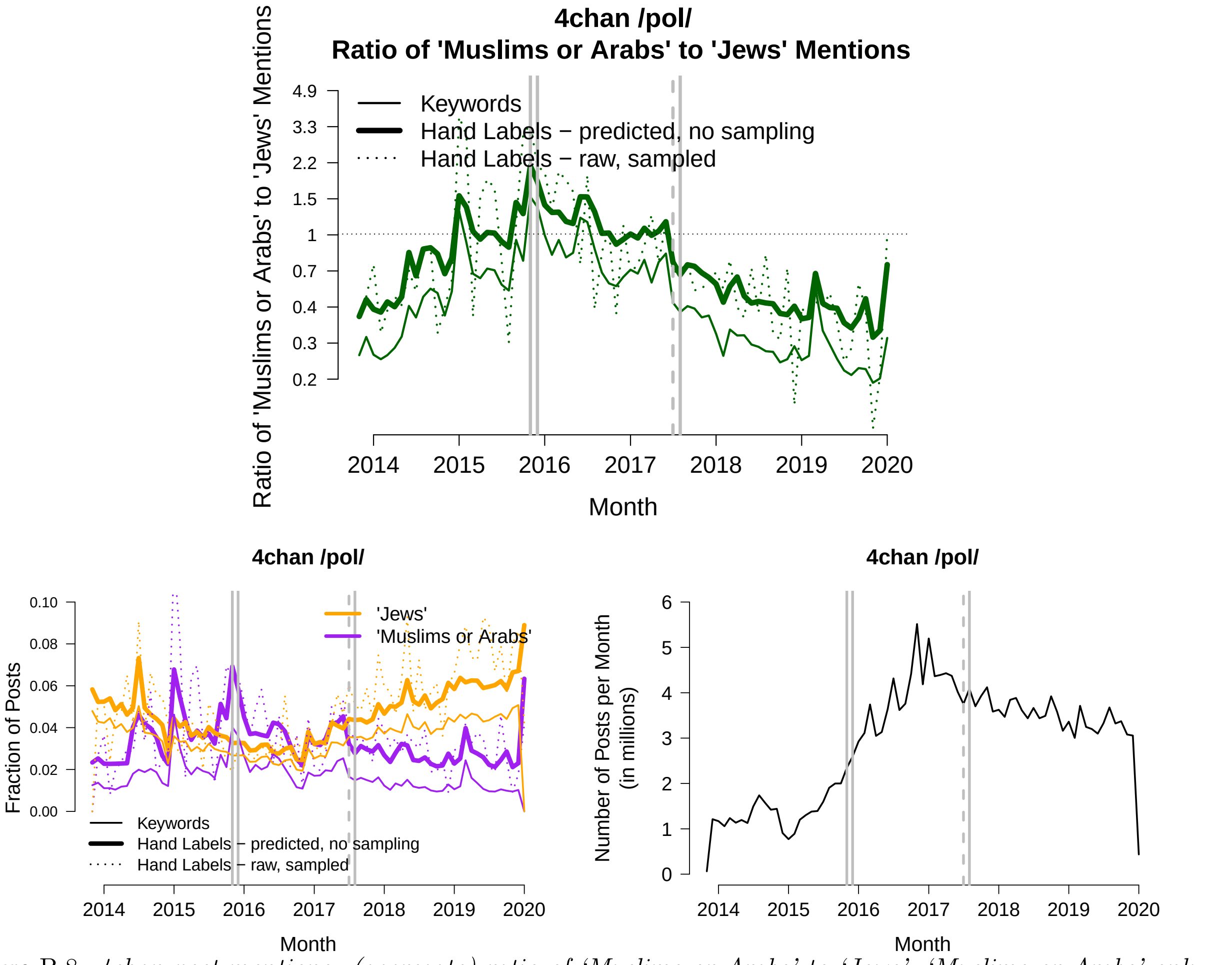


Figure B.8: *4chan post mentions: (aggregate) ratio of ‘Muslims or Arabs’ to ‘Jews’, ‘Muslims or Arabs’ only, ‘Jews’ only.* This figure displays the fractions of posts by month that mention either ‘Muslims or Arabs’ or ‘Jews’ on 4chan /pol/.

B.6 Aggregate Count Models

This section displays logged incidence rate ratio estimates (coefficients from quasi-Poisson count models by month) for the shifts in social media content shown in figures in the prior sections. Each model includes indicator variables for on or after December 2015 (during which there was a terrorist attack in San Bernardino, CA and a statement by then presidential candidate Donald Trump "calling for a total and complete shutdown of Muslims entering the United States") and on or after August 2017 (Unite the Right). In this specification, the coefficient for on or after August 2017 estimates the relative change for levels from December 2015 through July 2017. Dependent variables in these models use the monthly means of predicted group mentions (see Section B.2 and Section B.3).

DV: fraction of POSTS	'Muslims or Arabs'		'Jews'	
mentioning group	Gab	4chan /pol/	Gab	4chan /pol/
quasi-Poisson:	(1)	(2)	(3)	(4)
Dec. 2015 to July 2017 compared to Jan. 2014 to Dec. 2015		0.01 (−0.16, 0.18) p = 0.91		−0.31 (−0.43, −0.19) p < 0.01
Aug. 2017 to Dec. 2018 compared to Dec. 2015 to July 2017 (Gab start: Dec. 2016)	−0.29 (−0.38, −0.19) p < 0.01	−0.30 (−0.47, −0.12) p < 0.01	0.83 (0.70, 0.96) p < 0.01	0.54 (0.43, 0.66) p < 0.01
Intercept	−3.18 (−3.25, −3.11) p < 0.01	−3.30 (−3.41, −3.19) p < 0.01	−4.28 (−4.40, −4.17) p < 0.01	−3.11 (−3.18, −3.04) p < 0.01
Months	21	75	21	75

Table B.8: Shifts in mentions of 'Muslims or Arabs' and 'Jews' on Gab and 4chan – quasi-Poisson count model by month.

DV: fraction of USERS	'Muslims or Arabs'	'Jews'
mentioning group	Gab	
quasi-Poisson:	(1)	(2)
Aug. 2017 to Dec. 2018 compared to Dec. 2016 to July 2017	−0.51 (−0.67, −0.35) p < 0.01	0.59 (0.43, 0.75) p < 0.01
Intercept	−3.44 (−3.55, −3.33) p < 0.01	−4.69 (−4.83, −4.56) p < 0.01
Months	21	21

Table B.9: Shifts in fractions of users mentioning 'Muslims or Arabs' and 'Jews' on Gab – quasi-Poisson count model by month.

DV: logged fraction of	ln('Muslims or Arabs') - ln('Jews')	
POSTS mentioning group	Gab	4chan /pol/
OLS:	(5)	(6)
Dec. 2015 to July 2017 compared to 2014 to Dec. 2015		0.35 (0.16, 0.54) $p < 0.01$
Aug. 2017 to Dec. 2018 compared to Dec. 2015 to July 2017 (Gab start: Dec. 2016)	−1.11 (−1.23, −1.00) $p < 0.01$	−0.84 (−1.02, −0.66) $p < 0.01$
Intercept	1.10 (1.01, 1.19) $p < 0.01$	−0.23 (−0.35, −0.10) $p < 0.01$
Months	21	75

Table B.10: Shifts in mentions of Muslims or Arabs and Jews on Gab and 4chan – linear regression on logged counts by month.

B.7 Within-User Models

Our within-user models evaluate to what extent the same Gab users 1) mentioned Jews more often or mentioned Muslims less often and 2) shifted from one group to another (rather than to or from not mentioning a group at all). A standard difference-in-difference model will not separate these two effects. We first display results on the overall within-user shift, and then proceed to a model testing the substitution effect hypothesis.

For the substitution effect test, we use an errors-in-variables regression. In this, we assume that posts on Gab are a noisy measure of a user's long-run propensity to talk about a particular topic.

We assume this for two reasons. First, we often cannot reliably estimate an individual user's propensity to mention one group, for example, 4% of the time and then 2% of the time, and another group mentioning 1% to 2%. For the vast majority of users, we observe only one to a handful of posts; we cannot determine whether someone who posts 1/1 to 0/1 about Muslims before and after Unite the Right and then posts 0/1 and 1/1 about Jews did so coincidentally or if this reflects a shift in their propensity to mention one group or another.

Second, more generally and perhaps more importantly, our target substitution estimates should account for how consistently we can expect a user to talk about the *same* group over time, even when posting very frequently. Once we estimate the persistence of shifts in mentions of the same group, we can then assess the magnitude of shifts from mentions of one group to another relative to within-group mention persistence.

To accomplish this comparison, our within-user estimates of target substitution draw on errors-in-variables techniques for panel data (Griliches and Hausman 1986) to estimate an adjusted effect and calculate appropriate standard errors. In this technique, two-stage least squares regression allows us to compare the changes in cross group coefficients to the same group coefficients over the same time periods – the association between 'waves' 1 and 4 and 'waves' 2 and 3 both within and across groups. For example, this approximates the coefficients we observe, when the coefficient for changes in waves 2 and 3 with waves 1 and 4 *for the same group* is 0.10 and the coefficient for changes of mentions *one group* in waves 2 and 3 *with another* in waves 1 and 4 is -0.04, then the final target substitution coefficient will be -0.40.

Because we are interested in estimating a unique effect for a specific period, we further allow for the possibility that the shifts in group mentions are correlated within individual over time absent any major event. We estimate the target substitution effect around Unite the Right (an effect that is not cancelled out in aggregate as shown in other results) using an interaction term for the June to August (instrumented using May to September) compared to equally spaced periods from December 2016 through October 2018. Further, to allow for the possibility that drops in anti-Muslim shifts occurred in both July and August, while anti-Jewish substitution effects began in August, we estimate the shift for waves 2 and 3 using the shift from June to August.

We report robust standard errors clustered at the user level. These models were estimated using the R packages AER and ivpack.

This analysis has some inherent limitations. In particular, we can only analyze these within-user shifts for individuals who posted on this far-right platform before *and* after Unite the Right, and we further limit analyses to users who posted in the four months used to estimate an adjusted effect.

Before showing the results for this within-user analysis, Figure B.9 first displays shifts in mentions of Jews and Muslims or Arabs for users who posted every month on Gab. For each point estimate, we use quasi-Poisson regression to model the number of a user's posts that mentioned Jews or Muslims/Arabs and that contained predicted hate speech (see Section B.4 – a user is included only once for each of these estimates. This result establishes that the shift from anti-Muslim or Arab to anti-Jewish speech on Gab was not solely driven by an influx of new users.

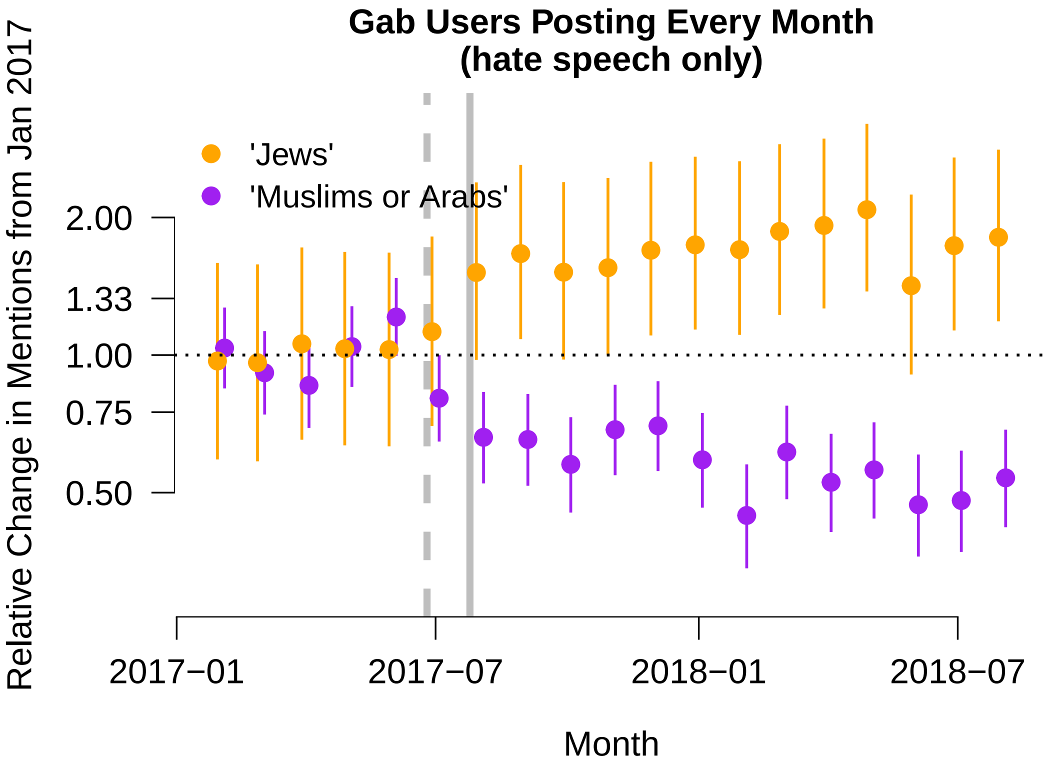


Figure B.9: Relative Change for Continuously Monthly-Active Users (n=1,359) from January 2017 through August 2018. This figure shows that the same users increased mentions of Jews and decreased mentions of Muslims and/or Arabs – i.e. the change cannot be entirely attributed due to an influx of new users. Labels of hate speech were drawn from the Gab Hate Corpus (Kennedy et al. 2018) and predicted probabilities of hate speech trained on this data (see Section B.4) were multiplied by group mention predicted probabilities to create the outcome in this figure. Table B.11 evaluates whether users shifted from discussing one group to another.

Next, Table B.11 displays the shifts in anti-Jewish and anti-Muslim mentions surrounding Unite the Right compared to other 4-wave periods on both side of the event. Like the rest of our main analyses, we log the fraction (or number – we display both models) of posts about a specific group to estimate a change in the ratio of group mentions over time. We are able to log these fractions and counts by using predicted probabilities from our supervised label prediction model – all of which are non-zero.

Before v After Unite the Right

Gab: within-user	Δln(‘Jews’$_t$)			
	predicted probabilities		keywords + 0.001	
	fraction	count	fraction	count
	(1)	(2)	(3)	(4)
Δln(‘Muslims or Arabs’$_t$)	0.44 (0.39, 0.49) $p < 0.01$	0.23 (0.14, 0.31) $p < 0.01$	0.44 (0.37, 0.50) $p < 0.01$	0.35 (0.27, 0.42) $p < 0.01$
June to Aug 2017	0.18 (0.03, 0.33) $p = 0.02$	−0.13 (−0.34, 0.08) $p = 0.22$	0.16 (0.06, 0.27) $p < 0.01$	0.08 (−0.17, 0.33) $p = 0.51$
Δln(‘Number of Posts’$_t$)		1.09 (0.94, 1.24) $p < 0.01$		0.54 (0.44, 0.65) $p < 0.01$
Δln(‘Muslims or Arabs’$_t$): June to Aug 2017	−0.36 (−0.51, −0.20) $p < 0.01$	−0.45 (−0.69, −0.22) $p < 0.01$	−0.35 (−0.51, −0.18) $p < 0.01$	−0.42 (−0.64, −0.21) $p < 0.01$
Δln(‘Number of Posts’$_t$): June to Aug 2017		0.75 (0.32, 1.18) $p < 0.01$		0.42 (0.10, 0.75) $p = 0.02$
Intercept	0.03 (0.02, 0.04) $p < 0.01$	0.06 (0.05, 0.07) $p < 0.01$	0.01 (0.01, 0.02) $p < 0.01$	0.09 (0.07, 0.10) $p < 0.01$
Number of Users	24,106	24,106	24,106	24,106

Table B.11: *Gab: within-user shifts in ‘Jews’ and ‘Muslims/Arabs’ mentions.* This table displays an errors-in-variables regression that compares the within-user consistency of shifts in same group mentions over time to shifts from one group to another (see text for more details). In models 1 and 2, variables ‘Jews’ and ‘Muslims or Arabs’ are non-zero average predicted probabilities for the ‘fraction’ model and sums of predicted probabilities for the ‘count’ model. In models 3 and 4, variables ‘Jews’ and ‘Muslims or Arabs’ are fractions of posts containing group keyword mentions for the ‘fraction’ model and counts of posts containing group keyword mentions for the ‘count’ model. Overall, when users shift toward mentioning one group they tend to also mention the other. Around Unite the Right, we estimate a separate effect where users who mentioned Muslims and/or Arabs less mentioned Jews substantially more than they otherwise would.

B.8 Time Series Tests, including Granger Causality Tests

We display results from monthly and weekly time series analyses below.

The first of these results considers contemporaneous coverage (i.e. same month) of news with keyword ‘muslim’ and terror attacks (keywords ‘muslim’, ‘islam’, ‘isis’, ‘isil’) by whether the attacks occurred during the 2016 U.S. presidential campaign (December 2015 through November 2016). This analysis is estimated on logged and differenced time series.

The second set of results tests for Granger causality between group mentions on Gab and 4chan and hate crimes/bias incidents recorded by ADL, CAIR, and the FBI.

Prior to these analyses, we tested the order of integration for each variable in the data using both augmented Dickey-Fuller (ADF) tests and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) tests. All variables were I(1). As reference, these tests involve rejecting the hypothesis that a series contains a unit root (ADF) and failing to reject the hypothesis that a series is non-stationary (KPSS). We specified our analyses of the I(1) series’ accordingly – differencing for the contemporaneous coverage model and adding a surplus (i.e. untested) lag for the Toda-Yamamoto Granger causality model.

B.8.1 Salience of Muslims/Arabs/Islam and Terrorism in News Media

Month-to-Month Change in Counts of News Articles Containing ‘Muslim’, ‘Islam’
excluding articles also mentioning ‘Islamic State’, ‘ISIS’, ‘ISIL’

	Global attacks		Attacks in the United States	
	(1)	(2)	(3)	(4)
Δln(‘Deadly Terror Attacks’$_{\mathrm{t}}$) (logged, in sd’s)	0.23 (0.01, 0.46) p = 0.05	0.27 (0.03, 0.51) p = 0.04	0.11 (−0.13, 0.34) p = 0.37	−0.08 (−0.34, 0.18) p = 0.56
U.S. Presidential Campaign		0.23 (−0.38, 0.84) p = 0.47		0.13 (−0.47, 0.73) p = 0.68
Δln(‘Deadly Terror Attacks’$_{\mathrm{t}}$): U.S. Presidential Campaign (logged, in sd’s)		−0.47 (−1.36, 0.43) p = 0.31		0.73 (0.21, 1.25) p = 0.01
Intercept	0.00 (−0.23, 0.23) p = 1.00	−0.04 (−0.29, 0.21) p = 0.75	0.00 (−0.23, 0.23) p = 1.00	−0.04 (−0.28, 0.21) p = 0.77
Months	71	71	71	71

Table B.12: *Deadly terror attacks and terrorism coverage.* This table shows that occurrences of deadly terror attacks as recorded in the Global Terrorism Database with attack description keywords ‘muslim’, ‘islam’ (including ‘Islamic State’), ‘isis’, ‘isil’ were more strongly associated with terrorism coverage containing keywords ‘muslim’ or ‘islam’ (excluding ‘Islamic State’) during the U.S. presidential election campaign (December 2015 through November 2016) than in other periods. We observe no difference in these estimates when extending the ‘campaign period’ into mid-2017.

Month-to-Month Change in Counts of News Articles Containing 'Muslim', 'Islam'
excluding articles also mentioning 'Islamic State', 'ISIS', 'ISIL'

	Global attacks		Attacks in the United States	
	(1)	(2)	(3)	(4)
Δln('Number killed in terror attacks'$_{\mathrm{t}}$) (logged, in sd's)	0.12 (−0.12, 0.35) p = 0.34	0.22 (−0.02, 0.45) p = 0.09	0.23 (0.00, 0.46) p = 0.06	−0.05 (−0.41, 0.31) p = 0.80
U.S. Presidential Campaign		0.23 (−0.37, 0.83) p = 0.46		0.22 (−0.38, 0.82) p = 0.48
Δln('Number killed in terror attacks'$_{\mathrm{t}}$): U.S. Presidential Campaign (logged, in sd's)		−1.00 (−1.75, −0.25) p = 0.02		0.46 (−0.001, 0.92) p = 0.06
Constant	0.00 (−0.23, 0.23) p = 1.00	−0.04 (−0.29, 0.21) p = 0.76	0.00 (−0.23, 0.23) p = 1.00	−0.04 (−0.28, 0.21) p = 0.78
Months	71	71	71	71

Table B.13: *Number killed in terror attacks and terrorism coverage.* This table shows that number of people killed in terror attacks as recorded in the Global Terrorism Database with attack description keywords 'muslim', 'islam' (including 'Islamic State'), 'isis', 'isil' were more strongly associated with terrorism coverage containing keywords 'muslim' or 'islam' (excluding 'Islamic State') during the U.S. presidential election campaign (December 2015 through November 2016) than in other periods. We observe no difference in these estimates when extending the 'campaign period' into mid-2017.

B.8.2 Fringe Hate Speech and Bias Incident/Hate Crimes Granger Causality Analysis

Our Granger causality tests use an approach by Toda and Yamamoto (1995) which provides valid results whether or not the studied time series are cointegrated. These tests can be used for difference stationary time series, and 'differencing' is accomplished through surplus lags for both levels of the dependent variable and levels of the independent variable(s) in the null model. These models further include linear trends and intercepts. The number of considered lags for each model (the same number of lags for both the dependent variable and the added independent variable) was selected using the Akaike information criterion (AIC). This lag was one week in the terrorism coverage models, and two weeks in the hate speech and hate crimes models. The two week lag for hate speech and hate crimes tests was drawn from the largest model (selected by AIC for the ADL/CAIR analyses and the FBI analyses separately) which included controls for terror attacks and for the ratio of media coverage of hate crimes committed against Jews to media coverage of hate crimes committed against Muslims.

In models that test the association between online hate speech and hate crimes, we included logged counts of terror attacks and media coverage of hate crimes (logged ratio of anti-Jewish to anti-Muslim hate crime articles) in the null model. These models evaluate whether lagged online hate speech significantly predicted hate crimes and bias incidents above and beyond a model that included lagged hate crimes, lagged terror attacks, and lagged hate crime coverage.

Because the ADL records many bias incidents only at the monthly level and lists those incidents as occurring on the first day of the month, we necessarily limit our analyses to ADL and CAIR hate crimes and bias incidents that were recorded on days *other than* the first day of the month. The ADL began recording some incidents at a daily level in 2017, and so our advocacy organization database analysis begins in January 2017 and continues through the end of the Gab data in August 2018.

The FBI hate crime data covers a longer period, but we have fringe social media data for all of 2015, 2016, and 2019 only for 4chan. Repeating the time frames in our other news and social media analyses, we test associations between 4chan group mentions and FBI recorded hate crimes for December 2015 through July 2017 and August 2017 through December 2018.

As in other analyses of social media data, these models use means based on predicted probabilities of group mentions on Gab and 4chan. Logged ratios of mentions were combined into a single variable for both Gab and 4chan by adding the logged ratios from each site.

log ratio of group mentions on **fringe/extremist social media** (Gab, 4chan)
↓
subsequent log ratio of **hate crimes, bias incidents** (ADL, CAIR)

	controlling for number killed in terror attacks (‘muslim’, ‘islam’, ‘isis‘/isil’)	controlling for number killed in terror attacks (‘muslim’, ‘islam’, ‘isis‘/isil’) and hate crime media coverage
χ^2 = 8.7, p = 0.01	χ^2 = 6.6, p = 0.04	χ^2 = 9.2, p = 0.01
Using the same data to test for the reverse: **hate crimes/bias incidents → group mentions**		
χ^2 = 0.8, p = 0.67	χ^2 = 0.1, p = 0.94	χ^2 = 1.1, p = 0.58

Toda-Yamamoto Granger causality test on logged time series
Order of integration/number of surplus lags=1
Number of weeks: 87
Number of lags tested: 2 (selected by AIC)

Table B.14: This table demonstrates that mentions of Muslims or Arabs or Jews on Gab and 4chan significantly predicted future hate crimes and bias incidents in the CAIR and ADL data. Number killed in terror attacks includes all attacks in the Global Terrorism Database which contained the keywords ‘muslim’, ‘islam’ (including ‘Islamic State’), or ‘isis’/‘isil’ in their attack descriptions.

log ratio of group mentions on **fringe/extremist social media** (4chan)
↓
subsequent log ratio of **hate crimes** (FBI)

		controlling for terror attacks	controlling for terror attacks and hate crime coverage
12/2015 - 07/2017	$\chi^2 = 0.0$, p = 0.99	$\chi^2 = 0.1$, p = 0.97	$\chi^2 = 0.2$, p = 0.91
08/2017 - 12/2018	$\chi^2 = 10.4$, p = 0.01	$\chi^2 = 10.0$, p = 0.01	$\chi^2 = 11.8$, p < 0.01

Using the same data to test for the reverse:
hate crimes → group mentions

12/2015 - 07/2017	$\chi^2 = 2.0$, p = 0.36	$\chi^2 = 2.1$, p = 0.35	$\chi^2 = 2.0$, p = 0.37
08/2017 - 12/2018	$\chi^2 = 6.1$, p = 0.05	$\chi^2 = 5.4$, p = 0.07	$\chi^2 = 5.7$, p = 0.06

Toda-Yamamoto Granger causality test on logged time series
Order of integration/number of surplus lags=1

Number of weeks - Dec. 2015 through July 2017: 87
Number of weeks - Aug. 2017 through Dec. 2018: 75

Number of lags tested: 2 (selected by AIC)

Table B.15: This table demonstrates that mentions of Muslims or Arabs or Jews on 4chan significantly predicted future violent and ‘likely to be investigated’ hate crimes in the FBI data. Number of terror attacks includes all attacks in the Global Terrorism Database which contained the keywords ‘muslim’, ‘islam’ (including ‘Islamic State’), or ‘isis’/‘isil’ in their attack descriptions. We use 4chan data for this test because, unlike the Gab data, it covers the longer timeframe in the FBI data.

log ratio of group mentions on **fringe/extremist social media** (4chan)
↓
subsequent log ratio of **hate crimes** (FBI)

		controlling for number killed in terror attacks	controlling for number killed in terror attacks and hate crime coverage
12/2015 - 07/2017	$\chi^2 = 0.0$, p = 0.99	$\chi^2 = 0.0$, p = 0.99	$\chi^2 = 0.1$, p = 0.94
08/2017 - 12/2018	$\chi^2 = 10.4$, p = 0.01	$\chi^2 = 8.9$, p = 0.01	$\chi^2 = 10.6$, p = 0.01
Using the same data to test for the reverse: **hate crimes → group mentions**			
12/2015 - 07/2017	$\chi^2 = 2.0$, p = 0.36	$\chi^2 = 2.3$, p = 0.31	$\chi^2 = 2.3$, p = 0.31
08/2017 - 12/2018	$\chi^2 = 6.1$, p = 0.05	$\chi^2 = 3.8$, p = 0.15	$\chi^2 = 4.8$, p = 0.09

Toda-Yamamoto Granger causality test on logged time series
Order of integration/number of surplus lags=1

Number of weeks - Dec. 2015 through July 2017: 87
Number of weeks - Aug. 2017 through Dec. 2018: 75

Number of lags tested: 2 (selected by AIC)

Table B.16: This table demonstrates that mentions of Muslims or Arabs or Jews on 4chan significantly predicted future violent and 'likely to be investigated' hate crimes in the FBI data. Number killed in terror attacks includes all attacks in the Global Terrorism Database which contained the keywords 'muslim', 'islam' (including 'Islamic State'), or 'isis'/'isil' in their attack descriptions. We use 4chan data for this test because, unlike the Gab data, it covers the longer timeframe in the FBI data.

B.9 Hate Crime and Bias Incident Models

B.9.1 ADL and CAIR Hate Crime Models

This section displays logged incidence rate ratio estimates (coefficients from quasi-Poisson count models by month) for shifts in hate crimes and bias incidents. Each model includes indicator variables for on or after December 2015 (during which there was a terrorist attack in San Bernardino, CA and a statement by then presidential candidate Donald Trump "calling for a total and complete shutdown of Muslims entering the United States") and on or after August 2017 (Unite the Right). In this specification, the coefficient for on or after August 2017 estimates the relative change for levels from December 2015 through July 2017. Because the ADL data begins in 2016 (before which the ADL currently only releases murders and extremist attacks, rather than all bias incident reports), we truncate both the CAIR and ADL models to begin on January 2016.

	CAIR	ADL	CAIR	ADL
quasi-Poisson:	(1)	(2)	(3)	(4)
CAIR (logged)			0.41 (0.19, 0.63) $p < 0.01$	
ADL (logged)				0.51 (0.07, 0.96) $p = 0.04$
Aug. 2017 to Dec. 2018 compared to Jan. 2016 to July 2017	-0.71 $(-0.96, -0.47)$ $p < 0.01$	0.16 $(-0.13, 0.45)$ $p = 0.28$	-0.80 $(-1.01, -0.59)$ $p < 0.01$	0.52 (0.11, 0.94) $p = 0.02$
Constant	4.39 (4.25, 4.52) $p < 0.01$	4.88 (4.67, 5.08) $p < 0.01$	2.40 (1.35, 3.46) $p < 0.01$	2.64 (0.68, 4.61) $p < 0.01$
Months	36	36	36	36

Table B.17: Changes in rates of hate crimes in the ADL and CAIR crime data by period (quasi-Poisson).

	ln(CAIR) - ln(ADL)
OLS:	(3)
Aug. 2017 to Dec. 2018 compared to Jan. 2016 to July 2017	-0.91 $(-1.17, -0.66)$ $p < 0.01$
Intercept	-0.44 $(-0.61, -0.26)$ $p < 0.01$
Months	36

Table B.18: Changes in rates of hate crimes in the ADL and CAIR crime data by period (OLS).

B.9.2 FBI Uniform Crime Reports (UCR) Hate Crime Models

This section displays logged incidence rate ratio estimates (coefficients from quasi-Poisson count models by month) for shifts in violent/likely to be reported hate crimes (aggravated assault, murder, arson, kidnapping). Each model includes indicator variables for on or after December 2015 (during which there was a terrorist attack in San Bernardino, CA and a statement by then presidential candidate Donald Trump "calling for a total and complete shutdown of Muslims entering the United States") and on or after August 2017 (Unite the Right). In this setup, the coefficient for on or after August 2017 estimates the relative change for levels from December 2015 through July 2017.

Below these models, we display comparisons of ADL, CAIR, and UCR data by bias incident and hate crime category, along with a table testing the significance of shifts in reporting to advocacy organizations after the 2016 election and the presidential transition.

DV: counts of aggravated assault,		
murder, arson, kidnapping	anti Muslim or Arab	anti Jewish
quasi-Poisson:	(1)	(2)
Dec. 2015 to July 2017	0.99	−0.07
compared to	(0.60, 1.41)	(−0.72, 0.58)
Jan. 2014 to Dec. 2015	p < 0.01	p = 0.85
Aug. 2017 to Dec. 2018	−0.46	0.60
compared to	(−0.84, −0.09)	(−0.01, 1.23)
Dec. 2015 to July 2017	p = 0.02	p = 0.07
Intercept	0.42	0.16
	(0.06, 0.74)	(−0.30, 0.56)
	p = 0.02	p = 0.47
Months	60	60

Table B.19: Changes in rates of hate crimes in the FBI UCR hate crime data by period.

DV: counts of aggravated assault, manslaughter, murder, arson, kidnapping	anti Muslim or Arab	anti Jewish	anti Muslim or Arab	anti Jewish
quasi-Poisson:	(1)	(2)	(3)	(4)
Dec. 2015 to July 2017 compared to Jan. 2014 to Dec. 2015	0.95 (0.57, 1.35) p < 0.01	−0.05 (−0.71, 0.60) p = 0.90	0.90 (0.53, 1.29) p < 0.01	−0.07 (−0.68, 0.54) p = 0.83
Aug. 2017 to Dec. 2018 compared to Dec. 2015 to July 2017	−0.53 (−0.90, −0.17) p = 0.01	0.62 (0.01, 1.26) p = 0.06	−0.47 (−0.84, −0.11) p = 0.02	0.47 (−0.11, 1.06) p = 0.13
All other race/ethnicity related hate crimes (logged)	0.55 (−0.03, 1.14) p = 0.08	−0.24 (−1.11, 0.65) p = 0.59	0.63 (−0.15, 1.40) p = 0.12	0.24 (−0.93, 1.40) p = 0.69
Intercept	−1.53 (−3.70, 0.56) p = 0.17	1.02 (−2.18, 4.08) p = 0.53	−1.68 (−4.41, 1.01) p = 0.24	−1.80 (−5.99, 2.27) p = 0.40
Observations	60	60	60	60
Month of Year Fixed Effects	N	N	Y	Y

Table B.20: Changes in rates of hate crimes in the FBI UCR hate crime data by period (with controls). Monthly fixed effects are included here because hate crimes in the UCR data are seasonal – there are fewer in winter.

	anti Muslim or Arab			
DV:	Agg. Assault, Manslaughter, Murder, Arson Kidnapping	Sim. Assault	Vandal.	Intim.
Dec. 2015 to July 2017 compared to Jan. 2014 to Dec. 2015	0.99 (0.60, 1.41) p < 0.01	0.61 (0.29, 0.93) p < 0.01	0.78 (0.47, 1.11) p < 0.01	0.71 (0.41, 1.02) p < 0.01
Aug. 2017 to Dec. 2018 compared to Dec. 2015 to July 2017	−0.46 (−0.84, −0.09) p = 0.02	−0.23 (−0.54, 0.08) p = 0.16	−0.58 (−0.92, −0.26) p < 0.01	−0.25 (−0.54, 0.04) p = 0.10
Intercept	0.42 (0.06, 0.74) p = 0.02	1.42 (1.16, 1.66) p < 0.01	1.49 (1.22, 1.74) p < 0.01	1.78 (1.52, 2.01) p < 0.01
Observations	60	60	60	60
Month of Year Fixed Effects	N	N	N	N

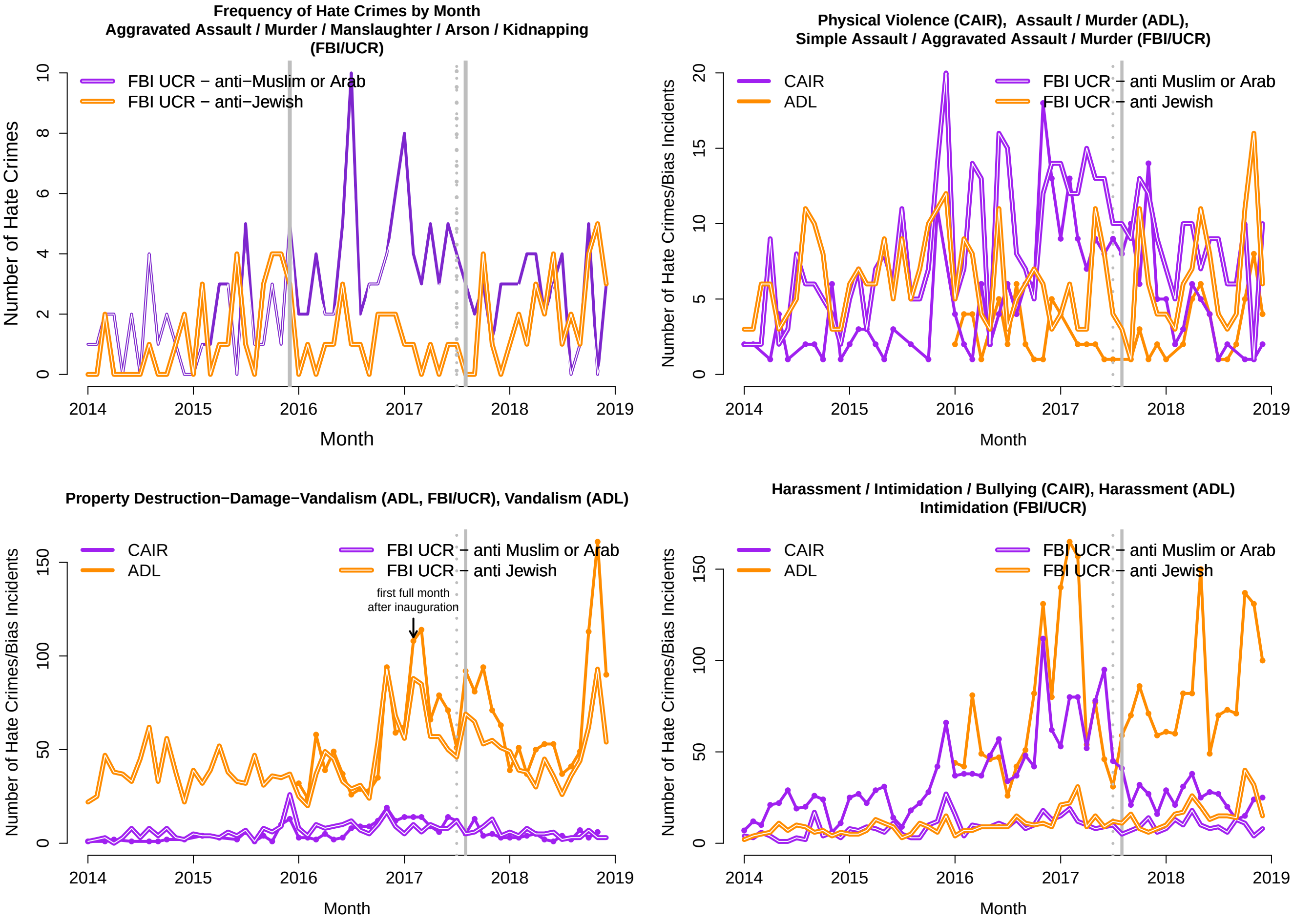
Table B.21: Changes in rates of hate crimes in the FBI UCR data by period and offense category (anti-Muslim or Arab).

	anti Jewish			
DV:	Agg. Assault, Manslaughter Murder, Arson Kidnapping	Sim. Assault	Vandal.	Intim.
Dec. 2015 to July 2017 compared to Jan. 2014 to Dec. 2015	−0.07 (−0.72, 0.58) p = 0.85	−0.03 (−0.35, 0.28) p = 0.84	0.26 (0.04, 0.48) p = 0.03	0.56 (0.22, 0.90) p < 0.01
Aug. 2017 to Dec. 2018 compared to Dec. 2015 to July 2017	0.60 (−0.01, 1.23) p = 0.07	−0.06 (−0.41, 0.29) p = 0.75	0.01 (−0.21, 0.23) p = 0.93	0.31 (0.02, 0.60) p = 0.05
Intercept	0.16 (−0.30, 0.56) p = 0.47	1.70 (1.48, 1.90) p < 0.01	3.64 (3.47, 3.79) p < 0.01	1.95 (1.67, 2.20) p < 0.01
Observations	60	60	60	60
Month of Year Fixed Effects	N	N	N	N

Table B.22: Changes in rates of hate crimes in the FBI UCR data by period and offense category (anti-Jewish).

B.9.3 Comparison of ADL, CAIR, UCR Data

This section compares the governmental hate crime data with advocacy organization hate crime and bias incidents data. These data sets largely match, but the FBI UCR data records more physical violence than ADL and CAIR (top-right panel), while ADL and CAIR record more harassment and intimidation (bottom-right panel). Beyond this, we see increases in reports to advocacy organizations not also reported in the governmental data after the 2016 election for anti-Muslim or Arab assaults (purple lines in top-right) and from 2017 on for anti-Jewish vandalism (orange lines in bottom-left). ADL and CAIR do not distinguish aggravated assault from simple assault and do not systematically record forms of property destruction, and so we cannot compare their data in the top-left panel for UCR reports of aggravated assault, murder, arson, and kidnapping.



Frequency of Hate Crimes by Month
Aggravated Assault / Murder / Manslaughter / Arson / Kidnapping
(FBI/UCR)

Physical Violence (CAIR), Assault / Murder (ADL),
Simple Assault / Aggravated Assault / Murder (FBI/UCR)

Property Destruction−Damage−Vandalism (ADL, FBI/UCR), Vandalism (ADL)

Harassment / Intimidation / Bullying (CAIR), Harassment (ADL)
Intimidation (FBI/UCR)

Figure B.10: *Comparison of FBI/UCR hate crime data with ADL and CAIR hate crime/bias incident data.* This figure compares governmental versus advocacy organization hate crime and bias incident reports. The advocacy organization data records many more instances of harassment, bullying, and intimidation than intimidation in the UCR data, while the UCR data records more assaults and physical violence. We observe larger discrepancies in reports of vandalism and assault from 2017 on, where the advocacy organizations record larger increases in vandalism (for anti-Jewish attacks) and assault (for anti-Muslim or Arab attacks) than the governmental data – see Table B.23 below. Declines in anti-Muslim or Arab physical violence (CAIR) and simple assault (FBI/UCR) appear to decline more slowly – continuing into 2018 – than other hate crimes and bias incidents (see top-right – this primarily represents simple assault rather than aggravated assault or murder, which are shown separately in the top-left panel).

	Physical Violence (CAIR)	Vandalism (ADL)
	(1)	(2)
Nov. 2016 to Dec. 2018 compared to Jan. 2014 to Oct. 2016	0.57 (0.14, 1.01) p = 0.02	
anti-Muslim or Arab assaults and murders (logged, FBI/UCR)	0.59 (0.17, 1.02) p = 0.01	
Feb. 2017 to Dec. 2018 compared to Jan. 2016 to Jan. 2017		0.29 (0.13, 0.45) p < 0.01
anti-Jewish property destruction/damage/vandalism (logged, FBI/UCR)		1.01 (0.82, 1.20) p < 0.01
Constant	−0.07 (−1.00, 0.75) p = 0.88	-0.03 (−0.78, 0.72) p = 0.63
Observations	60	36

Table B.23: *Reporting differences between advocacy organizations and governmental data after the 2016 election.* We observe larger discrepancies in reports of vandalism and assault from 2017 on, where the advocacy organizations record larger increases in vandalism (for anti-Jewish attacks) and assault (for anti-Muslim or Arab attacks) than the governmental data.