Hate in Word and Deed: The Temporal Association Between Online and Offline Islamophobia

Hate in word and deed: The temporal association between online and offline Islamophobia

Abstract

Objectives: The aim of the present study is to explore whether there is a temporal association between anti-Islamic online and offline hate, and if so in which direction.

Methods: We used data on hateful Twitter content, and hate incidents and hate crimes/offences recorded by the Metropolitan Police Service in the United Kingdom to analyse this association, using time-series analysis. This study is unique in its use of newly developed technology to undertake big data analysis with recent, disaggregated online and offline hate data.

Results: Our study examined the 'everyday' incidents of (online and offline) hate that affect communities throughout the United Kingdom and we found that anti-Islamic hate speech followed rather than preceded Islamophobic hate offline.

Conclusions: Our findings likely point to what we have referred to as *compound retaliation*, which suggests that media and social media dissemination about offline acts of hate compound already tense intergroup hostilities, providing further permission for those to express hatred online. Such a situation represents the compounding of hate and hostility through offline and online networks that are likely to be reinforcing.

Keywords

Hate crime, hate speech, Islamophobia, social media, *Twitter*

1. Introduction

Intergroup hostility (or 'hate') is a prevalent and growing phenomenon in many societies (Home Office, 2018; Federal Bureau of Investigation, 2018). Here we focus on 'hate' in its legal or

criminological sense and adopt a working definition of 'hate crime' and 'hate incident' used by the criminal justice system in England and Wales, the site of this study, which defines a hate crime as: 'any criminal offence which is perceived by the victim or any other person to be motivated by a hostility or prejudice based on a person's [perceived] race… religion… sexual orientation… disability [or]… transgender [identity]' (College of Policing, 2020). Police services are obliged to collect data on both hate crimes and hate incidents. Hate incidents are 'non-crime incidents' which are 'perceived by the victim or any other person to be motivated by hostility or prejudice based on a person's [perceived] race… religion… sexual orientation… disability [or]… transgender [identity]' (College of Policing, 2020). Examples of hate incidents include verbal abuse or acts of intimation that do not meet the threshold of a criminal offence.[1]

In England and Wales, 105,090 hate crimes were recorded by the police in 2019/2020, an increase of 8% from the previous year (Home Office, 2020). Previous data suggests that approximately 2% of recorded hate crime offences are flagged as having an online element (Home Office, 2018).[2] The vast majority of recorded offences are perpetrated 'offline' – the most common being public order offences (e.g. face-to-face verbal threatening and abusive language), physical assaults, and criminal damage. Further, police-recorded hate crime data from 2019/2020 suggests that Muslims were the most common targets of hate crimes where the perceived religion of the victim was a factor (Home Office, 2020).

Despite online hate making up a small proportion of all recorded hate crimes, it is likely that online hate is much more prevalent than offline hate and can be linked to certain offline

[1] In this paper, the terms hate crime and hate offences will be used interchangeably and are to be distinguished from hate incidents.

[2] However, such statistics are not fully developed yet and need to be viewed with caution at this point (Home Office, 2018). See outline of criminal offences in The Law Commission (2014, ch. 2) report. Even though not specifically designed to deal with online hate, the stirring up of racial, religious, or sexual orientation hatred offences under Part 3 and 3A of the Public Order Act 1986 can be used to prosecute offences committed via social media. Further, the Communications Act 2003, s. 127 or the Malicious Communications Act 1988, s. 1 can also be used to proscribe 'grossly offensive' online communication to which section 145 or 146 Criminal Justice Act 2003 sentence enhancements may be applied. However, according to Bakalis (2018, p. 87), legislation is in place 'to tackle various aspects of cyberhate, but in practice, the existing offences are difficult to use'.

events. For instance, Paterson et al. (2018, p. 22) reported that around 80 percent of Muslim respondents in their survey had experienced at least one hate incident online in the past three years (see also Zuleta and Burkal, 2017). Online incidents are likely to be most prevalent after certain 'trigger' events (e.g., post 9-11, post-Lee Rigby, post-Brexit; O'Neill, 2017; Kaplan, 2006; Cuerden and Rogers, 2017; Hanes and Machin, 2014; Williams and Burnap, 2016). Miller et al. (2016), for example, found that in a single eight-day period following the Brussels terrorist attacks in 2016, *Twitter* users sent 58,074 tweets containing an anti-Islamic slur.[3] A recent report by Sadique et al. (2018) using data from TellMAMA similarly indicated growing concern over anti-Muslim rhetoric/language on social media following trigger events.

Following terrorist attacks and the refugee crisis in Europe, a rise in Islamophobia[4] and hate crime directed at Muslims or those perceived to be Muslim has been observed in Western countries (Alam and Husband, 2013; Ciftci, 2012; Kaplan, 2006; Zunes, 2017). Recent UK-based research suggests that hate crimes motivated by religion increased significantly in England, Wales and Scotland after the 2017 terrorist attacks (Piatkowska and Lantz, 2021). Hate speech is common on news posts covering topics such as religion and foreigners, and hate speech is over-represented in debates around news posts that themselves include hateful content or quotes (Zuleta and Burkal, 2017). At the same time, the increasing diffusion of new

[3] Other research studies have also attempted to examine the nature of online hate (Awan, 2014), as well as the impacts that such incidents are likely to have (see Awan and Zempi, 2016; Paterson et al., 2018).

[4] We want to acknowledge here that there is still an ongoing debate about using the term *Islamophobia* (see, e.g., Irfan, 2021). Concerns include the vague nature of the term and that it 'target[s] expressions of Muslimness or 'perceived Muslimness', rather than bigotry against Muslim individuals themselves' (Malik, 2019). Critique of using the term *Islamophobia* to describe our *Twitter* dataset included 'conflat[ing] criticism of an idea (Islam) with abuse of people (Muslims), and that words like Islamophobia are used to shut down any kind of criticism of Islam' (Miller, 2016). Recently, the debate has shifted towards considering the term *anti-Muslim*, which refers more specifically to antipathy towards Muslims, but which has also been criticized for not encompassing non-Muslims (e.g. Sikhs) or mosques/schools as targets of such hate crimes/incidents (Irfan, 2021). This debate is currently still ongoing and is in need of further scrutiny in the future; however, as Irfan (2021) states: 'If the debate continues to focus on which term better describes the same phenomenon, we run the risk of getting lost in specifics of wordings and of policies and actions never being approved because no one can define them in one way'. In this paper, we use the terms Islamophobic (used by the Metropolitan Police Service when flagging such hate crimes) and anti-Islamic (terminology that describes our Twitter data) predominantly. In addition, the term anti-Muslim will be used when referring to the work of authors who have made specific reference to such terminology within their scholarship. Collectively, we use the terms Islamophobic, anti-Islamic and anti-Muslim to refer to online and offline incidents that are perceived to express prejudice or hostility towards Muslim people.

communication technologies, especially social networks, as alternative sources of information facilitate the emerging rise of online anti-Muslim hate (Awan, 2014). UK-based research on the media impact of online Islamophobia showed that the dissemination of topics, such as the Woolwich attack, by the print media was easily distributed online, with newspapers also having access to social media outlets and followers often tweeting and retweeting such news stories (Rahman, 2016).

In this article, we enrich the growing body of research on the link between online and offline hate by investigating whether there is a temporal association between anti-Islamic online hate and Islamophobic offline hate crimes and incidents, and if so in which direction. We first review the relevant literature on the relationship between online and offline hate and then develop the hypotheses.[5]

2. Understanding the connection between online and offline hate

Previous studies have tried to assert some connection between the online and offline hate context, which suggest that victims and their families rarely isolate experiences of online hate from offline hate (see, e.g., Awan and Zempi, 2016, 2017). Quantitative research on linking online to offline hate increasingly suggests that hateful online content and activity happens before offline hate crimes and incidents (e.g., Benesch, 2013; Hawdon, 2012; Williams et al., 2020). Maynard and Benesch (2016) suggest that 'dangerous' ideology and speech can play a critical role in the possible escalation to mass atrocities, as individual or group ideologies are communicated through speech, expressing ideological claims and providing key motives for violence. For violence to take place, 'how many [people] *partially internalize* the ideology enough to see violence as permissible or even desirable' is more important than the volume of people identifying with the communicated ideology (Maynard and Benesch, 2016, p. 74;

[5] This research was first presented at the International Network for Hate Studies Conference in Canada in May 2018 and at the Law Commission Hate Crime Research Conference in March 2019.

emphasis in the original). Research conducted by Chan et al. (2016) explored if an increase in internet access to online hate is linked to hate offline and found that, on average, internet availability has an effect on racial hate crimes, predominantly in areas with higher levels of segregation and racism. Chan et al. (2016) further suggest that access to the internet did not increase hate group operations, but did increase hate crimes committed by lone-wolf perpetrators.

Unpublished research by Müller and Schwarz (2020) also recently explored the link between social media and hate crime in Germany and the United States, hand-collecting data from Donald Trump's *Twitter* account and from Germany's right-wing party, *Alternative für Deutschland*. Specifically, the unpublished study investigated the link between anti-refugee sentiments on Facebook and hate crimes against refugees in Germany, and explored the time-series relationship between Donald Trump's anti-Hispanic and anti-Muslim tweets and hate crime against these groups (Müller and Schwarz, 2020). The study found that right-wing anti-refugee sentiment on the Alternative for Germany's Facebook page was predictive of hate crimes against refugees on the street and that Trump's *Twitter* activity was linked to hate offline (Müller and Schwarz, 2020). Müller and Schwarz (2020, pp. 40--41) therefore suggest 'social media has not only become a fertile soil for the spread of hateful ideas but also motivates real-life action'.

Finally we draw reference to Williams et al.'s (2020) study using data from an eight-month period during 2013/14 in London. The authors indicate a positive association between hate targeting race and religion online and offline, which they assert is independent of trigger events. There are, however, some concerns regarding this analysis, including the age of the recorded data and the short time period used for time-series analysis. Although *Twitter* has introduced some mechanisms to reduce the amount and spread of hate since 2015 (Twitter, 2015), our study shows that hateful content has not disappeared from *Twitter*, and the need to

research hate on this platform continues globally (see, e.g., Jaki and De Smedt, 2019; Oriola and Kotzé, 2020; Vidgen et al., 2019).

Hawdon (2012) asserts that hate-inspired actions are likely to increase through the principles of differential association, especially through associations in the virtual world. Differential association theory (Sutherland and Cressey, 1974) proposes that criminal behaviour is learned and shaped through communicative interaction with others. Such online and offline interaction can include the reading of information or seeing of images, or debate and dialogue with others, but also provides the motives, techniques and rationalisations for offline violence (Hawdon, 2012). It is a unique feature of social media communication that, for example, *Twitter* content can easily be endorsed and distributed by other *Twitter* users (Williams and Burnap, 2016).

From what we know about the effects of antilocution on people's conduct (Allport, 1954), it seems highly likely that online hate speech helps to create a normative climate in which hate behaviours become 'normalised' and acceptable. Differences exist here between 'descriptive' and 'injunctive' social norms (Cialdini et al., 1990, p. 1015). Descriptive norms are perceptions of behaviour that are seen as being typical or normal, which motivate people to also adapt that behaviour; while injunctive norms relate to perceptions of morally approved or disapproved behaviour motivated by social sanctions (Cialdini et al., 1990). It is likely that hate speech online works in both these ways, with people seeing other *Twitter* users spreading hateful content online or seeing others committing hate offline and perceiving such behaviour as typical and normal, therefore justifying their own hateful tweets and also their acting out the hate offline. If some *Twitter* users who spread hateful messages online, for example against Muslim communities, are members of elites and are hence perceived as 'legitimate' sources of influence (e.g., the President of the United States) other *Tweeters* may approve such behaviour morally and justify their own hate online and also offline (see, e.g., Müller and Schwarz, 2020). The Rohingya crisis beginning in 2016, for example, stands out as the most 'egregious and

harmful use of social media by a government', leading to the killing of 10,000 and the forced displacement of approximately 780,000 Rohingya civilians as a consequence of anti-Muslim rhetoric posted by high-ranking members of Myanmar's military and Buddhist nationalists on Facebook (Wilson and Land, 2021, p. 1043).

The direction of the online/offline hate link may however be less straightforward, based on the fact that differences exist between trigger events and various contributing factors associated with the event, such as the time lag online and offline, and the intensity and duration of the spike (Sadique et al., 2018). Hate online and offline seem to have different patterns, depending on the cultural impact of the incidents as well as its influence on media discourse, the target location and the scale of fatalities, as well as the political narrative around such incidents (Sadique et al., 2018). The myriad factors that influence both online and offline behaviour may mean that it is difficult to discern with much certainty whether one is causal of the other. Indeed, some recent research has suggested that the link between social media use and actual racist behaviour may be weaker than assumed. Alsaad et al. (2018, p. 2) explored if 'social media increase[s] racist behaviour through promoting biased beliefs among users' and found that there was no strong link between social media use and racist behaviour.

These results give rise to further questions about the predictive direction between the online and offline worlds of hate. One such question is, contrary to common assumption, whether offline hate incidents/crimes may in fact predict online hate abuse. After all, there is now a cogent body of research to suggest that physical acts of aggression ('trigger events') can spark online debate about the causes and impacts of incidents (O'Neill, 2017; Kaplan, 2006; Cuerden and Rogers, 2017; Hanes and Machin, 2014; Williams and Burnap, 2016). Indeed, the 'descriptive' and 'injunctive' power of observing something hateful happening offline may assist more cogently in the normalisation of the message expressed by such physical conduct; potentially leading some people to react to it through their everyday online world. Recent U.S. research, conducting focus groups with Muslim social media users exploring their experiences

with and responses to online and offline Islamophobia, also found a blurring of such online and offline spaces, with participants indicating that online Islamophobia increases not only after the media reports news of trigger events, but also after reports emerge of 'any Muslim in mainstream news media' (Eckert et al., 2021, p. 86).

Research on triggers for aggressive behaviour suggests that exposure to violent media can increase the propensity to commit aggressive behaviours (Anderson and Bushman, 2001). There are numerous examples of news stories about hate crime offline resulting in further incidents of similar identity-based hostility. This was noticeable after the Finsbury Park Mosque attack in London on 19 June 2017 when further Islamophobic attacks shortly followed (O'Neill, 2017), as well as the Christchurch Mosque shootings in New Zealand on 15 March 2019, where the Muslim community was re-targeted in various locations (including the UK (Dodd, 2019) and Norway (Burke, 2019)) in attempted copycat attacks (see, e.g., Roy, 2019), as well as online. The proliferation of hate-based abuse after these trigger events results in what we label as 'compound retaliation'.[6]

According to Cialdini, Kallgren, and Reno's (1991) focus theory of normative conduct, normative information has an influence on behaviours when such information is highlighted prominently in consciousness. In addition, news about hate crime or incidents offline may depict the phenomenon as alarming and widespread, creating fear and panic. This phenomenon has been attributed to *security threat perception* combined with the *ultimate attribution error* (Pettigrew, 1979; Hewstone, 1990), where perpetrators of ethnic violence (the in-group) pose a security threat, with in-group members erroneously attributing the negative behaviours by the in-group on the out-group (the victim group), negatively shaping the in-group's attitudes towards the victim group (Igarashi, 2020). Igarashi (2020, p. 14) found 'that ethnic violence

[6] Levin and McDevitt (1993) originally proposed three types of offender motivation (thrill, defensive, and mission) and later added retaliatory motivation to their original offender typology, where 'retaliatory offenders are inspired by a desire to avenge a perceived degradation or assault on their group' (McDevitt et al., 2002, p. 306). Compound retaliation is an expansion of this offender motivation.

increases the negative attitudes of the perpetrator group towards the victim group', while exploring the association between anti-refugee violence and anti-refugee attitudes in Germany. The cause of the threat to security is incorrectly attributed to the victims of such violence by believing, for example, that the responsibility of hate crimes against refugees lies within the influx of refugees in Germany, with the consequence of the in-group (in this case native Germans) forming negative attitudes towards refugees. The impact of salient social norms on behaviours is strong immediately following message reception, but negligible in the longer term as the salience of the norm reduces (Cialdini and Goldstein, 2004). The instant ease of access, availability, and anonymity of online platforms like *Twitter* provide a convenient forum to enact the proper norm-congruent behaviour (e.g., to express hostility and hate) in the shorter term.

The Cronulla riot in Australia in 2005 is one example of the media further mobilizing hate. The riot, involving 5,000 white Australian vigilantes who inflicted violence against the Lebanese community (and other non-White Australians) to regain control of 'their' beach, showed similarities to a pogrom as it was deemed similar to 'a violent attack by members of a dominant ethnic group against a minority, in order to put them back in their place' (Poynting, 2006, p. 85). The riot occurred in the aftermath of a fight between surf lifesavers and a group of young Lebanese men, popular media outlets published 'populist racialisation, even incitement' (Poynting, 2006, p. 86). The media 'dutifully' reprinted a message circulated in a text messaging campaign inciting racial violence (Poynting, 2006, p. 86). Two-thirds of people calling into a radio station the day after expressed their support for the riots, which Poynting (2006, p. 88) suggests points to a more deep-seated problem of the state giving people 'permission to hate'.

Although the studies above provide important insights into connecting the experiences of online to offline hate and the likely impact of political figures and parties on offline hate, they do not fully attempt to explain the temporal association between such hate activities. In this paper, we empirically test the temporal association between online and offline hate,

exploring both directions. We specifically focus on the temporal association between Islamophobic online and offline hate by utilising time-series analysis. Time-series analysis allows for the analysis of data that have been collected over time at equally spaced time intervals and allows the drawing of associations between past events with future events. We test the following two alternative hypotheses:

H_1: Online anti-Islamic hate speech will be temporally predictive of anti-Islamic offline hate crime/incidents in the United Kingdom.

H_2: Offline anti-Islamic hate crime/incidents will be temporally predictive of anti-Islamic online hate speech in the United Kingdom.

We test the viability of the two alternative hypotheses above, using approximately one year's worth of recorded hate crime/incident data provided by the Metropolitan Police Service (London) and an equivalent period of hate speech expressed via *Twitter*.

3. Methods

3.1. Data and data collection

To test these hypotheses, we used the following two datasets: (1) daily hateful *Twitter* content and (2) weekly hate crimes/incidents recorded by the Metropolitan Police Service. The first dataset included daily online anti-Islamic *Twitter* content in the period from 24 February 2016 to 14 March 2017. These data were collected and filtered in multiple steps. First, a search term list including anti-Islamic slurs was collated through a literature and website search, including consultation with charities and experts in the field.[7] Second, a search was undertaken on *Twitter* to better understand the ways in which each of these terms were being used on the platform, and checked for areas of unexpected usage or ambiguity, with terms judged likely to produce

[7] We acknowledge that there are difficulties with automatically classifying online hate speech using a collection of slurs and that we may have missed hateful content; however, 'certain terms are particularly useful for distinguishing between hate speech and offensive language' (Davidson et al., 2017, p. 515) and our search term list includes many hateful slurs (see Appendix). Since the end of our project in 2017, data collection methods have also further evolved (see, e.g., Alorainy et al., 2019; Vidgen and Yasseri, 2020).

overwhelmingly irrelevant documents removed. Third, *Twitter*'s public REST APIs[8] were used to collect tweets containing one or more of the terms appearing in the list. Fourth, these tweets were processed by a series of natural language processing (NLP) classifiers and keyword-based filters to remove from the dataset tweets sent from outside the United Kingdom, non-English tweets, and tweets containing irrelevant phrases. Finally, a further series of NLP classifiers, described in full below, were trained to remove tweets that were not judged to express an explicitly anti-Islamic sentiment.

All collection and filtering were conducted using Method52, a text and analytics platform developed by the University of Sussex, the Centre for Analysis of Social Media and Demos. Method52 is based on DUALIST, a well-documented framework for linguistic analysis (Settles, 2011) and is the successor to Method51 (Wibberley et al., 2014; Wibberley et al., 2013). DUALIST is 'an active learning annotation paradigm which solicits and learns from labels on both features (e.g., words) and instances (e.g., documents)' (Settles, 2011, p. 1467). It allows for a more accurate, efficient and cost-effective annotation process and has been applied to language filtering and sentiment classification of tweets (Settles, 2011). Method51 extends the DUALIST framework and adds 'significant additional functionality including collaborative gold standard and classifier construction, processing pipeline construction, data collection and storage, data visualisation, various filtering and processing modules, and time-based data selection' (Wibberley et al., 2014, p. 115). Its successor Method52 incorporates this additional functionality with a number of further improvements, the most important being the development of a clear user interface, allowing non-technical users to build and test classifiers based on textual datasets, as well as making the platform non-*Twitter* specific. This made

[8] In particular, the 'Track API' (https://developer.twitter.com/en/docs/tweets/filter-realtime/overview/statuses-filter) was used to establish an ongoing, real-time collection of tweets matching a keyword, and the 'Standard search' API (https://developer.twitter.com/en/docs/tweets/search/overview/standard) was used to return tweets matching a keyword sent in the seven days prior to beginning the collection. Each of these APIs can be accessed free of charge.

Method52 the appropriate software to collect and analyze hateful Islamophobic content on *Twitter*.

The real time aspect of *Twitter* data allows for a temporal but also spatial 'temperature check' of hate locally (Lightowlers et al., 2018, p. 10). Although access to the internet is mostly not restricted to geographic location, and tweets sent from all over the world can be consumed locally in the UK, *Twitter* is a great medium for the detection of community tensions locally when such 'temperatures' run hot online (Lightowlers et al., 2018). Further, an estimated 500 million tweets are sent each day,[9] of which further research indicates the use of approximately 10,000 racial, religious and ethnic slurs used daily within such tweets (Bartlett et al., 2014). This means that there is a lot of noise that may wash out what happens locally within the UK, with research suggesting an estimated 393 derogatory and anti-Islamic tweets sent per day from within the UK (Miller and Smith, 2017), we therefore restricted our dataset to hateful tweets sent within the United Kingdom, which may also allow for online and offline intervention opportunities locally, based on what we find within our results.

The most common language used on Twitter is English, with an estimated 38.25% of all tweets using the English language (followed by Japanese with 11.84% and Spanish with 11.37 %; Leetaru et al., 2013). English language identification is possible via the preferred language setting chosen by *Twitter* users or via identification of tweets using the English language (Sloan et al., 2013). In this case, a classifier trained to identify English language within the body of tweets was used to remove non-English content from the dataset. In addition, geographical user information can be collected by *Twitter* users indicating (1) a location on their *Twitter* profiles (which may include incorrect or made-up information), (2) via geo-tagged tweets (only rarely used due to privacy and safety concerns), and (3) via tweet content (Sloan et al., 2013). We took a combination of the second and third approaches, using Method52 to

[9] More *Twitter* statistics can be found here: https://www.internetlivestats.com/twitter-statistics/.

determine the country a user was likely to be tweeting from, based on various metadata fields relevant to that user included with their tweet – i.e., the contents of free text 'user description' and 'placename' fields, their time zone, and included latitude and longitude coordinates. This filtering was carried out in order to help identify content likely to be relevant to the UK, rather than to a global discussion dominated by America.

Tweets classified as being sent from the UK and written in English were first passed through an initial relevancy classifier. In order to determine which English tweets sent from the UK were likely to express an anti-Islamic sentiment, we used Method52 to train a number of NLP classifiers to remove irrelevant tweets from the dataset. We then trained Method52 to remove tweets which were discussing prominently irrelevant and easily identified themes in the dataset; for example, using a term in a context unrelated to discussion of Islam or Muslims. This first filter took a broad view of relevance, and was designed to remove clearly irrelevant data – for example, tweets discussing Pakistan's cricket team. Remaining tweets were then classified as follows: All tweets containing the string 'p*ki' were passed through a classifier trained specifically to remove uses of this term not judged to be hateful.[10] Similarly, all tweets containing the string 'terroris' were passed through two classifiers.[11] The first of these was trained to remove mentions of terrorism not connected to Islam, with tweets which were connected passing through a second classifier, trained to remove tweets which were either defending Islam or Muslims (e.g., with reference to reports of offline terror attacks) or which did not express an opinion.

These two terms, 'terroris' and 'p*ki', both of which were included in the initial list of potential slur terms used to collect data from *Twitter*, were classified separately as they were both prevalent within the dataset and highly contested, each being used in a wide variety of

[10] Uses of the term 'p*ki' which were not deemed to be hateful included uses of the term as shorthand for Pakistan, particularly with reference to cricket matches, or as an abbreviation of 'Pakistani' where no other hateful messaging was included. We also termed non-hateful some colloquial uses, which could be argued to be culturally problematic, such as the use of 'p*ki' as a term for a corner shop.

[11] This string allowed us to capture different terms, such as terrorism, terrorist and terrorists.

contexts across the collection period. Tweets which used other slur terms in the collection, including 'muzrat', 'muzzie' and 'rag-head' were found, in general, to be less contested – it was easier in these cases to make a decision about whether they were likely to be hateful.[12] Tweets which contained neither one of these remaining terms – i.e., a collection term which was neither 'p*ki' or 'terroris', were therefore used to train a third classifier designed to remove non-hateful uses of any of these remaining terms used in the collection. In this way, the final classifier for each stream ('p*ki', 'terroris' and all other terms) was built to identify tweets which were judged to be using terms in a hateful sense, and any tweets identified as hateful by any of these three classifiers was thereby classified as expressing an anti-Islamic sentiment.

The second dataset was provided by the Metropolitan Police Service (MPS), who we partnered with on the project. This dataset recorded weekly offline Islamophobic incidents and offences from the first week in January 2014 and ending on 5 March 2017. The two datasets were combined to produce a time span of 53 weeks between the week ending 28 February 2016 (2016w9) and the week ending 05 March 2017 (2017w9). During this time period, the MPS recorded a total of 1,239 Islamophobic incidents and 1,246 Islamophobic offences, while the *Twitter* dataset included 159,309 anti-Islamic tweets. We used the statistical programme STATA to analyse the data, using time-series analysis. There was one missing observation date of 17 January 2017 from the *Twitter* dataset, consequently resulting in a slightly lower tweet count than expected in 2017w3. This study has been approved by the Sciences & Technology Cross-Schools Research Ethics Committee at the University of Sussex (ER/DAVIDW/9).

3.2. Statistical analysis

Vector autoregressive (VAR) modelling — originally proposed by Sims (1980) — was used to investigate the time-series (VAR; Amisano and Giannini, 1997; Hamilton, 1994; Lütkepohl,

[12] A full list of collection terms can be found in the Appendix or via the following paper, published by Demos: https://www.demos.co.uk/wp-content/uploads/2017/04/Results-Methods-Paper-MOPAC-SUMMIT-Demos.pdf

2005). A VAR model is especially useful for investigating the temporal order of the effects between two or more time-series (Hamilton, 1994; Lütkepohl, 2005). In VAR modelling, a list of series is regressed each on its own previous values as well as lags of all the other series in the list. Therefore, using a VAR model, we can estimate the long-run relationship among online hate on offline hate (see Dugan and Chenoweth, 2020). Here we used a three-variable VAR modelling. All variables in the system were treated as endogenous. Therefore, the three variables were both determinant and outcome. To identify the number of lags to include in the VAR model, we used four lag length selection criteria: likelihood ratio test (LR test), Akaike's information criterion (AIC), Schwarz's Bayesian information criterion (SBIC), and the Hannan and Quinn information criterion (HQIC).

To check whether the VAR model was correctly specified, we performed several diagnostic tests (Hamilton, 1994; Lütkepohl, 2005) to determine whether the VAR model satisfies the stability condition, and whether the residuals were white noise, there was no residual autocorrelation and VAR disturbances were normally distributed. Finally, Wald lag-exclusion statistics on the VAR model were obtained to test the significance of the lags in each equation.

It is difficult to interpret the relationships in a VAR model by inspecting the estimated parameters. Instead, to summarize the complex dynamics of the variables represented by the estimated parameters, we used Granger causality test and impulse response function (IRF) analysis. Granger causality tests are a well-established statistical method within the study of economics and has also found its way into criminological studies within recent decades (Carson et al., 2020). Granger causality tests allow for an examination of 'short-term causality between time trends among variables and to identify any reciprocal relationships' (Carson et al., 2020, p. 711). In the context of VAR modelling, Granger causality is based on the idea that determinant and outcome can be distinguished by temporal ordering. According to Carson et al. (2020, p. 711), 'to be considered "Granger causal," trend *X* at time *t* needs to contain

information that helps forecast trend *Y* at time *t+1*'. Orthogonalized impulse response functions (OIRFs) allow tracing the response of a variable to a shock to another variable, while taking into account the simultaneous correlations. In short, Granger causality tests can assess the direction of the relationship, while OIRFs explore how hate offline responds to the impulses of hate online and vice versa (see Dugan and Chenoweth, 2020). We will use these statistical tests to explore the temporal link between online and offline hate.

3.3. Limitations

A few limitations in this study are in need of acknowledgment. First, the overlapping dataset spans only a small timeframe (N=53) and although we can draw temporal correlations between online and offline hate, we do not claim causality and only some limited inference is possible. Second, Method52 has some technical limitations. Machine learning classifiers are inherently probabilistic, and the classifiers used in this study were not 100% accurate, and we can expect both to have mislabeled data as anti-Islamic that was not, and to have removed tweets from the dataset that were in fact anti-Islamic. To account for this, each of the five classifiers trained for this study was tested for accuracy on both their recall and precision, against a 'gold standard' dataset labelled by a human coder, with overall accuracies for labels relevant to hate lying between 69.3% and 85.8% (Miller et al., 2016). Furthermore, the classifiers used are strongly influenced by human judgement, and the subjectivity of the coders, both in training the classifier by labelling tweets, and building the 'gold standard' from which a measure of accuracy is derived. While coders worked together throughout the project in an attempt to define clear and consistent definitions to be used in building each classifier, inconsistencies in coder judgements about what in fact constitutes a hateful tweet are likely to affect the accuracies of these algorithms. Third, we also need to acknowledge the possibilities of *Tweeters* using Virtual Private Network (VPNs) when tweeting content, which may mask the true geographic location information of our collected tweets. In order to identify data likely to be relevant to the UK, we

use self-reported information – either in the form of geographic latitude/longitude data where this was included by a user, or through classification of user-provided, optional free-text fields which allow people to specify their location. This is clearly vulnerable to manipulation by users misrepresenting their location, and reflects a common challenge posed by the often anonymous, global nature of a platform like *Twitter*. Data from the beginning of 2018 estimates that around 26% of internet users globally, and 18% specifically within Europe indicated accessing the internet through VPNs or proxy servers within the previous month.[13] Recent data suggest that VPN usage has further increased since the global pandemic, with some data estimating that 41% of users in the United Kingdom (and the United States) use such services to either access restricted content or use VPNs for privacy concerns at least once a week, and 36% every or nearly every day.[14] This increase in VPN usage over the past years, due to smart phones and the recent surge in working from home, will likely need further consideration in future research. Fourth, the MPS data are likely to be a gross underestimate of the true extent of offline hate crime, with 105,090 hate crimes recorded by the police in 2019/2020, but an estimated 190,000 hate crimes happening in England and Wales every year (Home Office, 2020). We are also comparing UK national *Twitter* content with London-based offline hate; however, religious hate crimes cluster in London (Home Office, 2019), of which the majority are Islamophobic (Walters and Krasodomski-Jones, 2018), and due to national and social media, it is likely that readers elsewhere in the UK notice events that happen within London (and vice versa). Further, hate crimes may be recorded as such days or weeks after an incident has occurred, therefore, we need to be cautious with our interpretation of the findings around hate crimes in our dataset.

Although we would have liked to run additional statistical tests, there is only so much that can be done with our dataset. For example, we would have liked to add an additional analysis to test if hate has a longer lasting effect online than offline after trigger events within

[13] See https://www.statista.com/statistics/306955/vpn-proxy-server-use-worldwide-by-region/.
[14] See https://www.statista.com/statistics/1219770/virtual-private-network-use-frequency-us-uk/.

our dataset; however, prior research indicates that the effect online and offline will have dissipated days after the trigger event, and we have access only to weekly datasets.[15] An additional point of interest would have been to control for other possible factors that are likely to have an impact on hate crime at the local level (i.e., educational attainment, ethnic composition, or unemployment); however, we did not have the statistical power to add additional control variables. A dataset across multiple years would have also allowed us to check for seasonality and its effect on our data; however, the project only ran for 18 months.

4. Results

We first explored the data visually to identify any trends in the data. Figure 1 displays the anti-Islamic incidents and offences, as well as the anti-Islamic tweets. To be able to display the tweets, which were in thousands, in the same figure as the incidents and offences, we divided the *Twitter* data by 100. Figure 1 shows some observable correspondence between the online and offline data. Especially, Figure 1 displays a noticeable increase in the online *Twitter* data (long-dashed line) around spikes in the offline hate data (solid line and dashed line).

*** Insert Figure 1 about here ***

4.1. Estimation and Evaluation of the VAR Model

Next, a VAR model was estimated, evaluated (using diagnostic test), and interpreted (using Granger causality tests and IRF analysis). Concerning the identification of the number of lags, the likelihood-ratio tests selected a model with three lags. AIC, HQIC, and SBIC have selected a model with zero lags. We decided to use three lags because the selection of zero lag means that there is no contribution of past value of the exogenous variable, and we are not able to

[15] Due to restrictions placed by *Twitter* on the inaccessibility of deleted or removed tweets, some of the content identified as hateful by Method52 may since have been deleted or removed from the platform, and is no longer accessible to researchers. However, a Demos report by Miller et al. (2016) includes some more information and exploration of the data. Previous research also exists, exploring Islamophobic tweets (see, e.g., Awan, 2016c) and anti-Muslim hate on Facebook (see, e.g., Oboler, 2016).

obtain impulse response functions. Results (the parameters of the VAR model are available on request from the corresponding author) showed that in the anti-Islamic tweets equation, the two coefficients of anti-Islamic incidents and offences were statistically significant. Results from diagnostic tests on the VAR model revealed that (1) all the eigenvalues lay inside the unit circle; therefore, the VAR model satisfied the stability condition; (2) there was no autocorrelation in the residuals; thus, there was no hint of model misspecification; (3) disturbances in the VAR were normally distributed; (4) the three lags cannot be excluded. We concluded that the VAR model was correctly specified.

4.2. Granger Causality Tests

We performed Granger causality tests to determine whether anti-Islamic incidents and offences 'Granger caused' anti-Islamic tweets and whether the reverse was true, or whether all cross-lagged relationships were not significant. Table 1 shows the results of the Granger causality tests (based on small-sample *F* statistics). Anti-Islamic tweets did not 'Granger cause' anti-Islamic incidents (F =1.146; p = 0.342) or offences (F = 1.192; p = 0.352), concluding that online hate speech will not be temporally predictive of offline hate (H_1). Anti-Islamic incidents did not 'Granger cause' anti-Islamic offences (F = 2.459; p = 0.076) and also anti-Islamic offences did not 'Granger cause' anti-Islamic incidents (F = 1.667; p = 0.189). While anti-Islamic incidents (F = 4.422; p = 0.009) and offences (F = 3.984; p = 0.014) did 'Granger cause' anti-Islamic tweets, concluding that offline hate crimes/incidents will be temporally predictive of online hate speech (H_2). In short, offline hate crimes/incidents seem to predict online hate, while everyday online hate does not seem to predict offline hate. To determine the directionality of the impact of the observed effects, we should turn to the sign of the estimates in the VAR model. The influence of anti-Islamic incidents on anti-Islamic tweets was positive. The effect of anti-Islamic offences on anti-Islamic tweets was positive at the beginning, but then turned out to be negative in the later period.

*** Insert Table 1 about here ***

4.3. Impulse Response Analysis

To trace out the impact of a change in anti-Islamic incidents and offences on anti-Islamic tweets over time, we calculated OIRFs. Figure 2 displays the OIRFs with the 95 percent confidence bounds. The left panel of Figure 2 shows the response of anti-Islamic tweets to a shock in anti-Islamic incidents, whereas the right panel of Figure 2 displays the response of anti-Islamic tweets to a shock in anti-Islamic offences. The shock refers to the effect of a one-standard-deviation impulse to the anti-Islamic tweets equation. The magnitude of the shock corresponds to one unit standard deviation impulse to the anti-Islamic tweets equation. When their error bands do not include 0, responses are considered significant. We can see that both, Islamophobic incidents (left panel) and Islamophobic offences (right panel) seem to influence anti-Islamic tweets. In both OIRFs, the response is immediate and positive, with both incidents and offences being positively related to tweets immediately. While the response to Islamophobic incidents peaked at about 390 in the same week after a 1 standard deviation shock and then slowly tapered off to become 0, and after approximately two-three weeks the response to Islamophobic offences turned out to be negative for about one week and then disappeared. The effect is only significant in the same week as the impulse with a magnitude of about 100 in the same week after a 1 standard deviation shock.

*** Insert Figure 2 about here ***

5. Discussion

The study established that a temporal link between online and offline hate exists in the direction of association that is supportive of H_2 rather than H_1: anti-Islamic hate speech *followed* rather than preceded anti-Islamic hate offline. Given that recent other studies in this area have found that online hate speech is predictive of offline hate-based aggression (Müller and Schwarz,

2020; Williams et al., 2020), this finding may appear somewhat unexpected. However, our findings do not necessarily contradict the hypothesis that antilocution precedes physical acts of hate. Indeed, history shows that prolonged periods of hate speech targeted at a specific group, and which are spoken (typed) by the most powerful voices in society, will likely result in serious forms of violence. Our study, though, does not focus on such prolonged targeted speech, but instead examined the 'everyday' incidents of (online and offline) hate that affect a specific community throughout the United Kingdom.

Our finding that online hate follows offline hate points to other likely factors that may have an influence on this temporal relationship between online and offline hate. Most likely is what we have referred to as 'compound retaliation', which suggests that media and social media dissemination about offline acts of hate compound already tense intergroup hostilities, providing further permission for those to express hatred online. Such a situation represents the compounding of hate and hostility through offline and online networks that are likely to be reinforcing. As news media can play a role in 'instigat[ing] hate crime by formulating, propagating, and legitimating stereotypes about potential target populations' (Green et al., 2001, p. 486), it will also have a similar influence on online hate. News about hate crime offline may increase the salience of Islamophobic sentiment and social norms (Cialdini, Kallgren, and Reno, 1991). Online hostility and hate can be considered norm-congruent behaviours that can be easily acted on through online social networks like *Twitter*. We found that with an increase in hate incidents, hateful tweets increase and then gradually decrease to the same levels as before the increase in hate incidents, while an increase in hate offences displays a 'rebound' effect, which turned out to be negative after approximately two-three weeks and then disappeared. Therefore, our data display differences between offline hate incidents and hate crime/offences. There are, however, limitations with this part of the data, as hate crimes may be recorded as such days or weeks after an incident has occurred, these potential differences between hate incidents and hate crimes need further exploration and analysis. With much news

content distributed through social media platforms, such media representation and coverage of hateful crime events offline is likely to also have an influence on online communications (see, e.g., Zuleta and Burkal, 2017).

In particular, the narrative around real world hate incidents/crimes or what has been established as trigger events becomes important as these are likely to lead to 'compound retaliation', stirring up further hateful conversation online. This can happen, when perpetrators, for example, incorrectly attribute such hateful offline events to the victims, negatively shaping attitudes towards such victimized communities (see Igarashi, 2020), and in turn voicing such negative and hateful anti-Islamic rhetoric via *Twitter*. We see, for example, that when the Jewish community is victimized in a terror attack, there is a noticeable rise in Antisemitic hate, similar to when the Muslim community is victimized, a rise in Islamophobic hate. Such a situation was also especially noticeable with the recent Christchurch Mosques Shootings in New Zealand and the hateful response that followed by users of the online platform 8chan during and after the attack. The rhetoric behind and 'justifications' for hateful comments towards Muslims after such hate crimes and trigger events where Muslims are already targeted, however, needs to be explored further, especially how online perpetrators try to explain or 'justify' retaliating further against already victimized communities online.

A further interesting part of our finding is that the offline hate crime/incident data in London is associated with hateful *Twitter* activity throughout the United Kingdom as a whole (since the online hate is not restricted to – though we expect it to be concentrated in – London). This may suggest that national media coverage of London events is influencing people's social media use elsewhere, or that MPS hate crime incidence is similar to hate crime incidence in other UK localities. A study conducted by Kwon et al. (2019, p. 2652) on global *Twitter* conversations about the 2017 Quebec Mosque Shooting revealed 'that proximity influences global conversations related to hate crime news', and Eckert et al. (2021) further suggest that

what happens in the news locally quickly accelerates to the national level. Therefore, such compounding of hate has far-reaching consequences.

The fact that offline and online hate have a temporal order means that both phenomena should not be viewed or addressed in isolation. Muslim communities experience Islamophobia in both realms, but encountering online Islamophobia is experienced as more relentless, leading to high levels of stress for these communities (Eckert et al., 2018). As has been highlighted by Awan and Zempi (2017), both online and offline anti-Muslim crime have detrimental effects for victims, yet online victims are less 'visible' in the criminal justice system (Awan and Zempi, 2015, p. 4). Addressing this issue is likely to be a difficult task. The number of online hate incidents is likely to dwarf those offline, with our collected hateful online tweets being in the thousands compared with recorded weekly hate incidents/crimes in the tens. However, as we have seen in the beginning of this paper, only 2% of hateful content online makes it into police statistics. Further studies are therefore needed to explore the reasons for such under-recording of online hate by police. Reasons that may be linked to, for example, challenges identifying hate speech (both by police and victims) and tracking down perpetrators online, or likely due to a lack of resources and tools to focus in on the vast amount of hate online. The perpetrators of our collected online data and the offline data provided by the police are also not directly linked, which is an area of research that needs further exploration, with qualitative methods able to provide further insights into how hate perpetrators are influenced by online and offline material.

Further, our study employs VAR models, Granger causality and orthogonalized impulse response functions to explore the temporal link between online and offline hate. The novelty of this study and its methodology is in the examination of the temporal relationship between online and offline Islamophobia. Although these statistical tests have also recently been used to analyze the temporal link between positive vs. negative government attention and hateful violence (Dugan and Chenoweth, 2020) and to analyze the link between government action and

radical eco-movement incidents (Carson, Dugan and Yang, 2020), our study is currently the first to use such tests to explore time-series related to online and offline hate. In terms of analytical methods, the sophisticated models used in the present investigation did not allow us to establish a genuine "causal" relationship between online and offline hate. Indeed, similar to Carson et al.'s (2020) conclusion, we also suggest that true experimental designs are necessary to isolate such 'causes'.

5.1. Alternative explanations and further research

There are though other possible explanations that may have an impact on our findings, which have been beyond the ambit of this study. These include: (1) the possibility of a more organized network of hateful actors with a shorter attention span offline compared to online, (2) that different online environments may encourage different types of behavior, and (3) that differences in effect may depend on the target of online and offline hate. Further longitudinal research studies are required in order to test these possibilities.

Our findings of an anti-Islamic backlash online after Islamophobic hate on the street may also suggest that the internet is very organized. Although mission offenders, who may be part of an organized hate group seem to be rare in the offline world (see, e.g., McDevitt et al., 2002), they may be more frequently behind hateful online content (Williams and Burnap, 2016), with research indicating that almost 70% of online hate incidents are linked to the far-right (Copsey et al., 2013). Many hate groups prefer the use of the internet because such hateful communications often fall outside the law and are often defended by freedom of speech concerns (Eichhorn, 2001). Previous research on far-right groups suggests that these groups are using online platforms to gain support and use social media for cyberhate attacks on Muslim communities (Bartlett et al., 2011; Feldman and Littler, 2014). Awan's (2014) typology of online offenders targeting Muslims on social media also suggests the more organized and

calculated nature of online hate. Extremist and incendiary undertones are also often found on social networking sites, such as *Twitter* (Awan, 2016c).

We also need to consider our findings in light of *Twitter* as only one of many online platforms, with research suggesting that different online environments (e.g., Facebook, *Twitter*, Instagram, Quora, Reddit or 4Chan) encourage different types of behaviours, with websites that encourage anonymity (like Ask.fm) enabling more egregious forms of conduct (Binns, 2013, 2014). Although it is possible to create a fake profile on *Twitter*, people are often not anonymous. It is possible that the temporal association between offline and online hate occurrences is linked to other social media websites' ease of access, availability, and anonymity, compared with *Twitter's* online structure and rules of use. Online communities have previously mobilised offline action (Blakemore, 2016), as seen just recently with the attacks on two Mosques in New Zealand, the Poway synagogue shooting in California, the El Paso shooting in Texas and their direct link to the online platform 8chan (Wong, 2019). This type of online activity relates to just one of four harms of online hate: 'harm caused to society by the radicalisation of others' (Bakalis, 2018, p. 110). As not every hate offender will act on what they have been exposed to online in the offline world, the three remaining harms of online hate remain important: 'harm to individuals in a private forum, harm to individuals in a public forum, [and] harm to vulnerable groups' (Bakalis, 2018, p. 110). These harms still exist, even if our results suggest that hateful content on *Twitter* is not temporarily predictive of offline hate.

We also need to consider that differences in effect may depend on the target of online and offline hate. As Allen (2017) points out, anti-Muslim attitudes have become increasingly unquestioned and persist as 'white noise' in the public and political discourse. Bakalis (2018, p. 103) suggests 'that the hate perpetrated online is not equal, and some groups are suffering more than others', with Muslims enduring most of the hate after trigger events (e.g., related to terrorism). Differences may therefore exist in the temporal link between online and offline hate

for other targets of hate crime/incidents and hate speech. Research into other web-based platforms and different victim groups of online and offline hate is therefore necessary.

As we highlight above, our data stem from the day-to-day instances of hate online and offline, it focuses on anti-Islamic/Islamophobic sentiment, and is therefore different from prior analyses. Our data also spanned over a very important UK specific trigger event, the EU referendum, with a report by the House of Commons (2018) finding evidence for political manipulation and the specific targeting of online platform users, which had an effect on the EU referendum results and is a very new development. We also know that although still flawed, police hate crime recording practices have improved over the past years (see Home Office, 2018). In addition, our study uses current data, which allowed us to identify Islamophobic incidents.

In the current times of *fake news*, *alternative facts*, and the *Cambridge Analytica data scandal*, irresponsible sensationalism within the media, political figures and parties spreading intolerance and hatred, and rogue online platforms will continue to fan the flames of anti-Muslim hatred, and little will change until greater care is taken to ensure responsibility by journalists, politicians, platforms and individuals themselves. A recent development within the Metropolitan Police Service has been the creation of an Online Hate Crime Hub in London, which has responsibility for assisting in the investigation, and monitoring of reported online hate. Its programme includes a dedicated, trained police team that filters, identifies and responds to hate online (MOPAC, 2016). This new approach to tackle online hate sets out to not only strengthen the relationship between the police, impacted communities and social media providers, but also to hold online perpetrators accountable for spreading hateful messages online and providing targeted and effective victim services. It is hoped this new approach will increase public confidence in the police and the reporting of hate crime and incidents online and offline (MOPAC, 2016). The effectiveness of such an initiative is yet to be evaluated.

6. Conclusion

In this paper, we established that a temporal association exists between online and offline hate, concluding that online hate follows rather than leads offline hate. Our analyses suggest that hate offline does not simply end with the commission of one hateful physical act, but that it is likely to be part of an ongoing process of hate that will affect multiple people (Bowling 1993, 1998). Our research gives credence to the well-founded assertion that hate crimes/incidents on the street are message crimes (Perry and Alvi, 2012). We found here that this message is spread further by online perpetrators, compounding such hate and hostility further online. The establishment of global social media platforms means that what happens locally offline is likely to have far-reaching consequences, reaching into the online world and to an audience of potentially hundreds of millions of people.

The 2016 UK Government Home Office Hate Crime Action Plan makes specific reference to the exploitation of the internet as a tool to spread hatred. The police are now specifically instructed to apply online flags to hate crime with an online element (Home Office, 2016). According to Awan (2016b, p. 183), 'research which can help examine the link between online abuse with actual offline physical violence would be poignant and critical for government' and can assist agencies in better understanding what they are dealing with. Therefore, the results in this paper have important implications for the policing of hate crime. When spikes in Islamophobic crimes/incidents happen on the street, police can expect 'compound retaliation' whereby an anti-Islamic backlash occurs on *Twitter*. Having a greater understanding of this retaliation cycle should help police forces to allocate time and resources into the policing of hate, including for example, the operation of the Online Hate Crime Hub.[16]

Twitter also recently announced an update to its rules against hateful conduct, 'includ[ing] language that dehumanizes others on the basis of religion' (Twitter Safety, 2019),

[16] More information can be found here: https://www.gov.uk/government/news/home-secretary-announces-new-national-online-hate-crime-hub.

taking on some of the responsibility for hateful content placed on their platform. Furthermore, we believe that tools which enable machine learning algorithms to be trained on particular datasets by non-technical subject matter experts, as Method52 does here, represent a vital opportunity to inexpensively conduct the difficult work of detecting hateful content within the vast, noisy and very human datasets created by social media.

Table 1 Granger Causality Tests

Causality test	***F***	***df***	***p***	**VAR model estimates**		
				Lag 1	Lag 2	Lag 3
Anti-Islamic offences → anti-Islamic incidents	1.667	3	.189	-0.51	-0.48	-0.62
Anti-Islamic tweets → anti-Islamic incidents	1.146	3	.342	0.00	0.00	0.00
Anti-Islamic incidents → anti-Islamic offences	2.459	3	.076	0.64	0.63	0.59
Anti-Islamic tweets → anti-Islamic offences	1.192	3	.352	0.00	0.00	-0.00
Anti-Islamic offences → anti-Islamic tweets	3.984	3	.014	11.40	-35.60	-192.83
Anti-Islamic incidents → anti-Islamic tweets	4.422	3	.009	31.13	26.91	210.24

Note. A significant *F* value implies that the determinant (variable on the left of the arrow) 'Granger causes' the outcome (variable on the right of the arrow). The direction of the influence can be derived from the sign of the estimates in the VAR model (rightmost column).

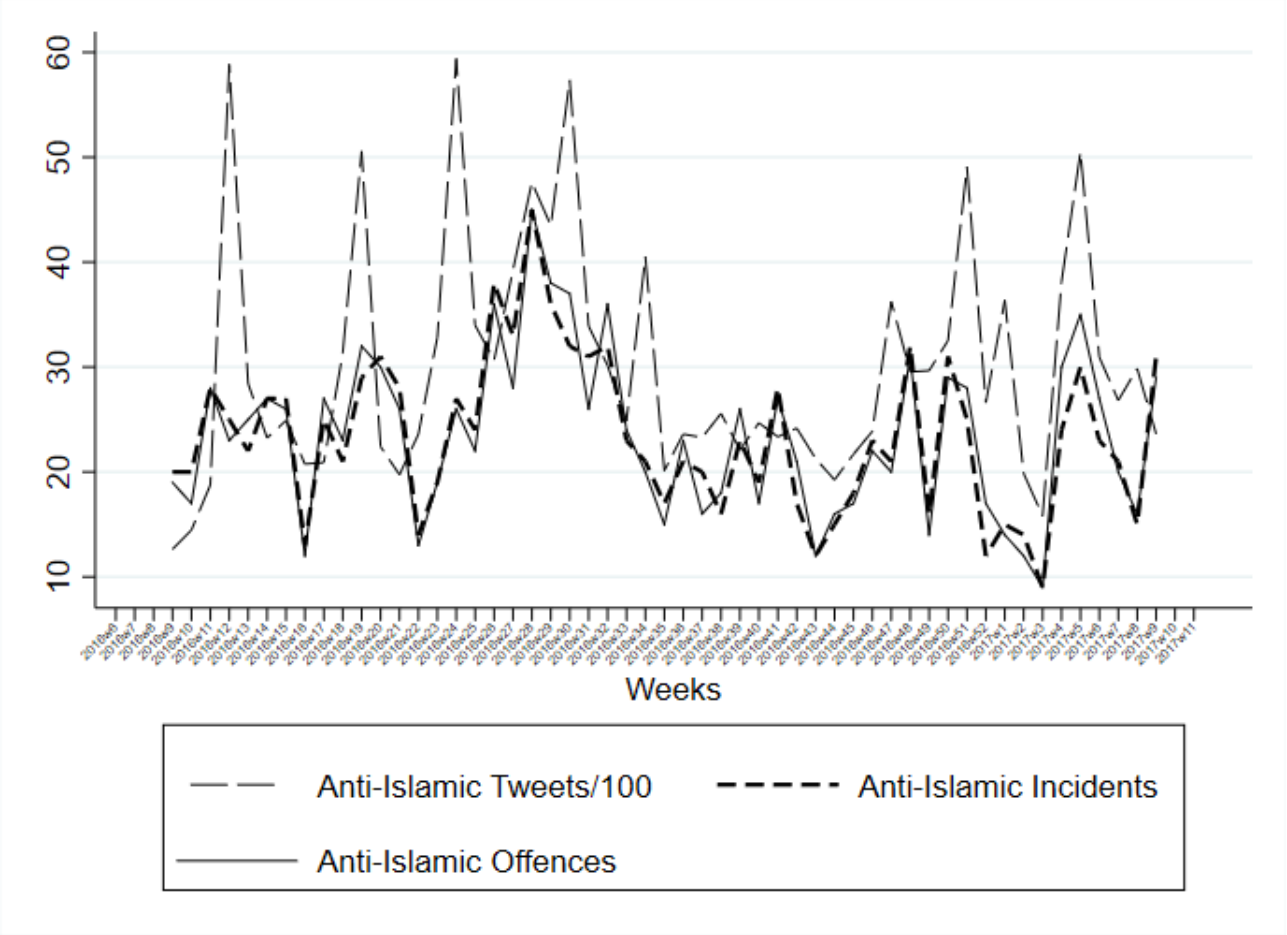

Fig. 1. Anti-Islamic incidents, anti-Islamic offences, and tweets/100 by weeks

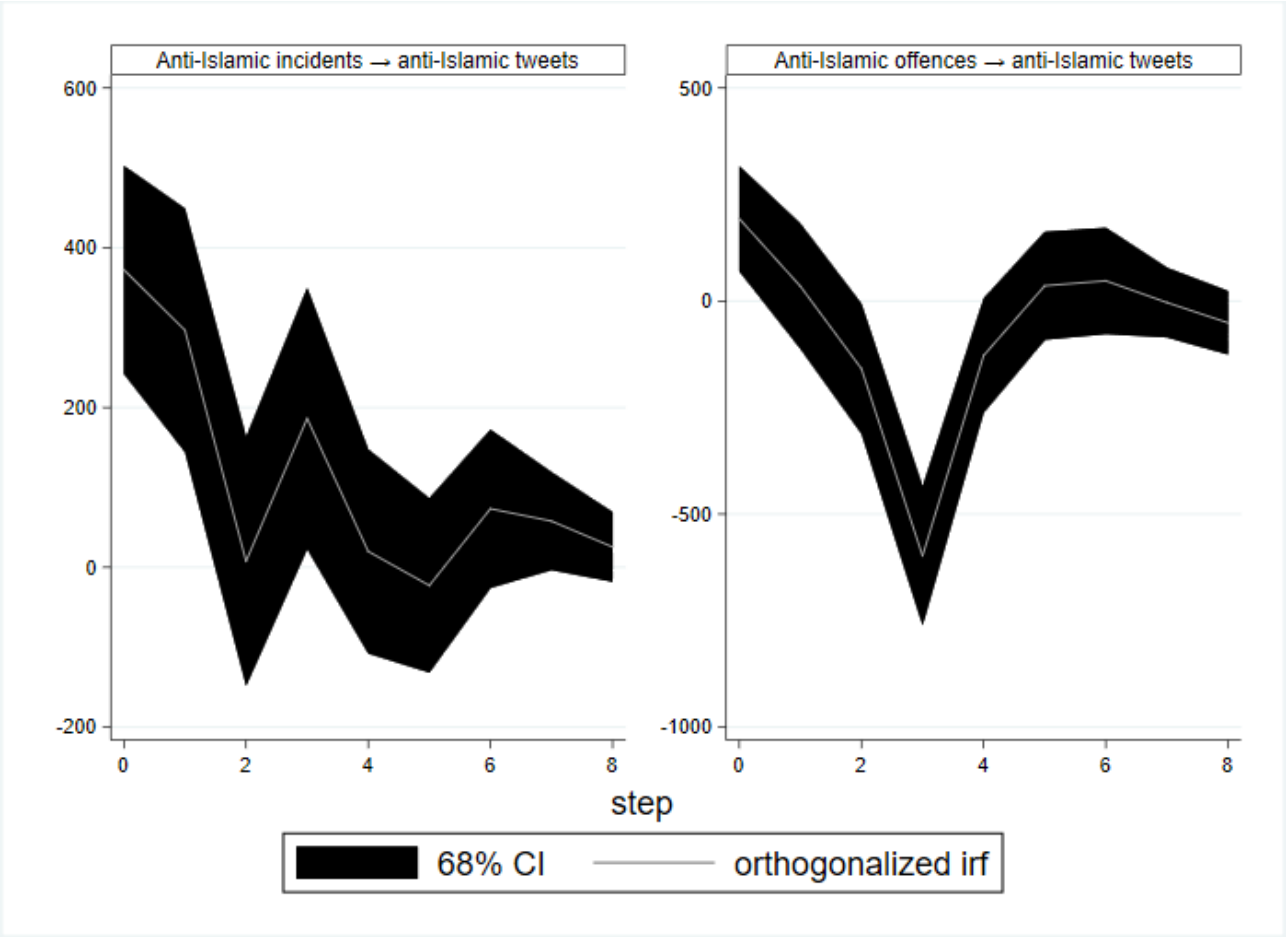

Fig. 2. Orthogonalized impulse response functions. Left panel: impulse = anti-Islamic incidents; right panel: impulse = anti-Islamic offences; *CI* = confidence interval

7. Appendix: List of *Twitter* data collection keywords

BNP
camel fucker
carpet pilot
clitless
derka derka
diaper head, diaper-head
dune coon
dune nigger
durka-durka
EDL
fuckmuslims
hijab
jig-abdul
jihad, jihadi
kaffir
kuffar
mudshark
muslim paedos
muslim pigs
muslim scum
muslim terrorists
muzrats
muzzie, muzzies
paki, pakis
Pegida
pisslam
q-tip head
rab
racoon
raghead, rag-head, ragheads
rapefugee, rapeugee
rug pilot
rug-rider
sand flea
sand monkey
sand moolie
sand nigger
sand rat
slurpee nigger
terrorist
towel head, towel-head
whitegenocide

8. References

Alam ,Y., and Husband, C. (2013). Islamophobia, community cohesion and counter-terrorism policies in Britain. Patterns of Prejudice 47: 235--252.

Allen, C. (2017). Britain must address the pervasive 'white noise' against Muslims. The Conversation. https://theconversation.com/britain-must-address-the-pervasive-white-noise-against-muslims-79770

Allport, G.W. (1954). The nature of prejudice, Addison-Wesley. Reading.

Alorainy, W., Burnap, P., Liu, H., and Williams, M.L. (2019). "The enemy among us": Detecting cyber hate speech with threats-based othering language embeddings. ACM Transactions on the Web 13(3, Article 14): 1--26.

Alsaad, A., Taamneh, A., and Al-Jedaiah, M.N. (2018). Does social media increase racist behavior? An examination of confirmation bias theory. Technology in Society 55: 41--46.

Amisano, G., and Giannini, C. (1997). Topics in structural VAR econometrics, Springer-Verlag, Heidelberg.

Anderson, C.A., and Bushman, B.J. (2001). Effects of violent video games on aggressive behavior, aggressive cognition, aggressive affect, physiological arousal, and prosocial behavior: A meta-analytic review of the scientific literature. Psychological Science 12: 353--359.

Awan, I. (2014). Islamophobia and Twitter: A typology of online hate against Muslims on social media. Policy & Internet 6: 133--150.

Awan, I. (2016a). Cyber-Islamophobia and internet hate crime. In Awan, I. (ed.), Islamophobia in cyberspace - Hate crimes go viral, 1st ed., Ashgate Publishing, Oxon, New York, pp. 7--22.

Awan, I. (2016b). Islamophobia, hate crime and the internet. In Awan, I. (ed.), Islamophobia in cyberspace - Hate crimes go viral, 1st ed, Ashgate Publishing, Oxon, New York, pp. 167--187.

Awan, I. (2016c). Virtual Islamophobia: The eight faces of anti-Muslim trolls on Twitter. In Awan, I. (ed.), Islamophobia in cyberspace - Hate crimes go viral, 1st ed., Ashgate Publishing, Oxon, New York, pp. 23--39.

Awan, I., and Zempi, I. (2015). We fear for our lives: Offline and online experiences of anti-Muslim hostility. TellMama, London. https://www.tellmamauk.org/wp-content/uploads/resources/We%20Fear%20For%20Our%20Lives.pdf

Awan, I., and Zempi, I. (2016). The affinity between online and offline anti-Muslim hate crime: Dynamics and impacts. Aggression and Violent Behavior 27: 1--8.

Awan, I., and Zempi, I. (2017). I will blow your face off' – Virtual and physical world anti-Muslim hate crime. The British Journal of Criminology 57: 362--380.

Bakalis, C. (2018). Rethinking cyberhate laws. Information & Communications Technology Law 27(1): 86--110.

Bartlett, J., Birdwell, J., and Littler, M. (2011). The new face of digital populism. London: Demos. https://demosuk.wpengine.com/files/Demos_OSIPOP_Book-web_03.pdf?1320601634

Bartlett, J., Reffin, J., Rumball, N., and Williamson, S. (2014). Anti-social media. Demos, London. https://www.demos.co.uk/files/DEMOS_Anti-social_Media.pdf

Benesch, S. (2013). Dangerous speech: A proposal to prevent group violence. http://dangerousspeech.org/guidelines/

Binns, A. (2013). Facebook's ugly sisters: Anonymity and abuse on Formspring and Ask.fm. Media Education Research Journal.

Binns, A. (2014). Twitter City and Facebook Village: Teenage girls' personas and experiences influenced by choice architecture in social networking sites. Journal of Media Practice 15: 71--91.

Blakemore, B. (2016). Online hate and political activist groups. In Awan, I. (ed.), Islamophobia in cyberspace - Hate crimes go viral, 1st ed., Ashgate Publishing, Oxon, New York, pp. 63--83.

Bowling, B. (1993). Racial harassment and the process of victimization: Conceptual and methodological implications for the local crime survey. The British Journal of Criminology 33: 231--250.

Bowling, B. (1998). Violent racism: Victimisation, policing, and social context, Ocford University Press, New York.

Burke, J. (2019). Norway mosque attack suspect 'inspired by Christchurch and El Paso shootings'. The Guardian. https://www.theguardian.com/world/2019/aug/11/norway-mosque-attack-suspect-may-have-been-inspired-by-christchurch-and-el-paso-shootings

Carson, J.V., Dugan, L., and Yang, SM. (2020). A comprehensive application of rational choice theory: How costs imposed by, and benefits derived from, the U.S. Federal Government affect incidents perpetrated by the Radical Eco-Movement. Journal of Quantitative Criminoligy 36**:** 701--724.

Chan, J., Ghose, A., and Seamans, R. (2016). The internet and racial hate crime: Offline spillovers from online access. MIS Quarterly 40: 381--404.

Cialdini, R.B., and Goldstein, N.J. (2004). Social influence: Compliance and conformity. Annual Review of Psychology 55: 591--621.

Cialdini, R.B., Kallgren, C.A., and Reno, R.R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In

Berkowitz, L. (ed.), Advances in experimental social psychology, Academic Press, San Diego, 201--234.

Cialdini, R.B., Reno, R.R., and Kallgren, C.A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. Journal of Personality and Social Psychology 58: 1015--1026.

Ciftci, S. (2012). Islamophobia and threat perceptions: Explaining anti-Muslim sentiment in the West. Journal of Muslim Minority Affairs 32: 293--309.

College of Policing (2020). Responding to hate. https://www.app.college.police.uk/app-content/major-investigation-and-public-protection/hate-crime/responding-to-hate/#agreed-definitions

Copsey, N., Dack, J., Littler, M., et al. (2013). Anti-Muslim hate crime and the far right. Teesside University Centre for Fascist, Anti-Fascist and Post-Fascist Studies. https://research.tees.ac.uk/en/publications/anti-muslim-hate-crime-and-the-far-right

Cuerden, G., and Rogers, C. (2017). Exploring race hate crime reporting in Wales following Brexit. Review of European Studies 9: 158--164.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. Proceedings of the Eleventh International AAAI Conference on Web and Social Media: 1--4.

Dodd, V. (2019). Anti-Muslim hate crimes soar in UK after Christchurch shootings. The Guardian. https://www.theguardian.com/society/2019/mar/22/anti-muslim-hate-crimes-soar-in-uk-after-christchurch-shootings

Dugan, L., and Chenoweth, E. (2020). Threat, emboldenment, or both? The effects of political power on violent hate crimes. Criminology 58: 714--746.

Eckert, S., O'Shay Wallace, S., Metzger-Riftkin, J., and Kolhoff, S. (2018). "The best damn representation of Islam:" Muslims, gender, social media and Islamophobia in the United States. CyberOrient 12(1): 4--30.

Eckert, S., Metzger-Riftkin, J., Kolhoff, S., and O'Shay-Wallace, S. (2021). A hyper differential counterpublic: Muslim social media users and Islamophobia during the 2016 US presidential election. New Media & Society 23(1): 78--98.

Eichhorn, K. (2001). Re-in/citing linguistic injuries: Speech acts, cyberhate, and the spatial and temporal character of networked environments. Computers and Composition 18: 293--304.

Federal Bureau of Investigation (2018). 2017 hate crime statistics. U.S. Department of Justice, Criminal Justice Information Services Division. https://ucr.fbi.gov/hate-crime/2017

Feldman, M., and Littler, M. (2014). TellMAMA reporting 2013/14: Anti-Muslim overview, analysis and 'cumulative extremism'. https://www.tellmamauk.org/wp-content/uploads/2014/07/finalreport.pdf

Green, D.P., McFalls, L.H., and Smith, J.K. (2001). Hate crime: An emergent research agenda. Annual Review of Sociology 27: 479--504.

Hamilton, J.D. (1994). Time series analysis, Princeton University Press, Princeton.

Hanes, E., and Machin, S. (2014). Hate crime in the wake of terror attacks. Journal of Contemporary Criminal Justice 30: 247--267.

Hawdon, J. (2012). Applying a differential association theory to online hate groups: A theoretical statement. Research on Finnish Society 5: 39--47.

Hewstone, M. (1990), The 'ultimate attribution error'? A review of the literature on intergroup causal attribution. European Journal of Social Psychology 20(4): 311--335.

Home Office (2016). Action against hate: The UK Government's plan for tackling hate crime. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/543679/Action_Against_Hate_-_UK_Government_s_Plan_to_Tackle_Hate_Crime_2016.pdf

Home Office (2018). Hate crime, England and Wales, 2017/18. Home Office, London. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/748598/hate-crime-1718-hosb2018.pdf

Home Office (2019). Hate crime, England and Wales, 2018/19 - Appendix tables (Table 1). Home Office, London. https://www.gov.uk/government/statistics/hate-crime-england-and-wales-2018-to-2019

Home Office (2020). Hate crime, England and Wales, 2019 to 2020. Home Office, London. https://www.gov.uk/government/statistics/hate-crime-england-and-wales-2019-to-2020/hate-crime-england-and-wales-2019-to-2020

House of Commons (2018). Disinformation and 'fake news': Interim report. https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/363/363.pdf

Igarashi, A. (2020). Hate begets hate: Anti-refugee violence increases anti-refugee attitudes in Germany. Ethnic and Racial Studies [online].

Irfan, A. (2021). Debating hatred: Islamophobia or anti-Muslim hate? Media Diversity Institute. https://www.media-diversity.org/debating-hatred-islamophobia-or-anti-muslim-hate/

Jaki, S., and De Smedt, T. (2019). Right-wing German hate speech on Twitter: Analysis and automatic detection. arXiv preprint. arXiv: 1910.07518.

Kaplan, J. (2006). Islamophobia in America?: September 11 and Islamophobic hate crime. Terrorism and Political Violence 18(1): 1--33.

Kwon, K.H., Chadha, M., and Wang, F. (2019). Proximity and networked news public: Structural topic modeling of global Twitter conversations about the 2017 Quebec Mosque Shooting. International Journal of Communication 13: 2652--2675.

Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., and Shook, E. (2013). Mapping the global Twitter heartbeat: The geography of Twitter. First Monday 18(5).

Levin, J., and McDevitt, J. (1993). Hate crimes: The rising tide of bigotry and bloodshed, Plenum, New York.

Lightowlers, C., Chenevoy, N., Malleson N., Beeley, S., Blair F., Keay, S., Bretherton, R., Stone, K., Eckersley, R., Chapman D., and Pascale F. (2018). Sharing insights on hate crime: New methods and forms of data. N8 Policing Research Partnership. https://www.liverpool.ac.uk/media/livacuk/law-and-social-justice/3research/Sharing,Insights,on,Hate,Crime.pdf

Lütkepohl, H. (2005). New introduction to multiple time series analysis, Springer-Verlag, Berlin.

Malik, N. (2019). Instead of Islamophobia, we should focus on defining anti-Muslim hatred. Forbes. https://www.forbes.com/sites/nikitamalik/2019/05/20/instead-of-islamophobia-we-should-focus-on-defining-anti-muslim-hatred/?sh=5f12b0ee69e5

Maynard, J.L., and Benesch, S. (2016). Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention. Genocide Studies and Prevention: An International Journal 9: 70--95.

McDevitt, J., Levin, J., and Bennett, S. (2002). Hate crime offenders: An expanded typology. Journal of Social Issues 58: 303--317.

Miller, C. (2016). Measuring Islamophobia on Twitter. Demos, London. https://demos.co.uk/blog/measuring-islamophobia-on-twitter/

Miller. C., and Smith, J. (2017). Anti-Islamic content on Twitter. Demos, London. https://demos.co.uk/project/anti-islamic-content-on-twitter/

Miller, C., Arcostanzo, F., Smith, J., et al. (2016) From Brussels to Brexit: Islamophobia, xenophobia, racism and reports of hateful incidents on Twitter. Demos, London. http://www.demos.co.uk/wp-content/uploads/2016/07/From-Brussels-to-Brexit_-Islamophobia-Xenophobia-Racism-and-Reports-of-Hateful-Incidents-on-Twitter-

Research-Prepared-for-Channel-4-Dispatches-%E2%80%98Racist-Britain%E2%80%99-.pdf

MOPAC (2016). Home Office Police Innovation Fund - Online Hate Crime Hub. https://www.london.gov.uk/sites/default/files/pcd_41_home_office_police_innovation_fund_-_online_hate_crime_hub_0.pdf

Müller, K., and Schwarz, C. (2020). Fanning the flames of hate: Social media and hate crime. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3082972

O'Neill, A. (2017). Hate crime, England and Wales, 2016 to 2017. Home Office, London. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/652136/hate-crime-1617-hosb1717.pdf

Oboler, A. (2016). The normalisation of Islamophobia through social media: Facebook. In Awan, I. (ed.), Islamophobia in cyberspace - Hate crimes go viral, Ashgate Publishing, Oxon, New York, pp. 41--61.

Oriola, O., and Kotzé, E. (2020). Evaluating machine learning techniques for detecting offensive and hate speech in South African tweets. IEEE Access 8: 21496--21509.

Paterson J., Walters M.A., and Brown R., et al. (2018). The Sussex Hate Crime Project - Final report. https://www.sussex.ac.uk/webteam/gateway/file.php?name=sussex-hate-crime-project-report.pdf&site=430

Perry, B., and Alvi, S. (2012). 'We are all vulnerable': The in terrorem effects of hate crimes. International Review of Victimology 18: 57--71.

Pettigrew, T. F. (1979). The ultimate attribution error: Extending Allport's cognitive analysis of prejudice. Personality and Social Psychology Bulletin 5(4): 461--476.

Piatkowska, S.J., and Lantz, B. (2021). Temporal clustering of hate crimes in the aftermath of the Brexit vote and terrorist attacks: A comparison of Scotland and England and Wales. The British Journal of Criminology 61(3): 648--669.

Poynting, S. (2006). What caused the Cronulla riot? Race & Class 48: 85--92.

Rahman, M. (2016). The media impact of online Islamophobia: An analysis of the Wollwich Murder. In Awan, I. (ed.), Islamophobia in Cyberspace: Hate Crimes Go Viral. Routledge, Oxon, New York, ch. 5.

Roy, E.A. (2019). 'It brings everything back': Christchurch despairs over white supremacist attacks. The Guardian. https://www.theguardian.com/world/2019/aug/14/it-brings-everything-back-christchurch-despairs-over-white-supremacist-attacks

Sadique, K., Tangen, J., and Perowne, A. (2018). The importance of narrative in responding to hate incidents following 'trigger' events. TellMAMA, London. https://tellmamauk.org/wp-content/uploads/resources/Tell%20MAMA%20-%20Report.pdf?utm_source=Report+Launch+Westminster+Bridge+09122018&utm_campaign=Westminster+Bridge+Report+09122018&utm_medium=email

Settles, B. (2011). Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Edinburgh, Scotland, 1467--1478.

Sims, C. (1980). Macroeconomics and reality. Econometrica 48: 1--48.

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., and Rana, O. (2013). Knowing the tweeters: Deriving sociologically relevant demographics from Twitter. Sociological Research Online, 18(3): 74--84.

Sutherland, E.H., and Cressey, D.R. (1974). Criminology,J. B. Lippincott, New York.

The Law Commission (2014). Hate crime: Should the current offences be extended? http://www.lawcom.gov.uk/app/uploads/2015/03/lc348_hate_crime.pdf

Twitter (2015). Fighting abuse to protect freedom of expression. https://blog.twitter.com/official/en_au/a/2015/fighting-abuse-to-protect-freedom-of-expression-au.html

Twitter Safety (2019). Updating our rules against hateful conduct. https://blog.twitter.com/en_in/topics/company/2019/updating-rules-against-hateful-conduct.html

Vidgen, B., and Yasseri, T. (2020). Detecting weak and strong Islamophobic hate speech on social media. Journal of Information Technology & Politics 17(1): 66--78.

Vidgen, B., Yasseri, T., and Margetts, H. (2019). Trajectories of Islamophobic hate amongst far right actors on Twitter. arXiv preprint. arXiv: 1910.05794.

Walters M.A., and Krasodomski-Jones, A. (2018). Patterns of hate crime: Who, what, when and where? Demos, London. https://www.demos.co.uk/wp-content/uploads/2018/08/PatternsOfHateCrimeReport-.pdf

Wibberley, S., Weir, D., and Reffin, J. (2014). Method51 for mining insight from social media datasets. COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations. Dublin, Ireland, 115--119.

Wibberley, S., Reffin, J., and Weir, D. (2013). Language technology for agile social media science. 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. Sofia, Bulgaria: Association for Computational Linguistics, 36--42.

Williams, M.L., and Burnap, P. (2016). Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data. The British Journal of Criminology 56: 211--238.

Williams, M.L., Burnap, P., Javed, A., et al. (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.

Wilson, R.A., and Land, M.K. (2021). Hate speech on social media: Content moderation in context. Connecticut Law Review 52(3): 1029--1076.

Wong, J.C. (2019). 8chan: The far-right website linked to the rise in hate crimes. The Guardian. https://www.theguardian.com/technology/2019/aug/04/mass-shootings-el-paso-texas-dayton-ohio-8chan-far-right-website

Zuleta, L., and Burkal, R. (2017). Hate speech in the public online debate.The Danish Institute for Human Rights, Copenhagen. https://www.humanrights.dk/sites/humanrights.dk/files/media/dokumenter/udgivelser/equal_treatment_2017/hate_speech_in_the_public_online_debate_eng_2017.pdf

Zunes, S. (2017). Europe's refugee crisis, terrorism, and Islamophobia. Peace Review 29: 1--6.