doi:10.1093/bjc/azz049 BRIT. J. CRIMINOL. (2020) **60**, 93–117
Advance Access publication 23 July 2019

# HATE IN THE MACHINE: ANTI-BLACK AND ANTI-MUSLIM SOCIAL MEDIA POSTS AS PREDICTORS OF OFFLINE RACIALLY AND RELIGIOUSLY AGGRAVATED CRIME

MATTHEW L. WILLIAMS*, PETE BURNAP, AMIR JAVED, HAN LIU and SEFA OZALP

*National governments now recognize online hate speech as a pernicious social problem. In the wake of political votes and terror attacks, hate incidents online and offline are known to peak in tandem. This article examines whether an association exists between both forms of hate, independent of 'trigger' events. Using Computational Criminology that draws on data science methods, we link police crime, census and Twitter data to establish a temporal and spatial association between online hate speech that targets race and religion, and offline racially and religiously aggravated crimes in London over an eight-month period. The findings renew our understanding of hate crime as a process, rather than as a discrete event, for the digital age.*

Keywords: hate speech, hate crime, social media, predictive policing, big data, far right

## *Introduction*

Hate crimes have risen up the hierarchy of individual and social harms, following the revelation of record high police figures and policy responses from national and devolved governments. The highest number of hate crimes in history was recorded by the police in England and Wales in 2017/18. The 94,098 hate offences represented a 17 per cent increase on the previous year and a 123 per cent increase on 2012/13. Although the Crime Survey for England and Wales has recorded a consistent decrease in *total* hate crime victimization (combining race, religion, sexual orientation, disability and transgender), estimations for race and religion-based hate crimes in isolation show an increase from a 112,000 annual average (April 13–March 15) to a 117,000 annual average (April 15–March 17) (ONS, 2017). This increase does not take into account the likely rise in hate victimization in the aftermath of the 2017 terror attacks in London and Manchester. Despite improvements in hate crime reporting and recording, the consensus is that a significant 'dark figure' remains. There continues a policy and practice need to improve the intelligence about hate crimes, and in particular to better understand the role community tensions and events play in patterns of perpetration. The HMICFRS (2018) inspection on police responses to hate crimes evidenced that forces remain largely ill-prepared to handle the dramatic increases in racially and religiously

*Matthew L. Williams, School of Social Sciences and HateLab, Cardiff University, King Edward VII Ave, Cardiff CF10 3WT, UK; WilliamsM7@cf.ac.uk; Pete Burnap, School of Computer Science and Informatics and HateLab, Cardiff University, Cardiff, UK; Amir Javed, School of Computer Science and Informatics, Cardiff University, Cardiff, UK; Han Liu, School of Computer Science and Informatics, Cardiff University, Cardiff, UK; Sefa Ozalp, School Social Sciences, Cardiff University, Cardiff, UK.

aggravated offences following events like the United Kingdom-European Union (UK-EU) referendum vote in 2016 and the terror attacks in 2017. Part of the issue is a significant reduction in Police Community Support Officers throughout England, and in particular London (Greig-Midlane (2014) indicates a circa 50 per cent reduction since 2010). Fewer officers in neighbourhoods gathering information and intelligence on community relations reduces the capacity of forces to pre-empt and mitigate spates of inter-group violence, harassment and criminal damage.

Technology has been heralded as part of the solution by transforming analogue police practices into a set of complementary digital processes that are scalable and deliverable in near real time (Williams *et al.*, 2013; Chan and Bennett Moses, 2017; Williams *et al.*, 2017a). In tandem with offline hate crime, online hate speech posted on social media has become a pernicious social problem (Williams *et al.*, 2019). Thirty years on from the Home Office (1989) publication '*The Response to Racial Attacks and Harassment*' that saw race hate on the streets become priority for six central Whitehall departments, the police, Crown Prosecution Service (CPS) and courts (Bowling, 1993), the government is now making similar moves to tackle online hate speech. The Home Secretary in 2016 established the National Online Hate Crime Hub, a Home Affairs Select Committee in 2017 established an inquiry into hate crime, including online victimization, and a review by the Law Commission was launched by the prime minister to address the inadequacies in legislation relating to online hate. Social media giants, such as Facebook and Twitter, have been questioned by national governments and the European Union over their policies that provided safe harbour to hate speech perpetrators. Previous research shows hate crimes offline and hate speech online are strongly correlated with events of significance, such as terror attacks, political votes and court cases (Hanes and Machin, 2014; Williams and Burnap, 2016). It is therefore acceptable to assume that online and offline hate in the immediate wake of such events are highly correlated. However, what is unclear is if a more general pattern of correlation can be found independent of 'trigger' events. To test this hypothesis, we collected Twitter and police recorded hate crime data over an eight-month period in London and built a series of statistical models to identify whether a significant association exists. At the time of writing, no published work has shown such an association. Our models establish a general temporal and spatial association between online hate speech targeting race and religion and offline racially and religiously aggravated crimes *independent of 'trigger' events*. Our results have the potential to renew our understanding of hate crime as a process, rather than a discrete event (Bowling, 1993), for the digital age.

### *Prevalence of Online Hate Speech on Social Media*

Since its inception, the Internet has facilitated the propagation of extreme narratives often manifesting as hate speech targeting minority groups (Williams, 2006; Perry and Olsson, 2009; Burnap and Williams, 2015, 2016; Williams and Burnap, 2016; Williams *et al.*, 2019). Home Office (2018) data show that 1,605 hate crimes were flagged as online offences between 2017 and 2018, representing 2 per cent of all hate offences. This represents a 40 per cent increase compared to the previous year. Online race hate crime makes up the majority of all online hate offences (52 per cent), followed by sexual orientation (20 per cent), disability (13 per cent), religion (12 per cent) and transgender online hate crime (4 per cent). Crown Prosecution Service data show that in the year April 2017/18, there were 435 prosecutions related to online hate, a 13 per cent increase on the previous year (CPS, 2018). These figures are a significant underestimate.[1] HMICFRS (2018) found that despite the Home Office introducing a requirement for police forces to flag cyber-enabled hate crime offences, uptake on this practice has been patchy and inconsistent, resulting in unreliable data on prevalence.

Hawdon *et al.* (2017), using representative samples covering 15- to 30-year-olds in the United States, United Kingdom, Germany and Finland, found on average 43 per cent respondents had encountered hate material online (53 per cent for the United States and 39 per cent for the United Kingdom). Most hate material was encountered on social media, such as Twitter and Facebook. Ofcom (2018b), also using a representative UK sample, found that near half of UK Internet users reported seeing hateful content online in the past year, with 16- to 34-year-olds most likely to report seeing this content (59 per cent for 16–24s and 62 per cent for 25–34s). Ofcom also found 45 per cent of 12- to 15-year-olds in 2017 reported encountering hateful content online, an increase on the 2016 figure of 34 per cent (Ofcom, 2018a; 2018c).

Administrative and survey data only capture a snapshot of the online hate phenomenon. Data science methods pioneered within Computational Criminology (see Williams and Burnap, 2016; Williams *et al.*, 2017a) facilitate a real-time view of hate speech perpetration in action, arguably generating a more complete picture.[2] In 2016 and 2017, the Brexit vote and a string of terror attacks were followed by significant and unprecedented increases in online hate speech (see Figures 1 and 2). Although the production of hate speech increased dramatically in the wake of all these events, statistical models showed it was least likely to be retweeted in volume and to survive for long periods of time, supporting a 'half-life' hypothesis. Where hate speech was retweeted, it emanated from a core group of like-minded individuals who seek out each other's messages (Williams and Burnap, 2016). Hate speech produced around the Brexit vote in particular was found to be largely driven by a small number of Twitter accounts. Around 50 per cent of anti-Muslim hate speech was produced by only 6 per cent users, many of whom were classified as politically anti-Islam (Demos, 2017).

The role of popular and politically organized racism in fostering *terrestrial* climates of intimidation and violence is well documented (Bowling, 1993). The far right, and some popular right-wing politicians, have been pivotal in shifting the 'Overton window' of online political discussion further to the extremes (Lehman, 2014), creating spaces where hate speech has become the norm. Early research shows the far right were quick to take to the Internet largely unhindered by law enforcement due to constitutional protections around free speech in the United States. The outcome has been the establishment of extreme spaces that provide a collective virtual identity to previously fragmented hateful individuals. These spaces have helped embolden domestic hate groups in many countries, including the United States, United Kingdom, Germany, the Netherlands, Italy and Sweden (Perry and Olsson, 2009).

In late 2017, social media giants began introducing hate speech policies, bowing under pressure from the German government and the European Commission (Williams *et al.*, 2019).

[1] For current CPS guidance on what constitutes an online hate offence see: https://www.cps.gov.uk/legal-guidance/social-media-guidelines-prosecuting-cases-involving-communications-sent-social-media.

[2] Not all hate speech identified reaches the threshold for a criminal offence in England and Wales.

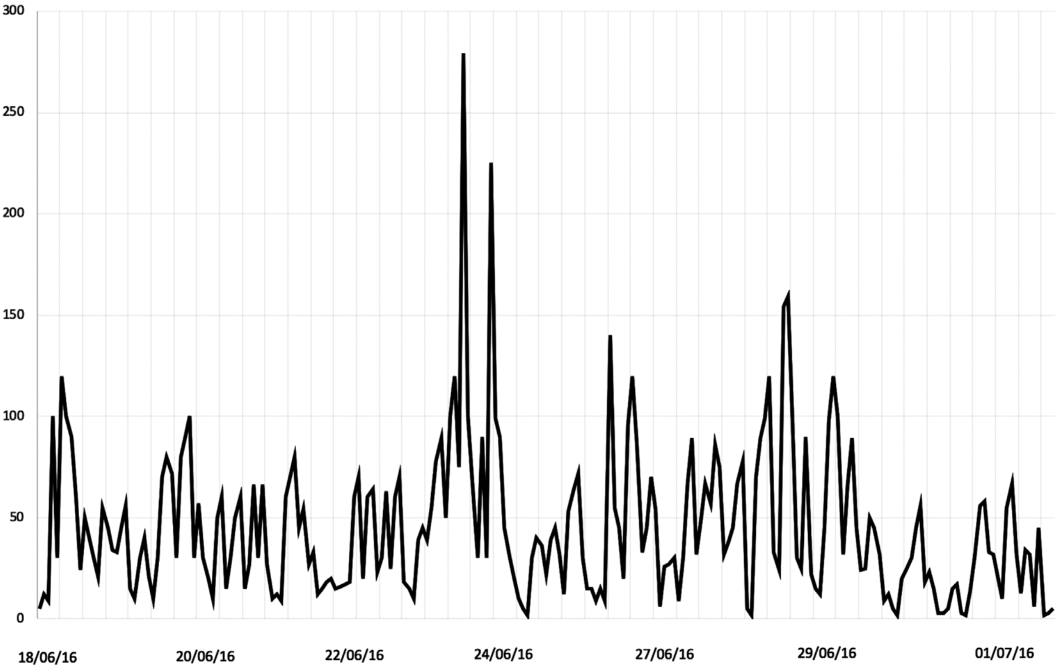


Fig. 1 UK anti-black and anti-Muslim hate speech on Twitter around the Brexit vote

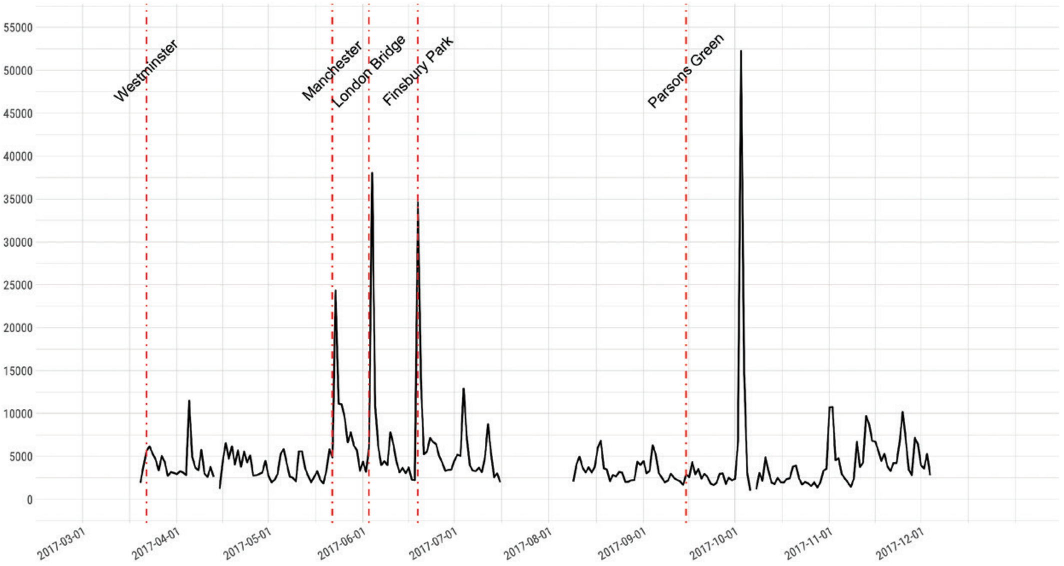
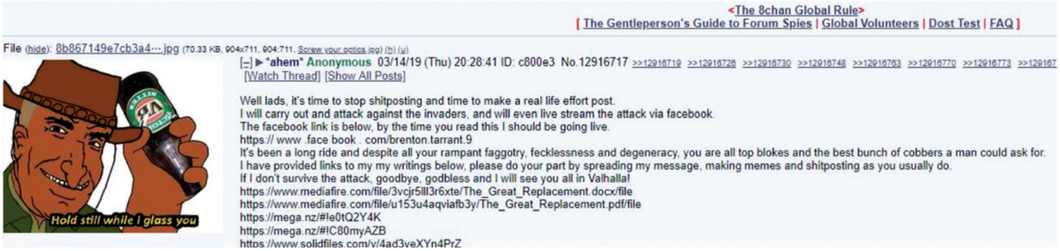


Fig. 2 Global anti-Muslim hate speech on Twitter during 2017 (gaps relate to breaks in data collection)

Up to this point, Facebook, Instagram, YouTube and Twitter were accused of 'shielding' far right pages as they generated advertising income due to their high number of followers. The 'Tommy Robinson' Facebook page, with 1 million followers, held the same protections as media and government pages, despite having nine violations of the platform's policy on hate speech, whereas typically only five were tolerated by the content review process (Hern, 2018). The page was eventually removed in March 2019, a year after Twitter removed the account of Stephen Yaxley-Lennon (alias Tommy Robinson) from their platform.

Social media was implicated in the Christchurch, New Zealand extreme-right wing terror attack in March 2019. The terrorist was an avid user of social media, including Facebook and Twitter, but also more subversive platforms, such as 8chan. 8chan was the terrorist's platform of choice when it came to publicizing his live Facebook video of the attack. His message opened by stating he was moving on from 'shit-posting'—using social media to spread hatred of minority groups—to taking the dialogue offline, into action. He labelled his message a 'real life effort post'—the migration of online hate speech to offline hate crime/terrorism (Figure 3). The live Facebook video lasted for 17 minutes, with the first report to the platform being made after the 12th minute. The video was taken down within the hour, but it was too late to stop the widespread sharing. It was re-uploaded more than 2 million times on Facebook, YouTube, Instagram and Twitter and it remained easily accessible over 24 hours after the attack. Facebook, Twitter, but particularly 8chan, flooded with praise and support for the attack. Many of these posts were removed, but those on 8chan remain due to its lack of moderation.

In the days following the terror attack spikes in hate crimes were recorded across the United Kingdom. In Oxford, Swastikas with the words "sub 2 PewDiePie" were graffitied on a school wall. In in his video ahead of the massacre, the terrorist asked viewers to 'subscribe to PewDiePie'. The social media star who earned $15.5 million in 2018 from his online activities has become known for his anti-Semitic comments and endorsements of white supremacist conspiracies (Chokshi, 2019). In his uploaded 74-page manifesto, the terrorist also referenced Darren Osborne, the perpetrator of the Finsbury Park Mosque attack in 2017. Osborne is known to have been influenced by social media communications ahead of his attack. His phone and computers showed that he accessed the Twitter account of Stephen Yaxley-Lennon two days before the attack, who he only started following two weeks prior. The tweet from Robinson read 'Where was the day of rage after the terrorist attacks. All I saw was lighting candles'. A direct Twitter message was also sent to Osborne by Jayda Fransen of Britain First (Rawlinson, 2018). Other lone actor extreme right-wing terrorists, including Pavlo Lapshyn and Anders Breivik, are also known to have self-radicalized via the Internet (Peddell *et al.* 2016).



<The 8chan Global Rule>
[ The Gentleperson's Guide to Forum Spies | Global Volunteers | Dost Test | FAQ ]

File (hide): 8b867149e7cb3a4···.jpg (70.33 KB, 904x711, 904:711, Screw your optics.jpg) (h) (u)

[–] ▶ *ahem* Anonymous 03/14/19 (Thu) 20:28:41 ID: c800e3 No.12916717 >>12916719 >>12916726 >>12916730 >>12916748 >>12916763 >>12916770 >>12916773 >>129167
[Watch Thread] [Show All Posts]

Well lads, it's time to stop shitposting and time to make a real life effort post.
I will carry out and attack against the invaders, and will even live stream the attack via facebook.
The facebook link is below, by the time you read this I should be going live.
https:// www .face book . com/brenton.tarrant.9
It's been a long ride and despite all your rampant faggotry, fecklessness and degeneracy, you are all top blokes and the best bunch of cobbers a man could ask for.
I have provided links to my my writings below, please do your part by spreading my message, making memes and shitposting as you usually do.
If I don't survive the attack, goodbye, godbless and I will see you all in Valhalla!
https://www.mediafire.com/file/3vcjr5lll3r6xte/The_Great_Replacement.docx/file
https://www.mediafire.com/file/u153u4aqviafb3y/The_Great_Replacement.pdf/file
https://mega.nz/#!e0tQ2Y4K
https://mega.nz/#!C80myAZB
https://www.solidfiles.com/v/4ad3veXYn4PrZ

FIG. 3 Christchurch extreme right terror attacker's post on 8chan, broadcasting the live Facebook video

Far right and popular right-wing activity on social media, unhindered for decades due to free-speech protections, has shaped the perception of many users regarding what language is acceptable online. Further enabled by the disinhibiting and deindividuating effects of Internet communications, and the ineffectiveness of the criminal justice system to keep up with the pace of technological developments (Williams, 2006), social media abounds with online hate speech. Online controversies, such as Gamergate, the Bank of England Fry/Austen fiasco and the Mark Meechan scandal, among many others, demonstrate how easily users of social media take to antagonistic discourse (Williams *et al.*, 2019). In recent times, these users have been given further licence by the divisive words of popular right-wing politicians wading into controversial debates, in the hopes of gaining support in elections and leadership contests. The offline consequences of this trend are yet to be fully understood, but it is worth reminding ourselves that those who routinely work with hate offenders agree that although not all people who are exposed to hate material go on to commit hate crimes on the streets, all hate crime criminals are likely to have been exposed to hate material at some stage (Peddell *et al.*, 2016).

## *Theoretical Framework*

The study relates to conceptual work that examines the role of social media in political polarization (Sunstein, 2017) and the disruption of 'hierarchies of credibility' (Greer and McLaughlin, 2010). In the United States, online sources, including social media, now outpace traditional press outlets for news consumption (Pew Research Centre, 2018). The pattern in the United Kingdom is broadly similar, with only TV news (79 per cent) leading over the Internet (64 per cent) for all adults, and the Internet, in particular social media taking first place for those aged 16–24 (Ofcom, 2018b). In the research on polarization, the general hypothesis tested is disinformation is amplified in partisan networks of like-minded social media users, where it goes largely unchallenged due to ranking algorithms filtering out any challenging posts. Sunstein (2017) argues that 'echo chambers' on social media reflecting increasingly extreme viewpoints are breeding grounds for 'fake news', far right and left conspiracy theories and hate speech. However, the evidence on the effect of social media on political polarization is mixed. Boxell *et al.* (2017) and Debois and Blank (2017), both using offline survey data, found that social media had limited effect on polarization on respondents. Conversely, Brady *et al.* (2017) and Bail *et al.* (2018), using online and offline data, found strong support for the hypothesis that social media create political echo chambers. Bail et al. found that republicans, and to a lesser extent democrats, were likely to become more entrenched in their original views when exposed to opposing views on Twitter, highlighting the resilience of echo chambers to destabilization. Brady et al. found that emotionally charged (e.g. hate) messages about moral issues (e.g. gay marriage) increased diffusion within echo chambers, but not between them, indicating this as a factor in increasing polarization between liberals and conservatives.

A recently exposed factor that is a likely candidate for increasing polarization *around events* is the growing use of fake accounts and bots to spread divisive messages. Preliminary evidence shows that these automated Twitter accounts were active in the UK-EU referendum campaign, and most influential on the leave side (Howard and Kollanyi, 2016). Twitter accounts linked to the Russian Internet Research Agency (IRA) were also active in the Brexit debate following the vote. These accounts also spread fake news and promoted xenophobic messages in the aftermath of the 2017 UK terror attacks (Crest, 2017). Accounts at the extreme-end of right-wing echo chambers were routinely targeted by the IRA to gain traction via retweets. Key political and far right figures have also been known to tap into these echo chambers to drum-up support for their campaigns. On Twitter, Donald Trump has referred to Mexican immigrants as ‘criminals and rapists’ and retweeted far right activists after Charlottesville, and Islamophobic tweets from the far right extremist group, Britain First. The leaders of Britain First, and the ex-leader of the English Defence League, all used social media to spread their divisive narrative before they were banned from most platforms between December 2017 and March 2019. These extremist agitators and others like them have used the rhetoric of invasion, threat and otherness in an attempt to increase polarization online, in the hope that it spills into the offline, in the form of votes, financial support and participation in rallies. Research by Hope Not Hate (2019) shows that at the time of the publication of their report, 5 of the 10 far-right social media activists with the biggest online reach in the world were British. The newest recruits to these ideologies (e.g. Generation Identity) are highly technically capable and believe social media to be essential to building a larger following.

Whatever the effect of social media on polarization, and how this may vary by individual-level factors, the role of events, bots and far right agitators, there remains limited experimental research that pertains to the key aim of this article: its impact on the behaviour of the public offline. Preliminary unpublished work suggests a link between online polarizing activity and offline hate crime (Müller and Shwarz, 2018a, 2018b). But what remains under-theorized is why social media has salience in this context that overrides the effect of other sources (TV, newspapers, radio) espousing arguably more mainstream viewpoints. Greer and Mclaughlin (2010) have written about the power of social media in the form of citizen journalism, demonstrating how the initially dominant police driven media narrative of ‘protestor violence’ in the reporting of the G20 demonstration was rapidly disrupted by technology-driven alternative narratives of ‘police violence’. They conclude “the citizen journalist provides a valuable additional source of real-time information that may challenge or confirm the institutional version of events” (2010: 1059). Increasingly, far right activists like Stephen Yaxley-Lennon are adopting citizen journalism as a tactic to polarize opinion. Notably, Lennon live-streamed himself on social media outside Leeds Crown Court hearing the Huddersfield grooming trials to hundreds of thousands of online viewers. His version of events was imbued with anti-Islam rhetoric, and the stunt almost derailed the trial. Such tactics take advantage of immediacy, manipulation, partisanship and a lack of accountability rarely found in mainstream media. Such affordances can provide a veil of authenticity and realism to stories, having the power to reframe their original casting by the ‘official’ establishment narrative, further enabled by dramatic delivery of ‘evidence’ of events as they occur. The ‘hacking’ of the information-communications marketplace enabled by social media disrupts the primacy of conventional media, allowing those who produce subversive “fake news” anti-establishment narratives to rise up the ‘hierarchy of credibility’. The impact of this phenomenon is likely considerable knowing over two-thirds of UK adults, and eight in ten 16- to 24-year-olds now use the Internet as their main source of news (Ofcom, 2018b).

## *Hypotheses*

The hypotheses test if online hate speech on Twitter, an indicator of right-wing polarization, can improve upon the estimations of offline hate crimes that use conventional predictors alone.

*H1*: Conventional census regressors associated with hate crime in previous research will emerge as statistically significant.

'Realistic' threats are often associated with hate crimes (Stephan and Stephan, 2000; Roberts et al., 2013). These relate to resource threats, such as competition over jobs and welfare benefits. Espiritu (2004) shows how US census measures relating to economic context are statistically associated with hate crimes at the state level. In the United Kingdom, Ray *et al.* (2004) found that a sense of economic threat resulted in unacknowledged shame, which was experienced as rage directed toward the minority group perceived to be responsible for economic hardship. Demographic ecological factors, such as proportion of the population who are black or minority ethnic and age structure, have also been associated with hate crime (Green, 1998; Nandi *et al.*, 2017; Williams and Tregidga, 2014; Ray *et al.*, 2004). In addition, educational attainment has been shown to relate to tolerance, even among those explicitly opposed to minority groups (Bobo and Licari, 1989).

*H2*: Online hate speech targeting race and religion will be positively associated with police recorded racially and religiously aggravated crimes in London.

Preliminary unpublished work focusing on the United States and Germany has showed that posts from right-wing politicians that target minority groups, deemed as evidence of extreme polarization, are statistically associated with variation in offline hate crimes recorded by the police. Müller and Shwarz (2018a) found an association between Trump's tweets about Islam-related topics and anti-Muslim hate in US state counties. The same authors also found anti-refugee posts on the far-right Alternative für Deutschland's Facebook page predicted offline-violent crime against immigrants in Germany (Müller and Shwarz, 2018b). This hypothesis tests for the first time if these associations are replicated in the United Kingdom's largest metropolitan area.

*H3*: Estimation models including the online hate speech regressor will increase the amount of offline hate crime variance explained in panel-models compared to models that include census variables alone.

Williams *et al.* (2017a) found that tweets mentioning terms related to the concept of 'broken windows' were statistically associated with police recorded crime (hate crime was not included) in London boroughs and improved upon the variance explained compared to census regressors alone. This hypothesis tests whether these results hold for the estimation of hate crimes.

## *Data and Methods*

### *Data*

The study adopted methods from Computational Criminology (see Williams *et al.*, 2017a for an overview). Data were linked from administrative, survey and social media sources to build our statistical models. Police recorded racially and religiously aggravated offences data were obtained from the Metropolitan Police Service for an eight-month period between August 2013 and August 2014. UK census variables from 2011 were derived from the Nomis web portal. London-based tweets were collected over the eight-month period using the Twitter streaming Application Programming Interface via the COSMOS software (Burnap *et al.*, 2014). All sources were linked by month and Lower Layer Super Output Area (LSOA) in preparation for a longitudinal ecological analysis.

### *Dependent measures*

*Police recorded crime.* Police crime data were filtered to ensure that only race hate crimes related to anti-black/west/south Asian offences, and religious hate crimes related to anti-Islam/Muslim offences were included in the measures. In addition to total police recorded racially and religiously aggravated offences ($N = 6,572$), data were broken down into three categories: racially and religiously aggravated violence against the person, criminal damage and harassment reflecting Part II of the Crime and Disorder Act 1998.

### *Independent measures*

*Social media regressors.* Twitter data were used to derive two measures. *Count of Geo-coded Twitter posts*—21.7 million posts were located within the 4720 London LSOAs over the study window as raw counts (Overall: mean 575; s.d. 1,566; min 0; max 75,788; Between: s.d. 1,451; min 0; max 53,345; Within: s.d. 589; min –23,108; max 28,178). *Racial and Religious Online Hate Speech*—the London geo-coded Twitter corpus was classified as 'hateful' or not (Overall: mean 8; s.d. 15.84; min 0; max 522; Between: s.d. 12.57; min 0; max 297; Within: s.d. 9.63; min –120; max 440). Working with computer scientists, a supervised machine learning classifier was built using the Weka tool to distinguish between 'hateful' Twitter posts with a focus on race (in this case anti-black/middle-eastern) and religion (in this case anti-Islam/Muslim), and more general non-'hateful' posts. A gold standard dataset of human-coded annotations was generated to train the machine classifier based on a sample of 2,000 tweets. In relation to each tweet, human coders were tasked with selecting from a ternary set of classes ('yes', 'no', and 'undecided') in response to the following question: 'is this text offensive or antagonistic in terms of race, ethnicity or religion?' Tweets that achieved 75 per cent agreement and above from four human coders were transposed into a machine learning training dataset (undecided tweets were dropped). Support Vector Machine with Bag of Words feature extraction emerged as most accurate machine learning model, with a precision of 0.89, a retrieval of 0.69 and an overall F-measure of 0.771, above the established threshold of 0.70 in the field of information retrieval (van Rijsbergen, 1979). The final hate dataset consisted of 294,361 tweets, representing 1.4 per cent of total geo-coded tweets in the study window (consistent with previous research, see Williams and Burnap, 2016; Williams and Burnap, 2018). Our measure of online hate speech is not designed to correspond directly to online hate acts deemed as criminal in the UK law. The threshold for criminal hate speech is high, and legislation is complex (see CPS guidance and Williams *et al.*, 2019). Ours is a measure of online inter-group racial and/or religious tension, akin to offline community tensions that are routinely picked up by neighborhood policing teams. Not all manifestations of such tension are necessarily criminal, but they may be indicative of pending activity that may be criminal. Examples of hate speech tweets in our sample, include: 'Told you immigration was a mistake. Send the #Muzzies home!'; 'Integrate or fuck off. No Sharia law. #BurntheQuran'; and 'Someone fucking knifed on my street! #niggersgohome'.[3]

*Census regressors.* Four measures were derived from 2011 census data based on the literature that estimated hate crime using ecological factors (e.g. Green, 1998; Espiritu, 2004). These include proportion of population: (1) with no qualifications, (2) aged 16–24, (3) long-term unemployed, and (4) black and minority ethnic (BAME).[4]

### *Methods of estimation*

The estimation process began with a single-level model that collapsed the individual 8 months worth of police hate crime and Twitter data into one time period. Because of the skewed distribution of the data and the presence of over-dispersion, a negative binomial regression model was selected. These non-panel models provide a baseline against which to compare the second phase of modelling. To incorporate the temporal variability of police recorded crime and Twitter data, the second phase of modelling adopted a random- and fixed-effects regression framework. The first step was to test if this framework was an improvement upon the non-panel model that did not take into account time variability. The Breusch–Pagan Lagrange multiplier test revealed random-effects regression was favourable over single-level regression. Random effects modelling allows for the inclusion of time-variant (police and Twitter data) and time-invariant variables (census measures). Both types of variable were grouped into the 4720 LSOA areas that make up London. Using LSOA as the unit of analysis in the models allowed for an 'ecological' appraisal of the explanatory power of race and religious hate tweets for estimating police recorded racially and religiously aggravated offences (Sampson, 2012). When the error term of an LSOA is correlated with the variables in the model, selection bias results from time-invariant unobservables, rendering random effects inconsistent. The alternative fixed-effects model that is based on within-borough variation removes such sources of bias by controlling for observed and unobserved ecological factors. Therefore, both random- and fixed-effects estimates are produced for all models.[5] A Poisson model was chosen over negative binomial, as the literature suggests the latter does not produce genuine fixed-effects (FE) estimations.[6] In addition, Poisson random-/fixed-effects (RE/FE) estimation with robust standard errors is recognized as the most reliable option in the presence of over-dispersion (Wooldridge, 1999). There were no issues with multicollinearity in the final models.

[3] These are not actual tweets from the dataset but are instead constructed illustrations that maintain the original meaning of authentic posts while preserving the anonymity of tweeters (see Williams *et al.* 2017b for a fuller discussion of ethics of social media research).

[4] Other census measures were excluded due to multicollinearity, including religion.

[5] To determine if RE or FE is preferred, the Hausman test can be used. However, this has been shown to be inefficient, and we prefer not to rely on it for interpreting our models (see Troeger, 2008). Therefore, both RE and FE results should be considered together.

[6] See https://www.statalist.org/forums/forum/general-stata-discussion/general/1323497-choosing-between-xtnbreg-fe-bootstrap-and-xtpoisson-fe-cluster-robust.

### *Results*

Figures 4–7 show scatterplots with a fitted lined (95% confidence interval in grey) of the three types of racially and religiously aggravated offences (plus combined) by race and religious hate speech on Twitter over the whole eight-month period. Scatterplots indicated a positive relationship between the variables. Two LSOAs emerged as clear outliers (LSOA E01004736 and E01004763: see Figures 8–9) and required further inspection (not included in scatter plots). A jackknife resampling method was used to confirm if these LSOAs (and others) were influential points. This method fits a negative binomial model in 4,720 iterations while suppressing one observation at a time, allowing for the effect of each suppression on the model to be identified; in plain terms, it allows us to see how much each LSOA influences the estimations. Inspection of a scatterplot of dfbeta values (the amount that a particular parameter changes when

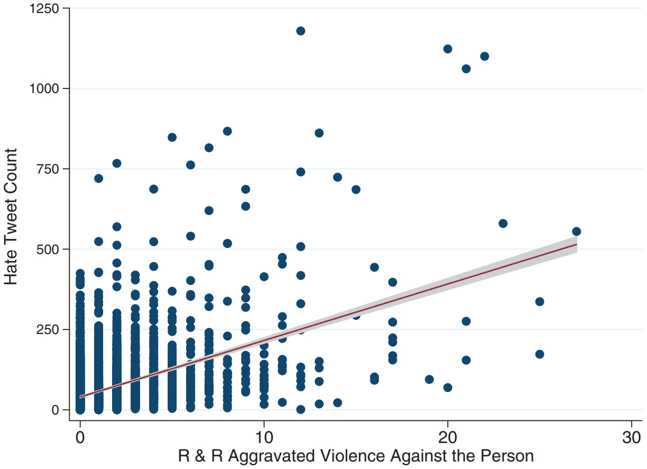


FIG. 4 Hate tweets by R & R aggravated violence against the person

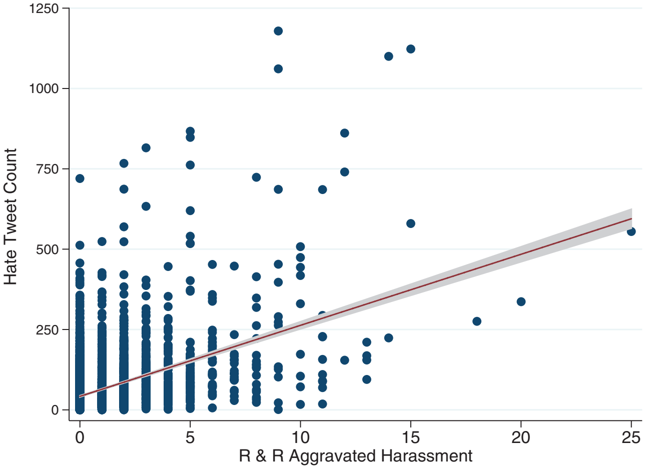


FIG. 5 Hate tweets by R & R aggravated harassment

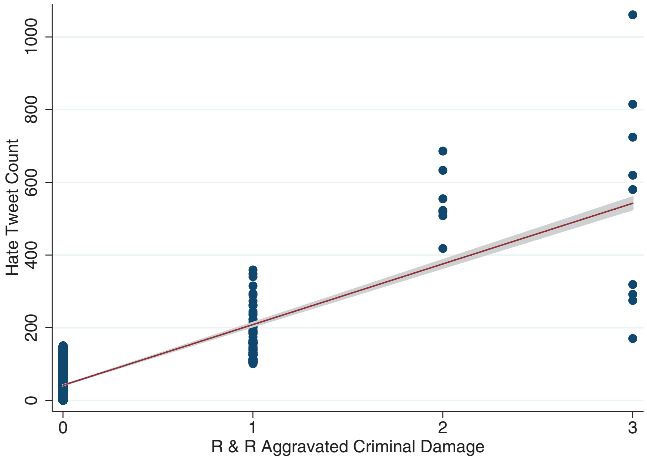


Fig. 6 Hate tweets by R & R aggravated criminal damage

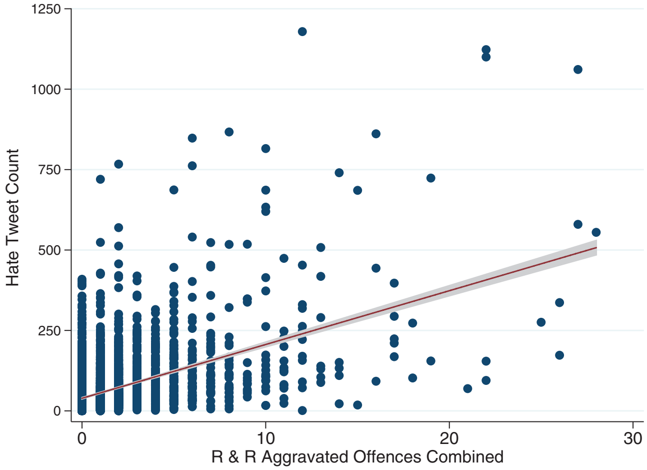


Fig. 7 Hate tweets by R & R aggravated offences combined

an observation is suppressed) confirmed the above LSOAs as influential points, and in addition E01002444 (Hillingdon, in particular Heathrow Airport) and E01004733 (Westminster). The decision was made to build all models with and without outliers to identify any significant differences. The inclusion of all four outliers did change the magnitude of effects, standard errors and significance levels for some variables and model fit, so they were removed in the final models.

Table 1 presents results from the negative binomial models for each type of racially and religiously aggravated crime category. These models do not take into account variation over time, so estimates should be considered as representing statistical associations covering the whole eight-month period of data collection, and a baseline against which to compare the panel models presented later. The majority of the census regressors emerge as significantly predictive of all racially and religiously aggravated crimes, broadly confirming previous hate crime research examining similar factors

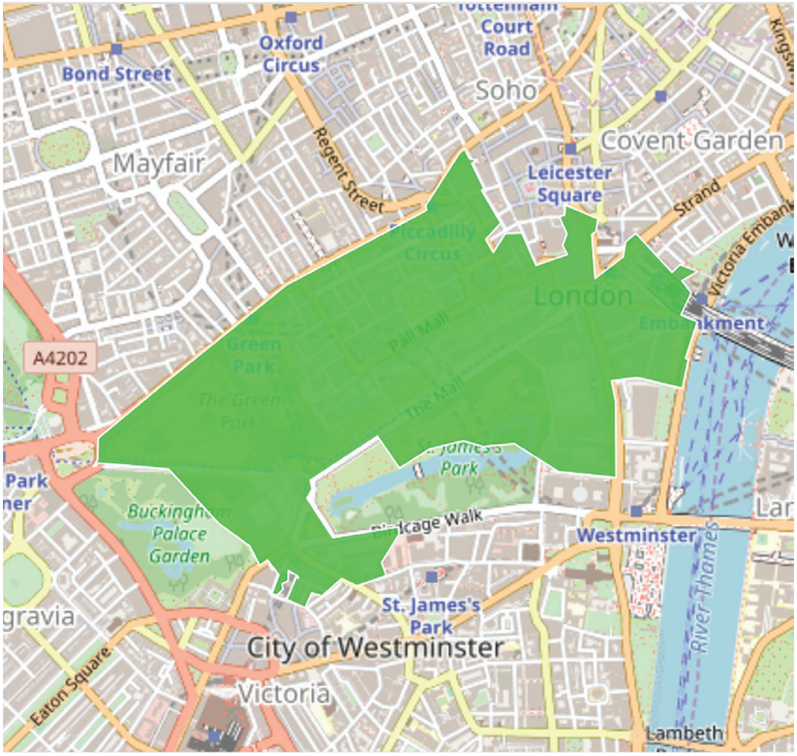


Fig. 8 Outlier LSOA E01004736

and partly supporting Hypothesis 1. Partly supporting Green (1998) and Nandi (2017) the proportion of the population that is BAME emerged as positively associated with all race and religious hate crimes, with the greatest effect emerging for racially or religiously aggravated violence against the person. Partly confirming work by Bobo and Licari (1989) models shows a positive relationship between the proportion of the population with no qualifications and racially and religiously aggravated violence, criminal damage and total hate crime, but the association only emerged as significant for criminal damage. Proportion of the population aged 16–24 only emerged as significant for criminal damage and total hate crimes, and the relationship was negative, partly contradicting previous work (Ray *et al.*, 2004; Williams and Tregidga, 2014). Like Espiritu (2004) and Ray *et al.* (2004), the models show that rates of long-term unemployment were positively associated with all race and religious hate crimes. Although this variable had the greatest effect in the models, we found an inverted U-shape curvilinear relationship (indicated by the significant quadratic term). Figure 10 graphs the relationship, showing as the proportion of the long-term unemployed population increases victimization increases to a mid-turning point of 3.56 per cent where victimization begins to decrease.

This finding at first seems counter-intuitive, but a closer inspection of the relationship between the proportion of the population that is long-term unemployed and the proportion of the population that is BAME reveals a possible explanation. LSOAs with very high long-term unemployment and BAME populations overlap. Where this

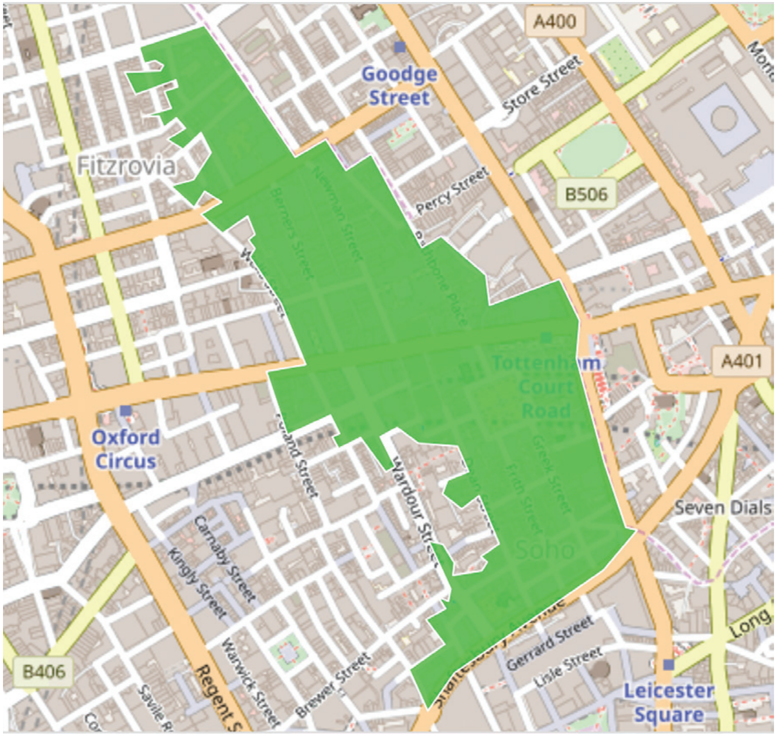


Fig. 9 Outlier LSOA E01004763

overlap is significant, we find relatively low rates of hate crime. For example, LSOA E01001838 in Hackney, in particular the Frampton Park Estate area has 6.1 per cent long-term unemployment, a 68 per cent BAME population and only 2 hate crimes, and LSOA E01003732 in Redbridge has 5.6 per cent long-term unemployment, a 76 per cent BAME population, and only 2 hate crimes. These counts of hate crime either are below or are only slightly above the mean for London (mean = 1.39, maximum = 390). We know from robust longitudinal analysis by Nandi *et al.* (2017) that minority groups living in very high majority white areas are significantly more likely to report experiencing racial harassment. This risk decreases in high multicultural areas where there is low support for far right groups, such as London. Simple regression (not shown here) where the BAME population proportion was included as the only regressor does show an inverted U-shape relationship with all hate crimes, with the risk of victimization decreasing when the proportion far outweighs the white population. However, this curve was smoothed out when other regressors were included in the models. This analysis therefore suggests that LSOAs with high rates of long-term unemployment but lower rates of hate crime are likely to be those with high proportions of BAME residents, some of whom will be long-term unemployed themselves but unlikely to be perpetrating hate crimes against the ingroup.

Supporting Hypotheses 2, all negative binomial models show online hate speech targeting race and religion is positively associated with all offline racially and religiously aggravated offences, including total hate crimes in London over an eight-month

Table 1 *Negative binomial models (full 8-month period, N = 4,270)*

| | Racially or religiously aggravated violence against the person | | | Racially or religiously aggravated harassment | | |
|---|---|---|---|---|---|---|
| | *Coef* | SE | IRR | *Coef* | SE | IRR |
| Prop. no qual | 0.00169 | 0.00236 | 1.00169 | −0.00023 | 0.00250 | 0.99977 |
| Prop. 16–24 | −0.00510 | 0.00371 | 0.99492 | −0.00724 | 0.00376 | 0.99279 |
| Prop. unmplyd | 0.62507*** | 0.05384 | 1.86838 | 0.63071*** | 0.05695 | 1.87894 |
| Prop. unmplydsqr | −0.08655*** | 0.00988 | 0.91709 | −0.08940*** | 0.01068 | 0.91448 |
| Prop. BAME | 0.00992*** | 0.00078 | 1.00997 | 0.00618*** | 0.00087 | 1.00620 |
| Tweet Freq. | 0.00005*** | 0.00001 | 1.00005 | 0.00003** | 0.00001 | 1.00003 |
| Hate Tweets | 0.00436*** | 0.00068 | 1.00437 | 0.00437*** | 0.00062 | 1.00438 |
| Constant | 1.20077 | 0.07082 | 3.32268 | 0.26735 | 0.07136 | 1.30650 |
| Pseudo $R^2$ | 0.53 | | | 0.44 | | |
| | Racially or religiously aggravated criminal damage | | | Racially or religiously aggravated offences combined | | |
| | *Coef* | SE | IRR | *Coef* | SE | IRR |
| Prop. no qual | 0.00893*** | 0.00222 | 1.00897 | 0.00372 | 0.00223 | 1.00372 |
| Prop. 16–24 | −0.00891** | 0.00354 | 0.99113 | −0.00692* | 0.00349 | 0.99310 |
| Prop. unmplyd | 0.47102*** | 0.05750 | 1.60162 | 0.58373*** | 0.05095 | 1.79271 |
| Prop. unmplydsqr | −0.06921*** | 0.01101 | 0.93313 | −0.08208*** | 0.00951 | 0.92120 |
| Prop. BAME | 0.00387*** | 0.00078 | 1.00388 | 0.00806*** | 0.00075 | 1.00809 |
| Tweet Freq. | 0.00002* | 0.00001 | 1.00002 | 0.00004*** | 0.00001 | 1.00004 |
| Hate Tweets | 0.00456*** | 0.00056 | 1.00457 | 0.00439*** | 0.00067 | 1.00440 |
| Constant | 0.69218 | 0.06849 | 1.99807 | 1.84826 | 0.06533 | 6.34879 |
| Pseudo $R^2$ | 0.39 | | | 0.52 | | |

Notes: Because of the presence of heteroskedasticity robust standard errors are presented. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$. All models significant at the 0.0000 level.

period. The magnitude of the effect is relatively even across offence category. When considering the effect of the Twitter regressors against census regressors, it must be borne in mind the unit of change needed with each regressor to affect the outcome. For example, a percentage change in the BAME population proportion in an LSOA is quite different from a change in the count of hate tweets in the same area. The latter is far more likely to vary to a much greater extent and far more rapidly (see later in this section). The associations identified in these non-panel models indicate a strong link between hateful Twitter posts and offline racially and religiously aggravated crimes in London. Yet, it is not possible with these initial models to state direction of association: We cannot say if online hate speech precedes rather than follows offline hate crime.

Table 2 presents results from RE/FE Poisson models that incorporate variation over space *and time*. RE/FE models have been used to indicate *causal pathways* in previous criminological research; however, we suggest such claims in this article would stretch the data beyond their limits. As we adopt an ecological framework, using LSOAs as our unit of analysis, and not individuals, we cannot state with confidence that area-level factors *cause* the outcome. There are likely sub-LSOA factors that account for causal pathways, but we were unable to observe these in this study design. Nevertheless, the results

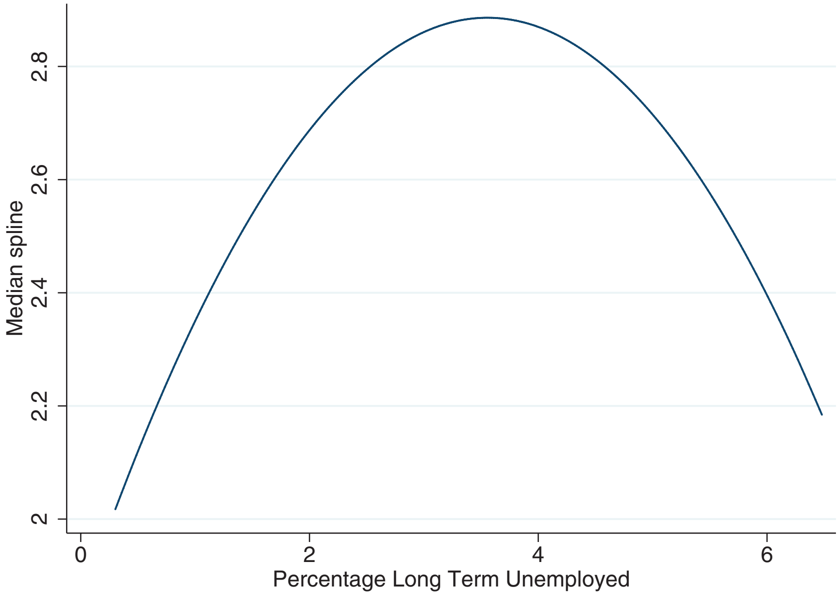


Fig. 10 Plot of curvilinear relationship between long term unemployment and racially and religiously aggravated crime.

of the RE/FE models represent a significant improvement over the negative binomial estimations presented earlier and are suitable for subjecting these earlier findings to a more robust test. Indeed, FE models are the most robust test given they are based solely on within-LSOA variation, allowing for the elimination of potential sources of bias by controlling for observed *and unobserved* ecological characteristics (Allison, 2009). In contrast, RE models only take into account the factors included as regressors. These models therefore allow us to determine if online hate speech precedes rather than follows offline hate crime.

The RE/FE modelling was conducted in three stages (Models A to C) to address Hypothesis 3—to assess the magnitude of the change in the variance explained in the outcomes when online hate speech is added as a regressor. Model A includes only the census regressors for the RE estimations, and for all hate crime categories, broadly similar patterns of association emerge compared to the non-panel models. The variance explained by the set of census regressors ranges between 2 per cent and 6 per cent. Such low adjusted R-square values are not unusual for time-invariant regressors in panel models (Allison, 2009).

Models B and C were estimated with RE and FE and introduce the Twitter variables of online hate speech and total count of geo-coded tweets. Model B introduces online hate speech alone, and both RE and FE results show positive significant associations with all hate crime categories. The largest effect in the RE models emerges for harassment (IRR 1.004). For every unit increase in online hate speech a corresponding 0.004 per cent unit increase is observed in the dependent. Put in other terms, an increase of 100 hate tweets would correspond to a 0.4 per cent increase, and an increase of 1,000 tweets would correspond to a 4 per cent increase in racially or religiously aggravated

TABLE 2 *Random and fixed-effects Poisson regression models*

| Racially or religiously aggravated violence against the person | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Model A | | | Model B | | | Model C | | |
| | *Coef* | SE | IRR | *Coef* | SE | IRR | *Coef* | SE | IRR |
| **Random model** | | | | | | | | | |
| Prop. no qual | −0.02371*** | 0.00322 | 0.97657 | −0.02094*** | 0.00308 | 0.97928 | −0.02010*** | 0.00320 | 0.98010 |
| Prop. 16–24 | 0.05212*** | 0.00833 | 1.05350 | 0.04451*** | 0.00728 | 1.04551 | 0.04265*** | 0.00742 | 1.04357 |
| Prop. unmplyd | 0.80908*** | 0.07804 | 2.24584 | 0.79596*** | 0.07534 | 2.21658 | 0.79509*** | 0.07483 | 2.21463 |
| Prop. unmplydsqr | −0.10414*** | 0.01490 | 0.90110 | −0.10288*** | 0.01435 | 0.90224 | −0.10287*** | 0.01425 | 0.90224 |
| Prop. BAME | 0.00328** | 0.00115 | 1.00329 | 0.00397*** | 0.00109 | 1.00398 | 0.00413*** | 0.00110 | 1.00414 |
| Tweet Freq. | | | | | | | 0.00001 | 0.00001 | 1.00001 |
| Hate Tweets | | | | 0.00226*** | 0.00049 | 1.00227 | 0.00134*** | 0.00029 | 1.00134 |
| Constant | −0.59539 | 0.10030 | 0.55135 | −0.58710 | 0.09520 | 0.55594 | −0.58547 | 0.09419 | 0.55685 |
| **Fixed model** | | | | | | | | | |
| Tweet Freq. | | | | | | | 0.00001* | 0.00000 | 1.00001 |
| Hate Tweets | | | | 0.00113*** | 0.00035 | 1.00113 | −0.00046 | 0.00086 | 0.99954 |
| Prop. BAME × Hate Tweets | | | | | | | 0.00009* | 0.00002 | 1.00009 |
| Adjusted $R^2$ | 0.0567 | | | 0.3039 | | | 0.3568 | | |

| Racially or religiously aggravated criminal damage | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Model A | | | Model B | | | Model C | | |
| | *Coef* | SE | IRR | *Coef* | SE | IRR | *Coef* | SE | IRR |
| **Random model** | | | | | | | | | |
| Prop. no qual | −0.00841** | 0.00268 | 0.99163 | −0.00543* | 0.00247 | 0.99459 | −0.00409 | 0.00253 | 0.99591 |
| Prop. 16–24 | 0.03228*** | 0.00574 | 1.03281 | 0.02482*** | 0.00473 | 1.02514 | 0.02234*** | 0.00473 | 1.02259 |
| Prop. unmplyd | 0.62621*** | 0.07859 | 1.87051 | 0.60606*** | 0.07492 | 1.83319 | 0.60389*** | 0.07420 | 1.82922 |
| Prop. unmplydsqr | −0.08545*** | 0.01581 | 0.91810 | −0.08326*** | 0.01507 | 0.92011 | −0.08313*** | 0.01491 | 0.92023 |
| Prop. BAME | 0.00010 | 0.00098 | 1.00010 | 0.00015 | 0.00091 | 1.00015 | 0.00021 | 0.00091 | 1.00021 |
| Tweet Freq. | | | | | | | 0.00003** | 0.00001 | 1.00004 |
| Hate Tweets | | | | 0.00353*** | 0.00065 | 1.00353 | 0.00133* | 0.00062 | 1.00133 |
| Constant | −1.20380 | 0.08824 | 0.30005 | | | | −1.20426 | 0.08319 | 0.29991 |
| **Fixed model** | | | | | | | | | |
| Tweet Freq. | | | | | | | 0.00004*** | 0.00001 | 1.00004 |
| Hate Tweets | | | | 0.00027 | 0.00039 | 1.00027 | −0.00167 | 0.00115 | 0.99833 |
| Prop. BAME × Hate Tweets | | | | | | | 0.00003* | 0.00003 | 1.00003 |
| Adjusted $R^2$ | 0.0242 | | | 0.1367 | | | 0.1537 | | |

Table 2 *Continued*

| Racially or religiously aggravated harassment | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Model A | | | Model B | | | Model C | | |
| | *Coef* | SE | IRR | *Coef* | SE | IRR | *Coef* | SE | IRR |
| **Random model** | | | | | | | | | |
| Prop. no qual | −0.02173*** | 0.00306 | 0.97851 | −0.01783*** | 0.00281 | 0.98232 | −0.01663*** | 0.00291 | 0.98351 |
| Prop. 16–24 | 0.04119*** | 0.00681 | 1.04205 | 0.03124*** | 0.00531 | 1.03173 | 0.02900*** | 0.00536 | 1.02943 |
| Prop. unmplyd | 0.80724*** | 0.07615 | 2.24172 | 0.78335*** | 0.07251 | 2.18880 | 0.78353*** | 0.07171 | 2.18918 |
| Prop. unmplydsqr | −0.10780*** | 0.01452 | 0.89781 | −0.10523*** | 0.01378 | 0.90012 | −0.10543*** | 0.01364 | 0.89993 |
| Prop. BAME | 0.00065 | 0.00111 | 1.00065 | 0.00157 | 0.00103 | 1.00157 | 0.00176 | 0.00103 | 1.00176 |
| Tweet Freq. | | | | | | | 0.00003* | 0.00001 | 1.00003 |
| Hate Tweets | | | | 0.00404*** | 0.00074 | 1.00405 | 0.00209*** | 0.00057 | 1.00209 |
| Constant | −1.59019 | 0.09197 | 0.20389 | −1.58503 | 0.08563 | 0.20494 | −1.58863 | 0.08445 | 0.20420 |
| **Fixed model** | | | | | | | | | |
| Tweet Freq. | | | | | | | 0.00004** | 0.00001 | 1.00004 |
| Hate Tweets | | | | 0.00080* | 0.00037 | 1.00080 | −0.00179 | 0.00142 | 0.99822 |
| Prop. BAME × Hate Tweets | | | | | | | 0.00008* | 0.00004 | 1.00008 |
| Adjusted $R^2$ | 0.0348 | | | 0.1692 | | | 0.1917 | | |

| Racially or religiously offences combined | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Model A | | | Model B | | | Model C | | |
| | *Coef* | SE | IRR | *Coef* | SE | IRR | *Coef* | SE | IRR |
| **Random model** | | | | | | | | | |
| Prop. no qual | −0.02009*** | 0.00297 | 0.98011 | −0.01806*** | 0.00285 | 0.98210 | −0.01727*** | 0.00295 | 0.98288 |
| Prop. 16–24 | 0.04632*** | 0.00746 | 1.04741 | 0.04084*** | 0.00672 | 1.04169 | 0.03908*** | 0.00681 | 1.03985 |
| Prop. unmplyd | 0.76556*** | 0.07448 | 2.15019 | 0.75562*** | 0.07247 | 2.12894 | 0.75444*** | 0.07197 | 2.12642 |
| Prop. unmplydsqr | −0.09988*** | 0.01447 | 0.90494 | −0.09892*** | 0.01406 | 0.90582 | −0.09887*** | 0.01396 | 0.90586 |
| Prop. BAME | 0.00196** | 0.00107 | 1.00196 | 0.00245* | 0.00103 | 1.00245 | 0.00260* | 0.00103 | 1.00261 |
| Tweet Freq. | | | | | | | 0.00001 | 0.00001 | 1.00001 |
| Hate Tweets | | | | 0.00172*** | 0.00037 | 1.00172 | 0.00093*** | 0.00026 | 1.00093 |
| Constant | 0.03797 | 0.09198 | 1.03871 | | 0.04329 | 0.08835 | 0.04491 | | 1.04593 |
| **Fixed model** | | | | | | | | | |
| Tweet Freq. | | | | | | | 0.00002** | 0.00001 | 1.00002 |
| Hate Tweets | | | | 0.00093*** | 0.00028 | 1.00094 | −0.00070 | 0.00071 | 0.99931 |
| Prop. BAME × Hate Tweets | | | | | | | 0.00004* | 0.00002 | 1.00004 |
| Adjusted $R^2$ | 0.0495 | | | 0.2937 | | | 0.3412 | | |

Notes: Table shows results of separate random and fixed effects models. To determine if RE or FE is preferred the Hausman test can be used. However, this has been shown to be inefficient, and we prefer not to rely on it for interpreting our models (see Troeger, 2008). Therefore, both RE and FE results should be considered together. Because of the presence of heteroskedasticity robust standard errors are presented. Adjusted $R^2$ for random effects models only. $*p < 0.05$; $**p < 0.01$; $***p < 0.001$. All models significant at the 0.0000 level.

harassment in a given month within a given LSOA. Given we know hate speech online increases dramatically in the aftermath of trigger events (Williams and Burnap, 2015), the first example of an increase of 100 hate tweets in an LSOA is not fanciful. The magnitude of the effect with harassment, compared to the other hate offences, is also expected, given hate-related public order offences, that include causing public fear, alarm and distress, also increased most dramatically in the aftermath the 'trigger' events alluded to above (accounting for 56 per cent of all hate crimes recorded by police in 2017/18 (Home Office, 2018)). The adjusted R-square statistic for Model B shows large increases in the variance explained in the dependents by the inclusion of online hate speech as a regressor, ranging between 13 per cent and 30 per cent. Interpretation of these large increases should be tempered given time-variant regressors can exert a significant effect in panel models (Allison, 2009). Nonetheless, the significant associations in both RE and FE models and the improvement in the variance explained provide strong support for Hypotheses 2 and 3.

Model C RE and FE estimations control for total counts of geo-coded Tweets, therefore eradicating any variance explained by the hate speech regressor acting as a proxy for population density (Malleson and Andresen, 2015). In all models, the direction of relationship and significance between online hate speech and hate crimes does not change, but the magnitude of the effect does decrease, indicating the regressor was likely also acting, albeit to a small extent, as proxy for population density. The FE models also include an interaction variable between the time-invariant regressor proportion of the population that is BAME and the time-variant regressor online hate speech. The interaction term was significant for all hate crime categories with the strongest effect emerging for racially and religiously aggravated violence against the person. Figure 11 presents a predicted probability plot combining both variables for the outcome of violent hate crime. In an LSOA with a 70 per cent BAME population with 300 hate tweets posted a month, the incidence rate of racially and religiously aggravated violence is predicted to be between 1.75 and 2. However, it must be borne in mind when interpreting these predictions, the skewed distribution of the sample. Just over 70 per cent of LSOAs have a BAME population of 50 per cent or less and 150 or less hate tweets per month, therefore the probability for offences in these areas is between 1 and 1.25 (lower-left dark blue region of the plot). This plot provides predictions based on the model estimates, meaning if in the future populations and hate tweets were to increase toward the upper end of the spectrums, these are the probabilities of observing the racially and religiously aggravated violence in London.

## *Discussion*

Our results indicate a consistent positive association between Twitter hate speech targeting race and religion and offline racially and religiously aggravated offences in London. Previous published work indicated an association around events that acted as 'triggers' for on and offline hate acts. This study confirms this association is consistent in the presence and absence of events. The models allowed us to provide predictions of the incidence rate of offline offences by proportion of the population that is BAME and the count of online hate tweets. The incidence rate for near three-quarters of LSOAs within London when taking into account these and other factors in the models remains

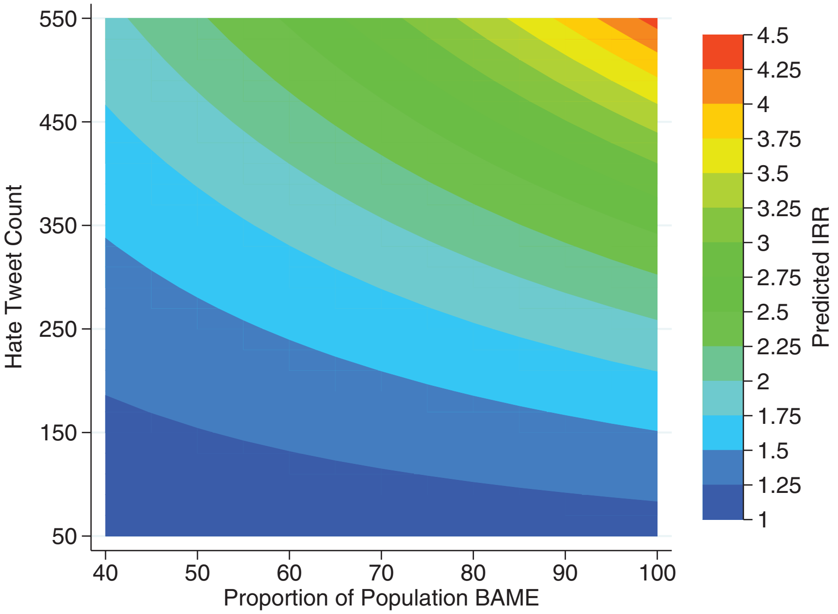


Fig. 11 Predicted probability of R & R agg. violence by BAME population proportion and hate tweet count

below 1.25. Were the number of hate tweets sent per month to increase dramatically in an area with a high BAME population, our predictions suggest much higher incidence rates. This is noteworthy, given what we know about the impact of 'trigger' events and hate speech, and indicates that the role of social media in the process of hate victimization is non-trivial.

Although we were not able to directly test the role of online polarization and far right influence on the prevalence of offline hate crimes, we are confident that our focus on online hate speech acted as a 'signature' measure of these two phenomena. Through the various mechanisms outlined in the theoretical work presented in this article, it is plausible to conclude that hate speech posted on social media, an indicator of extreme polarization, influences the frequency of offline hate crimes. However, it is unlikely that online hate speech is directly causal of offline hate crime in isolation. It is more likely the case that social media is only part of the formula, and that local level factors, such as the demographic make-up of neighbourhoods (e.g. black and minority ethnic population proportion, unemployment) and other ecological level factors play key roles, as they always have in estimating hate crime (Green, 1998; Espiritu, 2004; Ray *et al.*, 2004). What this study contributes is a data and theory-driven understanding of the relative importance of online hate speech in this formula. If we are to explain hate crime as a process and not a discrete act, with victimization ranging from hate speech through to violent victimization, social media must form part of that understanding (Bowling, 1993; Williams and Tregidga, 2014).

Our results provide an opportunity to renew Bowling's (1993) call to see racism as a continuity of violence, threat and intimidation. We concur that hate crimes must be conceptualized as a process set in geographical, social, historical and political context. We would add that 'technological' context is now a key part of this conceptualization. The enduring quality of hate victimization, characterized by repeated or continuous insult, threat, or violence now extends into the online arena and can be linked to its offline manifestation. We argue that hate speech on social media extends 'climates of unsafety' experienced by minority groups that transcend individual instances of victimization (Stanko, 1990). Online hate for many minorities is part and parcel of everyday life—as Pearson *et al.* (1989: 135) state 'A black person need never have been the actual victim of a racist attack, but will remain acutely aware that she or he belongs to a group that is threatened in this manner'. This is no less true in the digital age. Social media, through various mechanisms such as unfettered use by the far right, polarization, events, and psychological processes such as deindividuation, has been widely infected with a casual low-level intolerance of the racial *Other*.

Our study informs the ongoing debate on 'predictive policing' using big data and algorithms to find patterns at scale and speed, hitherto unrealizable in law enforcement (Kaufmann *et al.*, 2019). Much of the criminological literature is critical. The process of pattern identification further embeds existing power dynamics and biases, sharpens the focus on the symptoms and not the causes of criminality, and supports pre-emptive governance by new technological sovereigns (Chan and Bennett Moses, 2017). These valid concerns pertain mainly to predictive policing efforts that apply statistical models to data on crime patterns, offender histories, administrative records and demographic area profiles. These models and data formats tend to produce outcomes that reflect existing patterns and biases because of their historical nature. Our work mitigates some of the existing pitfalls in prediction efforts in three ways: (1) The data used in estimating patterns are not produced by the police, meaning they are immune from inherent biases normally present in the official data generation process; (2) social media data are collected in real-time, reducing the error introduced by 'old' data that are no longer reflective of the context; and (3) viewing minority groups as likely victims and not offenders, while not addressing the existing purported bias in ongoing predictive policing efforts, demonstrates how new forms of data and technology can be tailored to achieve alternative outcomes. However, the models reported in this article are not without their flaws, and ahead of their inclusion in real-life applications, we would warn that predictions alone do not necessarily lead to good policing on the streets. As in all statistics, there are degrees of error, and models are only a crude approximation of what might be unfolding on the ground. In particular, algorithmic classification of hate speech is not perfect, and precision, accuracy and recall decays as language shifts over time and space. Therefore, any practical implementation would require a resource-intensive process that ensured algorithms were updated and tested frequently to avoid unacceptable levels of false positives and negatives.

Finally, we consider the methodological implications of this study are as significant as those outlined by Bowling (1993). Examining the contemporary hate victimization dynamic requires methods that are able to capture both time and space variations in both online and offline data. Increasing sources of data on hate is also important due to continued low rates of reporting. We demonstrated how administrative (police records), survey (census) and new forms of data (Twitter) can be linked to study hate in the digital age. Surveys, interviews and ethnographies should be complemented by these new technological methods of enquiry to enable a more complete examination of the social processes which give rise to contemporary hate crimes. In the digital age, computational criminology, drawing on dynamic data science methods, can be used to study the patterning of online hate speech victimization and associated offline victimization. However, before criminologists and practitioners incorporate social media into their 'data diets', awareness of potential forms of bias in these new forms of data is essential. Williams *et al.* (2017a) identified several sources of bias, including variations in the use of social media (e.g. Twitter being much more popular with younger people). This is particularly pertinent given the recent abandonment of Twitter by many far right users following a clamp-down on hate speech in Europe. A reduction in this type of user may see a corresponding decrease in hate tweets, as they flock to more underground platforms, such as 8chan, 4chan, Gab and Voat, that are currently more difficult to incorporate into research and practical applications. The data used in this study were collected at a time before the social media giants introduced strict hate speech policies. Nonetheless, we would expect hate speech to be displaced, and in time data science solutions will allow us to follow the hate wherever it goes.

## *Conclusion*

The government publication of 'The Response to Racial Attacks and Harassment' in 1989 saw a sea-change in the way criminal justice agencies and eventually the public viewed hate crime the United Kingdom (Home Office, 1989). In 2019, the government published its *Online Harms White Paper* that tries to achieve the same with online hate (Cabinet Office, 2019). Over the past decade, online hate victims have failed to convince others that they are undeserved targets of harm that is sufficiently serious to warrant collective concern, due to insufficient empirical credibility and their subsequent unheard calls for recognition. This research shows that online hate victimization is part of a wider process of harm that can begin on social media and then migrate to the physical world. Qualitative work shows direct individual level links between online and offline hate victimization (Awan and Zempi, 2017). Our study extends this to the ecological level at the scale of the UK's largest metropolitan area. Despite this significant advancement, we were unable to examine sub-LSOA factors, meaning the individual level mechanisms responsible for the link between online and offline hate incidents remain to be established by more forensic and possibly qualitative work. The combination of the data science-driven results of this study and future qualitative work has the potential to address the reduced capacity of the police to gain intelligence on terrestrial community tensions that lead to hate crimes. Such a technological solution may even assist in the redressing of the bias reportedly present in 'predictive policing' efforts, by refocussing the algorithmic lens away from those historically targeted by police, onto those that perpetrate harms against minorities.

## *Funding*

This work was supported by the Economic and Social Research Council grant: 'Centre for Cyberhate Research and Policy: Real-Time Scalable Methods & Infrastructure for Modelling the Spread of Cyberhate on Social Media' (grant number: ES/P010695/1) and the US Department of Justice National Institute for Justice grant: 'Understanding Online Hate Speech as a Motivator for Hate Crime' (grant number: 2016-MU-MU-0009)

## References

Allison, D. P. (2009), *Fixed Effects Regression Models*. Sage.

Awan, I. and Zempi, I. (2017), 'I Will Blow Your Face Off'—Virtual and Physical World Anti-Muslim Hate Crime', *British Journal of Criminology*, 57: 362–80

Burnap, P., Rana, O., Williams, M., Housley, W., Edwards, A., Morgan, J., Sloan, L. and Conejero, J. (2014), 'COSMOS: Towards an Integrated and Scalable Service for Analyzing Social Media on Demand', *IJPSDS*, 30: 80–100.

Burnap, P. and Williams, M. L. (2015), 'Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making', *Policy & Internet*. 7: 223–42.

———. (2016), 'Us and Them: Identifying Cyber Hate on Twitter across Multiple Protected Characteristics'. *EPJ Data Science,* 5: 1–15

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. F., Lee, J., Mann, M., Merhout, F. and Volfovsky, A. (2018), 'Exposure to Opposing Views on Social Media Can Increase Political Polarization' *PNAS*, 115: 9216–21.

Bobo, L. and Licari, F. C. (1989), 'Education and Political Tolerance: Testing The Effects of Cognitive Sophistication and Target Group Affect', *Public Opinion Quarterly* 53: 285–308.

Bowling, B. (1993), 'Racial Harassment and The Process of Victimisation: Conceptual and Methodological Implications for The Local Crime Survey', *British Journal of Criminology*, 33: 231–50.

Boxell, L., Gentzkow, M., and Shapiro, J. M. (2017), 'Greater Internet Use Is Not Associated With Faster Growth In Political Polarization Among Us Demographic Groups', *PNAS*, 114: 10612–0617.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A. and Van Bavel, J. J. (2017), 'Emotion Shapes The Diffusion of Moralized Content in Social Networks', *PNAS*, 114: 7313–18.

Cabinet Office. (2019) *Internet Safety White Paper*. Cabinet Office

Chan, J. and Bennett Moses, L. (2017), 'Making Sense of Big Data for Security', *British Journal of Criminology*, 57: 299–319.

Chokshi, N. (2019), *PewDiePie in Spotlight After New Zealand Shooting*. New York Times.

CPS. (2018), *Hate Crime Report 2017–18*. Crown Prosecutions Service.

Crest (2017), *Russian Influence and Interference Measures Following the 2017 UK Terrorist Attacks*. Centre for Research and Evidence on Security Threats.

Debois, E. and Blank, G. (2017), 'The Echo Chamber is Over-Stated: The Moderating Effect of Political Interest and Diverse Media', *Information, Communication & Society*, 21: 729–45.

Demos. (2017), *Anti-Islamic Content on Twitter*. Demos

Espiritu, A. (2004), 'Racial Diversity and Hate Crime Incidents', *The Social Science Journal*, 41: 197–208.

Green, D. P., Strolovitch, D. Z. and Wong, J. S. (1998), 'Defended Neighbourhoods, Integration and Racially Motivated Crime', *American Journal of Sociology*, 104: 372–403.

Greer, C. and McLaughlin, E. (2010), 'We Predict a Riot? Public Order Policing, New Media Environments and the Rise of the Citizen Journalist', *British Journal of Criminology*, 50: 1041–059.

Greig-Midlane, J. (2014), *Changing the Beat? The Impact of Austerity on the Neighbourhood Policing Workforce*. Cardiff University.

Hanes, E. and Machin, S. (2014), 'Hate Crime in the Wake of Terror Attacks: Evidence from 7/7 and 9/11', *Journal of Contemporary Criminal Justice*, 30: 247–67.

Hawdon, J., Oksanen, A. and Räsänen, P. (2017), 'Exposure To Online Hate In Four Nations: A Cross-National Consideration', *Deviant Behavior*, 38: 254–66.

Hern, A. (2018), *Facebook Protects Far-Right Activists Even After Rule Breaches*. The Guardian.

HMICFRS. (2018), *Understanding the Difference: The Initial Police Response to Hate Crime*. Her Majesty's Inspectorate of Constabulary and Fire and Rescue Service.

Home Office. (1989), *The Response to Racial Attacks and Harassment: Guidance for the Statutory Agencies, Report of the Inter-Departmental Racial Attacks Group*. Home Office.

———. (2018), *Hate Crime, England and Wales 2017/18*. Home Office.

Hope Not Hate. (2019), *State of Hate 2019*. Hope Not Hate.

Howard, P. N. and Kollanyi, B. (2016), *Bots, #StringerIn, and #Brexit: Computational Propeganda during the UK-EU Referendum*. Unpublished Research Note. Oxford University Press.

Kaufmann, M., Egbert, S. and Leese, M. (2019), 'Predictive Policing and the Politics of Patterns', *British Journal of Criminology*, 59: 674–92.

Lehman, J. (2014), *A Brief Explanation of the Overton Window*. Mackinac Center for Public Policy.

Malleson, N. and Andresen, M. A. (2015), 'Spatio-temporal Crime Hotspots and The Ambient Population', *Crime Science*, 4: 1–8.

Müller, K. and Schwarz, C. (2018a), *Making America Hate Again? Twitter and Hate Crime Under Trump*. Unpublished working paper. University of Warwick.

———. (2018b), *Fanning the Flames of Hate: Social Media and Hate Crime*. Unpublished working paper. University of Warwick.

Nandi, A., Luthra, R., Saggar, S. and Benzeval, M. (2017), *The Prevalence and Persistence of Ethnic and Racial Harassment and Its Impact on Health: A Longitudinal Analysis*. University of Essex.

Ofcom. (2018a), *Children and Parents: Media Use and Attitudes*. Ofcom

———. (2018b), *News Consumption in the UK: 2018*. Ofcom.

———. (2018c), *Adults' Media Use and Attitudes Report*. Ofcom

ONS. (2017), *CSEW Estimates of Number of Race and Religion Related Hate Crime in England and Wales, 12 Months Averages, Year Ending March 2014 to Year Ending March 2017*. Office for National Statistics.

Pearson, G., Sampson, A., Blagg, H., Stubbs, P. and Smith, D. J. (1989), 'Policing Racism', in R. Morgan and D. J. Smith, eds., *Coming to Terms with Policing: Perspectives on Policy*. Routledge.

Peddell, D., Eyre, M., McManus, M. and Bonworth, J. (2016), 'Influences and Vulnerabilities in Radicalised Lone Actor Terrorists: UK Practitioner Perspectives', *International Journal of Police Science and Management*, 18: 63–76.

Perry, B. and Olsson, P. (2009), 'Cyberhate: The Globalisation of Hate', *Information & Communications Technology Law*, 18: 185–99.

Pew Research Centre. (2018), *Americans Still Prefer Watching to Reading the News*. Pew Research Centre.

Rawlinson, K. (2018), *Finsbury Park-accused Trawled for Far-right Groups Online, Court Told*. The Guardian.

Ray, L., Smith, D. and Wastell, L. (2004), 'Shame, Rage and Racist Violence', *British Journal of Criminology*, 44: 350–68.

Roberts, C., Innes, M., Williams, M. L., Tregidga, J. and Gadd, D. (2013), *Understanding Who Commits Hate Crimes and Why They Do It* [Project Report]. Welsh Government.

van Rijsbergen, C. J. (1979), *Information Retrieval* (*2nd ed.*), Butterworth.

Sampson, R. J. (2012), *Great American City: Chicago and the Enduring Neighborhood Effect*. University of Chicago Press.

Stanko. (1990), *Everyday Violence*. Pandora.

Stephan, W. G. and Stephan, C. W. (2000), *An Integrated Threat Theory of Prejudice*. Lawrence Erlbaum Associates.

Sunstein, C. R. (2017), *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.

Troeger, V. E. (2008), 'Problematic Choices: Testing for Correlated Unit Specific Effects in Panel Data', Presented at 25th Annual Summer Conference of the Society for Political Methodology, 9–12 July 2008.

Williams, M. L. (2006), *Virtually Criminal: Crime, Deviance and Regulation Online*. Routledge.

Williams, M. and Burnap, P. (2016), 'Cyberhate on Social Media in the Aftermath of Woolwich: A Case Study in Computational Criminology and Big Data', *British Journal of Criminology*, 56: 211–38.

———. (2018), *Antisemitic Content on Twitter*. Community Security Trust.

Williams, M. and Tregidga, J. (2014), 'Hate Crime Victimisation in Wales: Psychological and Physical Impacts Across Seven Hate Crime Victim-types', *British Journal of Criminology*, 54: 946–67.

Williams, M. L., Burnap, P. and Sloan, L. (2017a), 'Crime Sensing With Big Data: The Affordances and Limitations of Using Open-source Communications to Estimate Crime Patterns', *The British Journal of Criminology*, 57: 320–40.

———. (2017b), 'Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation', *Sociology*, 51: 1149–68.

Williams, M. L., Eccles-Williams, H. and Piasecka, I. (2019), *Hatred Behind the Screens: A Report on the Rise of Online Hate Speech*. Mishcon de Reya.

Williams, M. L., Edwards, A. E., Housley, W., Burnap, P., Rana, O. F., Avis, N. J., Morgan, J. and Sloan, L. (2013), 'Policing Cyber-Neighbourhoods: Tension Monitoring and Social Media Networks', *Policing and Society*, 23: 461–81.

Wooldridge, J. M. (1999), 'Distribution-Free Estimation of Some Nonlinear Panel Data Models', *Journal of Econometrics*, 90: 77–97.