doi:10.1093/bjc/azv059 BRIT. J. CRIMINOL. (2016) **56**, 211–238
Advance Access publication 25 June 2015

CYBERHATE ON SOCIAL MEDIA IN THE AFTERMATH OF WOOLWICH: A CASE STUDY IN COMPUTATIONAL CRIMINOLOGY AND BIG DATA

Matthew L. Williams* and Pete Burnap

This paper presents the first criminological analysis of an online social reaction to a crime event of national significance, in particular the detection and propagation of cyberhate on social media following a terrorist attack. We take the Woolwich, London terrorist attack in 2013 as our event of interest and draw on Cohen's process of warning, impact, inventory and reaction to delineate a sequence of incidents that come to constitute a series of deviant responses following the attack. This paper adds to contemporary debates in criminology and the study of hate crime in three ways: (1) it provides the first analysis of the escalation, duration, diffusion and de-escalation of cyberhate in social media following a terrorist event; (2) it applies Cohen's work on action, reaction and amplification and the role of the traditional media to the online context and (3) it introduces and provides a case study in 'computational criminology'.

Keywords: cyberhate, hate crime, social media, computational criminology, big data

Introduction

Recent research has shown that the prevalence and severity of *offline* crimes with a prejudicial component are influenced in the short term by singular or clusters of events. In particular, terrorist acts have been found to function as antecedent 'trigger' events that 'validate' prejudicial sentiments and tensions, opening up a space for the spread of hostile beliefs and the mobilization of action as a result of the desire for retribution in the targeted group. This can manifest in the amplification or escalation of deviance towards groups that share similar characteristics to the perceived terrorist perpetrators. A focus on events has allowed researchers to study the escalation, duration, diffusion and de-escalation of offline hate crimes (Legewie 2013; Hanes and Machin 2014; King and Sutton 2014). While the evidence from these studies is compelling, they say little about the ebb and flow of contemporary forms of hate that manifest online. Despite cyberhate being evident from the birth of the domestic Internet (initially with the launch of the Stormfront website in 1995[1]), it has only recently become identified as a social problem that requires addressing. The prominence of the problem is linked to the recognition that contemporary online spaces, such as social media platforms (e.g. Twitter), now represent a socio-technical assemblage that creates a new public sphere enabling digital citizenship through which key aspects of civil society are played out (Mossberger et al. 2008). The former UK Justice Secretary Chris Grayling announced

*Matthew L. Williams, Social Data Science Lab, School of Social Sciences, Cardiff University, King Edward VII Avenue, Cardiff CF10 3WT, UK; WilliamsM7@cf.ac.uk; Matthew L. Williams is a Reader in Computational Criminology at the Social Data Science Lab, School of Social Sciences, Cardiff University. Pete Burnap is a Lecturer in Computer Science at the Social Data Science Lab, School of Computer Science and Informatics, Cardiff University.

[1] Stormfront existed in bulletin board format in the early 1990s before being reformed as a website.

plans in 2014 to amend the Criminal Justice and Courts Bill to increase the maximum sentence for online abusive and hateful content from 6 to 24 months. Despite this recognition, almost all the evidence on the manifestation and prevalence of cyberhate is anecdotal. No research to date has empirically studied the *propagation* of cyberhate; the escalation, duration, diffusion and de-escalation of hate speech on computer networks.

This paper presents an analysis of cyberhate in social media networks following a 'trigger' event. Open and widely accessible social media technologies, such as Twitter, are increasingly being used by citizens on a global scale to publish content in reaction to real-world events. The rapid uptake of these technologies has resulted in a massive distributed 'social sensor net' that affords criminologists with the opportunity to identify, monitor and trace social reactions to events to the second in real time. The diffusion of information in these networks following events can manifest itself in a number of ways, ranging from support of social resilience through calls for assistance and advice (Morell et al. 2011) to the socially disruptive, through the production and contagion of misinformation and antagonistic and prejudiced commentary (Williams et al. 2013; Burnap et al. 2014). We take the murder of Lee Rigby in the terrorist attack in Woolwich, London in 2013 as our event of interest to study the manifestation, prevalence and propagation of cyberhate. We draw on Cohen's (1972) notion of action, reaction and amplification by applying the process of warning, impact, inventory and reaction to delineate a sequence of incidents that come to constitute a series of deviant social responses following the Woolwich terror attack. A computational criminological treatment of this process is possible through the collection and analysis of social media communications that are more voluminous and rapidly and continuously produced than newspaper headlines, interviews or surveys. Like Cohen, we advance the argument that the social response to an event is partly responsible for deepening its impact or causing new events through the cyclical process of action, reaction and amplification and postulate that social media reactions have the potential to act as a 'force amplifier' in contemporary forms of social response. This paper adds to contemporary debates in criminology and the study of hate in three distinct ways: (1) it provides the first analysis of the escalation, duration, diffusion and de-escalation of cyberhate in social media following a trigger event; (2) it applies Cohen's work on action, reaction and amplification and the role of the traditional media to the online social media context and (3) it introduces and provides a case study in 'computational criminology' using Big Social Data. At the time of writing, this paper represents the first criminological analysis of an online social reaction to a major crime event, in particular the detection and propagation of cyberhate on social media following a terrorist attack.

The manifestation and harms of cyberhate

Cyberhate[2] has manifested in online communications in various contexts since the Internet became popular among the general population in the mid-1990s (Williams 2006; Wall and Williams 2007). Defining cyberhate (in particular hate speech) is complex given cultural and linguistic variations. However, legal scholars have focussed on the expressive value of language in their attempts to classify hateful speech. The

[2] The practice of 'trolling' (the targeting of defamatory and antagonistic messages towards users of social media) has received press attention of late. We avoid using the term in this paper as it can encapsulate broader forms of online abuse not restricted to victims with minority or protected characteristics.

definition adopted in this study emanates from Greenawalt (1989) who states that any analysis of the law in regard to hate speech offline has to consider the extent to which this language has expressive value. He considers four criteria that might make such expressions criminal: (1) that they might provoke a response of violence; (2) that they may deeply wound those at whom the speech is directed; (3) that such speech causes offence to those that hear it and (4) that slurs and epithets have a degrading effect on social relationships within any one community. Several of these conditions are encapsulated within UK provisions including the Public Order Act of 1986, the Malicious Communications Act 1988, the Protection from Harassment Act 1997 and the Crime and Disorder Act 1998. The application of these laws and others that criminalize incitement on the basis of religion and sexual orientation[3] to the online context is relatively non-contentious as evidenced by several recent high profile social media cyberhate cases (see next section).

Despite these provisions, for over a decade much cyberhate that manifested online (pre-social media) met with little criminal justice response in the United Kingdom. Further afield, in countries like the United States, cyberhate continues largely unchallenged by law enforcement due to freedom of speech protections. Levin (2002) studied how US right-wing groups promoted their goals on the web largely unchallenged by law enforcement, concluding that the online medium has been useful to hatemongers because it is economic, far reaching and protected by the First Amendment. Perry and Olsson (2009) found that the web created a new common space that fostered a 'collective identity' for previously fractured hate groups, strengthening their domestic presence in counties such as the United States, Germany and Sweden. They warn a 'global racist subculture' could emerge if cyberhate is left unchallenged. Eichhorn (2001) focuses on how the online environment opens up the possibility for a more immediate and radical recontextualization of hate speech, while also highlighting its affordances for more effective modes of response, such as vigilantism and counter-speech. Leets (2001) in a study of the impacts of hate-related web pages found that respondents perceived the content of these sites as having an indirect but insidious threat, while Oksanen et al. (2014) show how 67 per cent of 15- to 18-year olds in their study had been exposed to hate material on Facebook and YouTube, with 21 per cent becoming victims of such material. This final study evidences how the rise of social media platforms has been accompanied by an exponential increase in cyberhate (see also Williams and Wall 2013).

Conceptual Framework

Antecedent or 'trigger' events and social media reactions

Historically criminologists have been preoccupied with *where* crimes take place. An abundance of research, particularly in the United States, focuses upon the spatial clustering of crimes in so-called hotspots, where reportedly over half of recorded crimes occur (Braga et al. 2012). Fewer studies have taken as their focus *when* crimes occur. Arguably, this is a result of a lack of fine-grained data on the temporal dimension of crime. Given the problems associated with interviewer recall, retrospective victimization surveys understandably neglect to ask detailed questions on the times crimes occurred. Police-recorded crime data may not accurately reflect the time of occurrence

[3] Racial and Religious Hatred Act 2008 and the Criminal Justice and Immigration Act 2008.

and are impoverished due to low reporting levels of hate crime (Williams and Tregidga 2014). It is also rare for researchers to gain access to such fine-grained data. Given this lack of data, and where they do exist difficulties of access, there is limited research on singular or clusters of events and their influence on the prevalence of hate crimes in the short term. Of what research exists, the focus has been upon acts of terrorism. On a European scale, Legewie (2013) established a significant association between anti-immigrant sentiment and the Bali terrorist bombing using Eurobarometer data. The usability of this data set to establish this correlation was dependent upon chance, as the terrorist event occurred during fieldwork, allowing a pre–post intervention design. King and Sutton (2014) found an association between terrorist acts and a rise in hate crime incidents in the United States. Convincingly, they show that following the 9/11 terrorist attack law enforcement agencies recorded 481 hate crimes with a specific anti-Islamic motive, with 58 per cent of these occurring within two weeks of the attack (4 per cent of the at-risk period of 12 months). In the United Kingdom, Hanes and Machin (2014) found significant increases in hate crimes reported to the police in London following 9/11 and 7/7. These latter two studies were reliant upon police-recorded hate crime data, and both sets of authors acknowledge the significant problems of non-reporting and lack of temporal granularity. The first two studies found that a sharp de-escalation was evident following the spike in hate crimes following the trigger event, indicating that event-specific motivated hate has a 'half-life'. These authors conclude hate crimes cluster in time and tend to increase, sometimes dramatically, in the aftermath of antecedent 'trigger' or galvanizing events, such as terrorist acts. They postulate that hate crimes are communicative acts, often provoked by events that incite a desire for retribution in the targeted group, towards the group that share similar characteristics to the perpetrators.

A focus on the temporal dimension of hate crimes allows for a study of their escalation, duration, diffusion and de-escalation following trigger events. However, as noted previously, there are limitations in offline data: (1) low temporal granularity; (2) issues with under reporting in official police data (particularly in the case of hate crimes) and (3) the retrospective nature of reporting and problems with witness and victim recall. Forms of naturally occurring online data, such as social media communications, while noisy and unstructured, lend themselves to temporal analysis. This is primarily due to the time-stamps that accompany all items of online social media communication. Further, researchers have argued that users of social media act like a distributed sensor network, often identifying events before the authorities and traditional media (Sakaki et al. 2010). Furthermore, users of social media are more likely to express emotional content due to phenomena such as deindividuation (Williams 2006). Therefore, we argue that following trigger events, such as terrorists acts, social media users are often first publish a reaction, and given there are now over 2.5 billion users of social media (Smith 2014), these online communications provide rapid (to the second) insight into social reaction on an unprecedented scale. Indeed, there is evidence to support this argument in the recent high profile prosecution of social media users who posted negative emotive reactions following various events. For example, in 2012, Liam Stacey was sentenced to 56 days in prison for posting racist comments on Twitter after footballer's cardiac arrest and Daniel Thomas was arrested after a homophobic message was sent to Olympic diver Tom Daley. In 2014, Isabella Sorley, John Nimmo and Peter Nunn were jailed for abusing feminist campaigner

Caroline Criado-Perez and MP Stella Creasy, and Declan McCuish was jailed for a year for tweeting racist comments about two Rangers Football Club players. In relation to the Woolwich terrorist attack, seven social media users were arrested after posting messages that were suspected of inciting racial or religious hatred (BBC 2013). While these examples of 'extreme' cyberhate are relatively rare, tens of thousands of other users posted less extreme hateful content in relation to these events, creating a dataset that can be subject to criminological inspection. What is particularly unique about social media communications, when compared to more traditional interviews or survey methods, is that the user posts can be endorsed and spread by other users (in the case of Twitter 'retweeted'), creating an information flow or propagation network that can be studied. For example, researchers can use such networks to identify what information or sentiment is being endorsed and propagated by users, and which users have the most or least influence in the spread of such messages. This locomotive, extensive and linked dimension of social media data allows criminologists to study the fine-grained (i.e. seconds instead of days, months or years) escalation, duration, diffusion and de-escalation of social reaction following events, often far in advance of research using conventional curated or administrative data.

These data therefore afford us with the possibility of computationally reconfiguring classic criminological theory. For example, in the study of action, reaction and amplification, Cohen (1972) outlined a process of warning, impact, inventory and reaction to delineate a sequence of incidents that come to constitute a deviant event or series of events. A computational criminological approach to this process is possible through the collection and analysis of social media communications that are more voluminous and rapidly and continuously produced than newspaper headlines, interviews or surveys. If, as Cohen suggests, the social response to an event is partly responsible for deepening its impact or causing new events through the cyclical process of action, reaction and amplification, then social media reactions have the potential to act as a force amplifier. The dynamics and features of this amplifier effect are directly observable in real time (down to the second) given the digital traces left behind by new 'online publics'. The initial transmission (in the case of Twitter, an original tweet) and subsequent diffusion (a retweet, considered an endorsement)[4] of a social media reaction can be used to gauge a 'warning stage'—tensions and sentiments expressed 'based on conditions out of which danger may arise' (1972: 22). Here, we may take as an example the general state of insecurity concerning the terrorist threat to the United Kingdom that may foster social media communications around potential associated risks. Following, the 'impact stage' 'during which the disaster strikes and the immediate unorganised response to the death, injury and destruction takes place' (in our example, the killing of Lee Rigby in Woolwich in 2013), social media communications are likely to spike, and further again in the 'inventory stage' 'during which those exposed to the disaster begin to form a preliminary picture of what has happened and of their own condition'. Original tweeters (and their retweeters) in these stages struggle to comprehend and clarify the impact, drawing comparisons with similar events in the past and making spurious links with potential perpetrator groups and possible motives. It is important to note that tweeters include individual citizens, groups and organizations; and media, government and police communications play a significant part in these stages, either appealing for information or attempting to shape reactions (e.g.

[4] Unless altered by the retweeter to convey their alternative view-point, then considered a modified retweet.

dispelling or fuelling rumours and speculation). During the 'reaction stage', where the 'images in the inventory [are] crystallized into more organised opinions and attitudes' social media communications focus less upon the event itself and more upon the wider implications and the 'issues'—domestic security, war in the Middle East and so on. Social media communications also reveal the extent to which 'online publics' become sensitized to individuals and groups associated with the event or implicated by it. This sensitization may represent itself as expressions in tweets of the overestimation of an increased risk of deviance and calls for an escalation of social control. Finally, a point of clear importance to this current paper, the social media reaction may affect the nature, extent and development of deviant activity from spectators, especially during the impact phase. In particular, contagion fuelled by 'rumours and the milling process' (1972: 19) in online social networks 'validate' sentiments and tensions, opening up a space for the galvanizing and spread of hostile beliefs and the mobilization of action as a result of the desire for retribution in the targeted group, manifesting in the escalation of deviance towards groups that share similar characteristics to the perceived perpetrators.

In this study, we examine the emergence and propagation of cyberhate following the Woolwich terror attack using part of Cohen's framework: impact, inventory and early reaction. We focus upon the Twitter social media network (see Data and Methods for the rationale) and analyse approximately half a million tweets two weeks following the event. We take as our focus cyberhate, and not offline hate crimes and incidents, which previous research and initial evidence show increased following the attack (see Feldman and Littler 2014; Hanes and Machin 2014). At the time of writing, this paper represents the first criminological analysis of an online social reaction to a major crime event, in particular the detection and propagation of cyberhate on social media following a terrorist attack.

Hypotheses

H1: The Woolwich terrorist attack will act as an antecedent trigger event for the publication of cyberhate on the Twitter social media network.

By examining the *presence* of cyberhate tied to the Woolwich event using a bespoke hate speech supervised machine classifier (see Burnap and Williams 2015), this first hypothesis extends the work of Legewie (2013), King and Sutton (2014) and Hanes and Machin (2014) that evidences *offline* hate incidents and crimes increase following a terrorist antecedent 'trigger' event as they operate to galvanize tensions and sentiments against the suspected perpetrators and groups associated with them.

H2: Agent Type will be significantly predictive of the production of cyberhate.

The second hypothesis tests whether the type of tweeter (media, police, political, far right political) is predictive of the production of cyberhate. We make no assumptions about the direction of associations due to a lack of previous research.

H3: The number of news headlines relating to the event will be positively associated with the production of cyberhate.

If, as Cohen suggests, the traditional media play a role in '*setting the agenda*', '*transmitting the images*' and '*claims making*' following deviant events of national interest, we anticipate a positive correlation between the number of news headlines and the production

of a deviant social reaction to the perpetrators and those that share similar individual characteristics (ethnicity, religion, etc.).

H4: Cyberhate will propagate in *size* during the impact stage, will begin to abate during the inventory stage and will die out in the reaction stage following the attack.

H5: Cyberhate will *survive over time* during the impact stage, will begin to abate during the inventory stage and will die out in the reaction stage following the attack.

The fourth and fifth hypotheses examine a *propagation* dimension by postulating cyberhate will spread as a result of contagion via 'rumours and the milling process' (Cohen 1972: 19) facilitated by the act of retweeting. Two propagation dimensions are tested: size and survivability over time. In particular, they test whether the diffusion of cyberhate is confined to Cohen's impact and inventory stages. During the impact stage, social media communications are likely to spike where online publics struggle to comprehend and clarify the impact, drawing comparisons with similar events in the past and making spurious links with potential perpetrator groups and possible motives, opening up a space for the germination of hate. During the inventory and reaction stages, details emerge of the victim, perpetrator and motive via official channels (media, police, government), and discussion moves on from the actors to issues, opening up a space for the countering of rumour and hate. These hypotheses also test whether the 'half-life' found in offline hate incident patterns following antecedent events also applies to cyberhate (Legewie 2013; King and Sutton 2014).

H6: Tweets emanating from particular Agents will be significantly predictive of information flow size and survival.

The final hypothesis tests whether the type of tweeter is predictive of the spread of non-cyberhate-related information following the event. Little is currently known about the relative influence of actors in social media networks on the flow of information following terrorist events. Therefore, this hypothesis tests the assumption that information emanating from some agents will spread significantly more in terms of size and survival, compared to other agents.

Data and Methods

Big 'social' data and computational criminology

The exponential growth and uptake of social media and the availability of vast amounts of information from these networks as interactional data to researchers has created a fundamental methodological and technical challenge for social science. The collection, analysis and representation of data for this study required collaboration with computer scientists to deal with the six Vs of Big Data: volume, variety, velocity, veracity, virtue and value (see Burnap et al. 2014; 2015; Williams et al. 2013). The authors consisting of a criminologist and a computer scientist developed an interdisciplinary methodology, dubbed *computational criminology*, that has its roots in computational social science (Edwards et al. 2013). Savage and Burrows (2007) argue that corporate giants such as Facebook, Google and Twitter have been using advanced computing to mine and interpret naturally occurring social data for half a decade. Until recently, social scientists in academia have been left behind, in an 'empirical crisis', lacking the access, infrastructure and skills to marshal these data. This paper is one of the first to report on the analysis of social media data using advanced computing techniques to answer a classic criminological question on social reactions to criminal events of national interest.

Data

The data collection period spanned a month following the terrorist event in Woolwich. Data were derived from the Twitter social media network. This network differs from others such as Facebook, in that it is public and the data are freely accessible by researchers. Twitter also has an open friendship network (non-reciprocal linking between users means that the followed are not required to follow their followers) resulting in a digital 'public agora' that promotes the free exchange of opinions and ideas. As a result, Twitter has become the primary space for online citizens to publicly express their reaction to events of national significance. A hashtag convention has emerged among Twitter users that allows tweets to be tagged to a topic that is searchable. The term 'trending' is used to describe hashtags that become popular within the tweet-stream, indicating a peak or pulse in discussion usually surrounding an event. Data were collected via the Twitter streaming Application Programming Interface based on a manual inspection of the highest trending keyword following the event (i.e. 'Woolwich'), the most common strategy in the field of information diffusion online (Yang and Counts 2010). This strategy produces robust samples due to the interactive nature of keywords and hashtags, where followers of events on Twitter actively seek out the most popular, or trending topics/hashtags in order to identify relevant information and subsequently add to the flow by replicating the keyword or hashtag in their posts. This selection procedure generates a census of tweets containing the most common keyword, and hence a large sample of all tweets about the event in question. An examination of web search trends using the 'Woolwich' keyword to query the Google Trends service indicated that an issue attention cycle around this event (the duration within which public attention to this event rises and falls away) spanned 15 days. This time window also maps onto the combined durations of Cohen's (1972) warning, impact and early reaction phases in our data set (see frequency distribution in Figure 4). This became the analysis sampling time frame for our study, during which we collected $N = 427{,}330$ tweets. The sample was subject to data preprocessing and recoding using high performance computational infrastructure prior to modelling. Given our sampling technique ensured the collection of all tweets containing the most popular term surrounding the event for 15 days, we are confident that the sample is representative of *non-trivial* information flows on Twitter.

Information propagation models

Dependent measures

We took the frequency of retweets of an original tweet surrounding the event as a *size of information flow*-dependent measure, and the duration between the first and last retweet as a *survival of information flow*-dependent measure. In terms of size, the number of retweets is a measure of public interest and endorsement of the information, while survival (or duration) is a measure of persistence of interest over time. This is consistent with previous work on modelling information diffusion in social networks (Yang and Counts 2010) (see Table 1).

Independent measures

Table 1 provides descriptive statistics of the independent measures that were input into the models. Several *Content Factors* were incorporated, including the sentiment expressed

TABLE 1 *Descriptive statistics (N = 210,807)*

Variables	Coding	Sample	
		M	SD
Dependent variables			
Size (retweets)	Range: 0–4,079	0.39	11.92
Survival (seconds)	Range: 0–1,295,876	2,399.80	27,416.34
Extremity of cyberhate	0 = none; 1 = moderate; 2 = extreme	0.01	0.10
Independent variables			
Content factors			
Sentiment	–1 = negative; 0 = neutral; 1 = positive	–0.59	0.66
Hashtag	0 = no; 1 = yes	0.33	0.47
URL	0 = no; 1 = yes	0.61	0.49
Social factors			
News Agent	0 = no; 1 = yes	0.05	0.22
Police Agent	0 = no; 1 = yes	0.01	0.02
Political Agent	0 = no; 1 = yes	0.01	0.03
Far Right Political Agent	0 = no; 1 = yes	0.01	0.06
Other Agent	0 = no; 1 = yes	0.94	0.23
Followers	Range: 0–8,820,174	6,030.70	109,986.8
Tweet Count	Range: 0–942,149	18,236.98	41,417.64
External factors			
Press Headlines	Range: 102–565	350.66	132.26
Google Searches	Range: 1–100	32.39	35.50
Control factors			
Commute Morning	0 = no; 1 = yes	0.11	0.31
Work	0 = no; 1 = yes	0.30	0.50
Commute Evening	0 = no; 1 = yes	0.17	0.39
Evening	0 = no; 1 = yes	0.25	0.43
Night	0 = no; 1 = yes	0.16	0.37
Sunday	0 = no; 1 = yes	0.14	0.35
Monday	0 = no; 1 = yes	0.10	0.30
Tuesday	0 = no; 1 = yes	0.08	0.27
Wednesday	0 = no; 1 = yes	0.06	0.23
Thursday	0 = no; 1 = yes	0.23	0.42
Friday	0 = no; 1 = yes	0.24	0.43
Saturday	0 = no; 1 = yes	0.15	0.36

Reduction in N due to removal of retweets, leaving only original tweets.

(using the established Sentistrength tool, Thelwall et al. 2010), the expression of hateful terms (see cyberhate models below) and the presence of hashtags and URLs (both of which improve discoverability and information sharing, Yang and Counts 2010). *Social Factors* of Twitter users were input, including type of tweeter (police, media, etc.), number of followers and total number of previous tweets. Frequencies of news stories that were published each day that included the term 'Woolwich' in the headline and Google Search Trends for the same keyword were both entered as *External Factors*. Based on previous research, we identified several *Control Factors* (time of day and day of week) that have been shown to influence the propagation of information flows in social media (Zarrella 2009).

Cyberhate model

Dependent measure

We entered cyberhate as a dependent measure in a third model to examine the features that enabled and inhibited hateful information flows (see Table 1). We built a

computational supervised machine classifier to learn the features of hateful tweets towards Black Minority Ethnic (BME) and religious groups following the event, allowing us to distinguish these from more general tweets. The overall precision of the classifier was 0.77, well in excess of the recommended 0.70 for scientific research (van Rijsbergen 1979). The cyberhate variable entered into the models was coded as 0 = no cyberhate; 1 = moderate cyberhate; 2 = extreme cyberhate (see Appendix and Burnap and Williams 2015 for more detail on the reliability of the classifier[5]).

Methods of estimation

Information propagation size model

A zero-inflated negative binomial model was used to fit to the data. This modelling strategy is appropriate where the dependent is skewed and over dispersed, and where there may be an excessive amount of zeros (91 per cent of the tweets had a zero count for retweets).

Information propagation survival model

The second dependent—survival—was a measure of the lifetime of an information flow. Our interest was to model the factors that affect the survival of information flows following the terrorist event through the impact, inventory and early reaction phases (Cohen 1972). For example, does expressing hateful content increase or decrease the lifetime of an information flow beyond Cohen's initial impact and inventory phases? This question can be posed as one of hazards to survival, thus we adopted Cox's proportional hazards model (Cox 1972). As the study was bounded by a 15-day data analysis window, we were mindful that some information flows may survive the study period. Based on previous research, tweets that were posted within 48 hours of the curtailment of data analysis were right-censored[6] (i.e. the last retweet may not have occurred and we assume the information flow to still be active or 'alive').

Cyberhate model

As the cyberhate-dependent variable is best described as ordinal, we adopted a generalized ordered logit model. This model is not bound by the proportional odds assumption that was violated by our data.

All independent variables were subject to exploration and all outliers were removed prior to analysis to ensure the robustness of all models.[7] Given the sample size, when interpreting the relevance of the various coefficients, we should not be over reliant on tests of statistical significance. Therefore, we have used odds ratios and calculated the incidence rate ratio (IRR) as measures of the magnitude of associations. The IRR is derived by the exponentiation of the negative binomial regression coefficients, allowing for the interpretation of retweet incidence rates (as opposed to logs of expected retweet counts).

[5] Direct link to the open access article: http://onlinelibrary.wiley.com/doi/10.1002/poi3.85/epdf.

[6] http://www.sysomos.com/insidetwitter/engagement/.

[7] The distribution of the extremity of cyberhate-dependent variable showed lower categories were more likely than higher. While there is no assumption of normality with ordered logistic regression, the proportional odds assumption must be met. As the data did not meet this assumption, we opted for a generalized ordered logit model as an alternative. For robustness, the extremity of cyberhate dependent was transformed into a binary variable (no hate/hate) and logistic regression was run. Broadly similar results were found to the generalized ordered logit model providing a degree of confidence in our model choice and results. In addition to these checks, we used the 'robust' command in Stata to obtain robust standard errors mitigating against potential data distribution problems.

Results

Data visualization

Given the uniqueness of the data set, we used the Collaborative Online Social Media Observatory (COSMOS) platform for initial visualization (Williams et al. 2011; Burnap et al. 2014). Figures 1–3 provide snapshots of the COSMOS Dashboard showing geographic and temporal distributions of Twitter communications in the 15-day analysis window. As the content of tweets is not directly quotable in academic research,[8] the COSMOS platform allows for the visualization of tweet content in the form of a WordCloud that presents an aggregate overview of the thousands of posts, while maintaining the anonymity and confidentiality of users (see Williams et al. 2013).[9] In the first hour of data collection (Figure 1), we can identify a relatively sparse geographic distribution of Twitter traffic across the United Kingdom (far left) and London (centre bottom). This is not unusual given that approximately 1 per cent of Twitter users enable their geo-location (Sloan et al. 2013; 2015). Nevertheless, the relative distribution over time allows us to monitor the spread of social reaction. The WordCloud (centre top) is based on the full Twitter sample in the first hour of analysis (10,080 original tweets), not just geo-located data. Therefore, this summary of content provides a window on the thousands of original commutations being sent during the initial reaction on social media, where size of word represents frequency of use in this period. The content in this early stage reflects the act ('attack', 'killing', 'murder', 'london'), speculation as to the perpetrators' nationalities ('nigerian') religious backgrounds and possible motivation ('islam', 'muslim', 'religion') and is devoid of any details on the victim apart from possibly gender ('man'). The focus on the perpetrator was likely fuelled by the YouTube video of the attacker that was uploaded within minutes of the event. Of particular salience to this paper is the presence of the terms 'edl' and 'hate'. A closer inspection shows that the former term was being used to discuss the English Defence League's (EDL) various activities, mostly in a negative tone (e.g. criticism of a speech made by Tommy Robinson about Woolwich, and the rejection of an EDL donation to the charity Help for Heroes), while the latter term was being used in counter-hate speech tweets (e.g. shame on the EDL and British National Party (BNP) for spreading hate), as opposed to hateful tweets. Given the relatively low number of hateful tweets compared to the overall volume of communications (see below), no racist or religious slurs appear in the WordCloud. This initial hour of the study window, characterized by a lack of firm details and a degree of speculation, is akin to Cohen's (1972: 22) impact stage in which citizens display an 'unorganised response to the death, injury and destruction'.

Figure 2 is a snapshot of the first four days in the study window. What is immediately apparent is the geo-tagged communications are more voluminous and widespread across the United Kingdom, with concentration in line with population density (Sloan et al. 2013). London emerges as a hotspot for communications around the event, in particular

[8] Twitter Terms of Service forbid the anonymization of tweet content (screen-name must always accompany tweet content), meaning that ethically, informed consent should be sought from each tweeter to quote their post in research outputs. However, this is impractical given the number of posts generated and the difficulty in establishing contact (a direct private message can only be sent on Twitter if both parties follow each other). Therefore, it is not ethical to directly quote tweets that identify *individuals* without prior consent.

[9] Tweets from public organizations, such as government departments and police services, are deemed quotable as no individual can be identified.

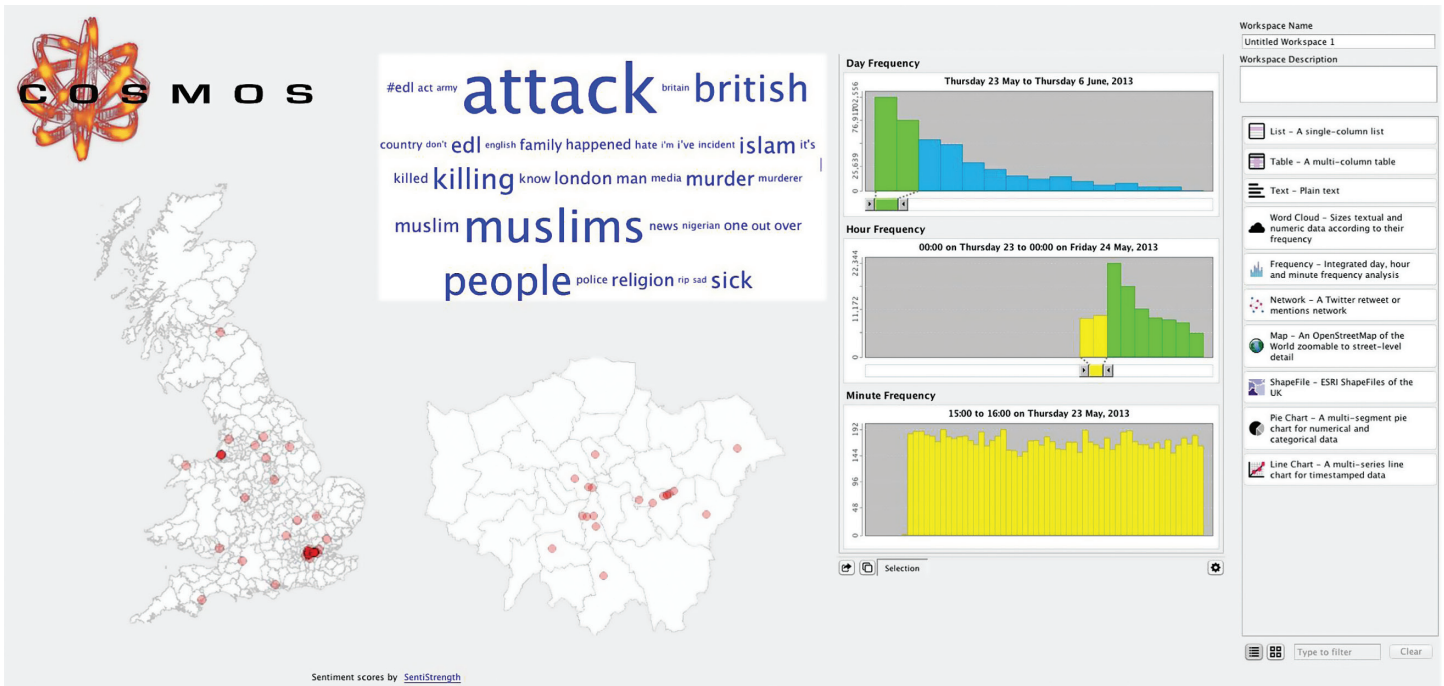

FIG. 1 Woolwich impact stage—COSMOS visualization of Twitter traffic during the first hour of collection (left to right: geo-located tweets in United Kingdom and London, WordCloud of tweet content and frequency of tweets in the 15-day collection window).

the Woolwich region. The WordCloud (133,275 original tweets) shows the content of communications has shifted from disorganized speculation to specific details on the identity of the victim ('lee', 'rigby', 'drummer'), a perpetrator ('michael', 'adebolajo') and the progress of the case ('arrested'). This more organized form of communication resonates with Cohen's inventory stage, where the mass media begin to play a central role, communicating to citizens the 'preliminary picture' of the event. However, in our case, it is worthy of note that the identity of both perpetrators was broadcast on Twitter hours ahead of conventional media. Figure 3 is a snapshot of the final three days of the study window. The geo-tagged communications plots represent the full 15 days and the cumulative volume of data now allows us to plot a more complete picture of the national social response, where we can observe dense clusters around London, the Midlands and Manchester (Lee Rigby's family home). The WordCloud (11,399 original tweets) shows Twitter communications have shifted from the specific details (victim, perpetrator, case status) onto boarder issues linked to the event ('british', 'muslims', 'islam', 'religion', 'terrorism', 'media', 'edl', 'cameron'). This shift was possibly influenced by widespread media converge of the speeches made by David Cameron and Tony Blair in the latter stages of the study window. Cameron's mention of the EDL in his speech (Cabinet Office 2013) is partly responsible for their presence in the WordCloud, while the presence of 'terrorism' for the first time is likely due to the use of the term in conjunction with radicalization and Islam in the speech by Tony Blair (Thompson 2013). This shift towards the broader issues related to the event is akin to Cohen's reaction stage, where the images emerging in the inventory are 'crystallized into more organised opinions and attitudes'. The potential influence of the media in 'transmitting the images' and 'setting the agenda' (Cohen 2002: xxiii) throughout these various stages is investigated later in the paper. In the following sections, we scope in from this '10,000 foot view' of the data by modelling the specific enablers and inhibiters of hateful and non-hateful information propagation following the event.

Cyberhate model

Of the 210,807 original tweets posted about the terrorist event in the 15 days following, 1,878 tweets (1 per cent) were identified by the validated supervised machine classifier (Burnap and Williams 2015) as containing BME or religious hate-related terms at the moderate (e.g. 'send them home', 'deport them', etc.) or extreme (e.g. 'niggers', 'muslim scum') level. This shows that *event-specific* cyberhate targeted towards the perpetrators and BME and religious groups associated with them was present in social media communications following the Woolwich terrorist attack, supporting the first hypothesis (H1).[10] The targeted nature of the cyberhate (e.g. containing the term Woolwich) demonstrates that the attack acted as an antecedent trigger event, galvanizing tensions and sentiments against groups that shared similar characteristics to the suspected perpetrators. This is the first evidence to suggest spikes in hate crimes and incidents following such events are not confided to offline settings (Legewie 2013; Hanes and Machin 2014; King and Sutton 2014). Table 2 details the results of the

[10] As we did not collect data before the event (as the schedules of terrorists are not made public knowledge), we cannot verify that this amount of cyberhate is an increase on the pre-event condition Twitter data could be purchased from a vendor (e.g. GNIP or DataSift) in an attempt to verify this hypothesis. However, the deletion of cyberhate tweets from this data set by perpetrators themselves and by Twitter as a result of complaints made by victims, means these purchased data sets are inaccurate reflections of the social media response to events. Our data set was not subject to deletions of this kind as we collected in real time.

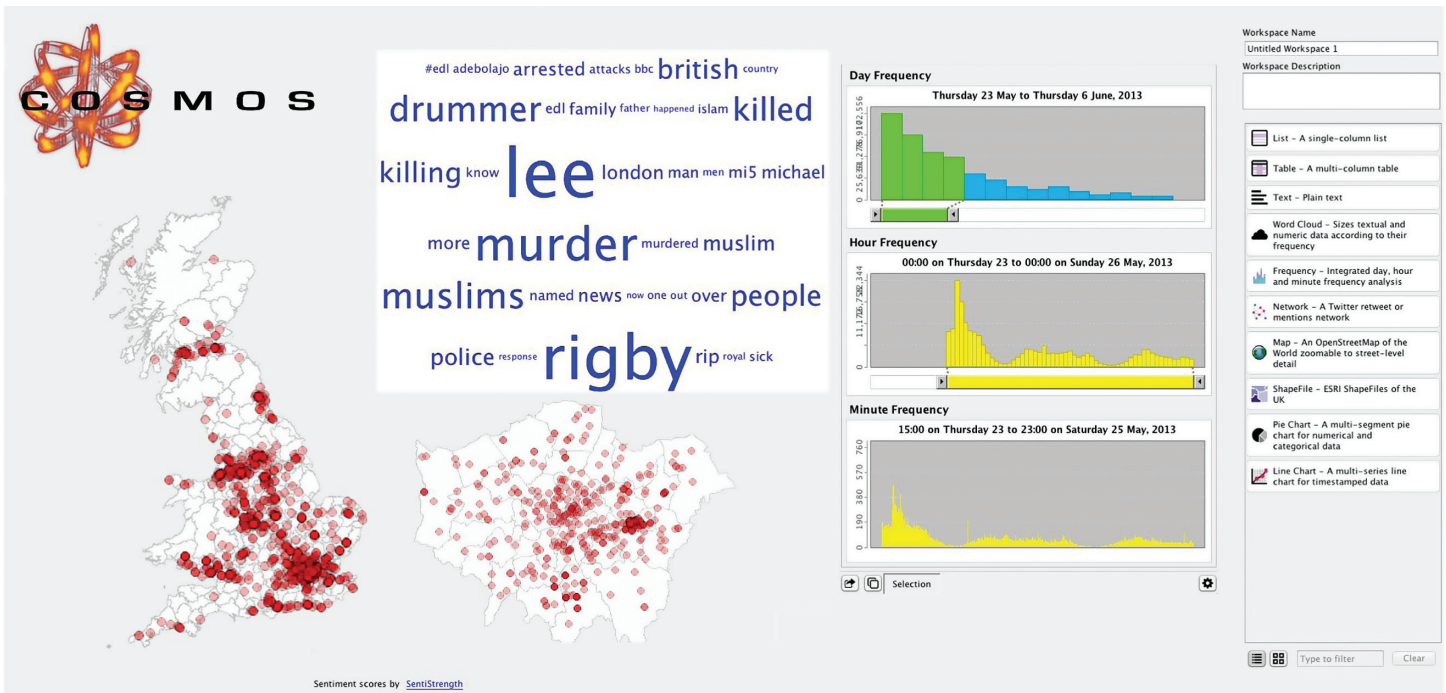

FIG. 2 Woolwich inventory stage—COSMOS visualization of Twitter traffic during the first four days of collection (left to right: geo-located tweets in United Kingdom and London, WordCloud of tweet content and frequency of tweets in the 15-day collection window).

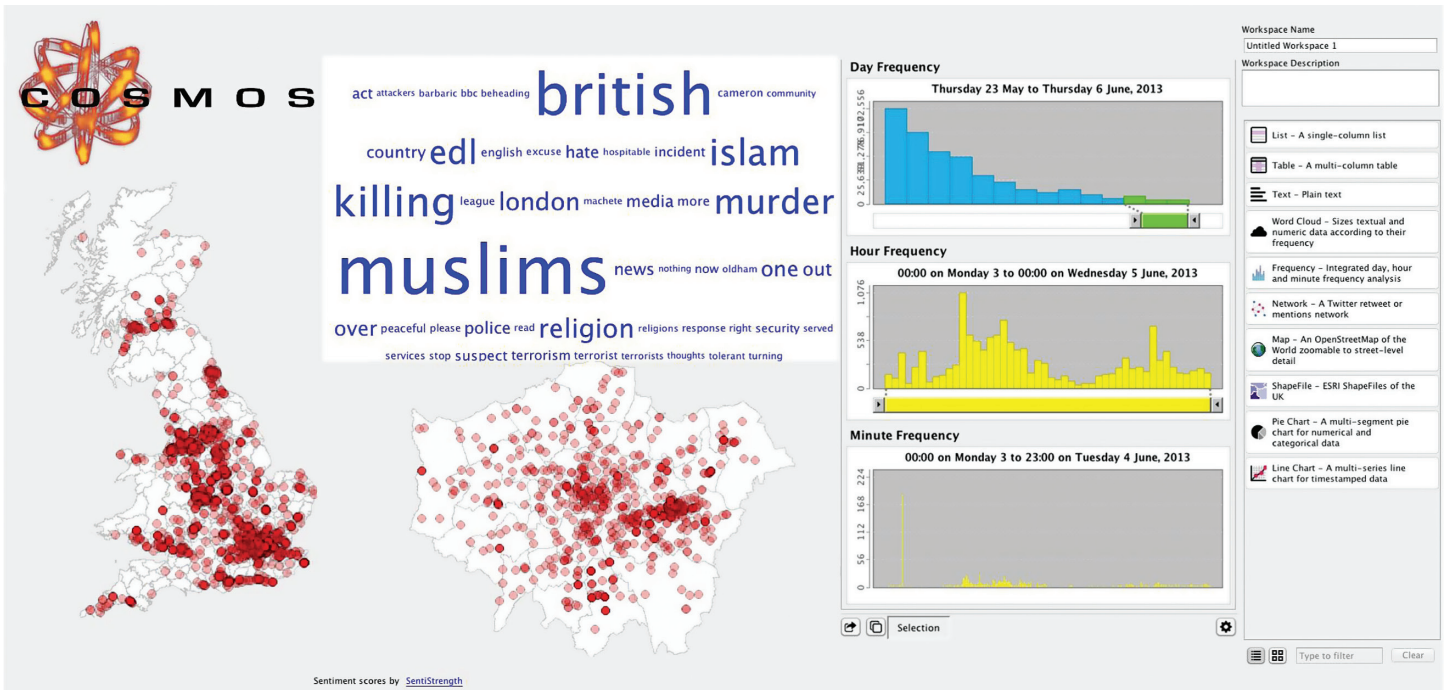

FIG. 3 Woolwich reaction stage—COSMOS visualization of Twitter traffic during the final three days of collection (left to right: geo-located tweets in United Kingdom and London, WordCloud of tweet content and frequency of tweets in the 15-day collection window).

generalized ordered logit model predicting production and extremity of cyberhate. Other Agent was entered as a reference category given their volume in the data set (94 per cent) and their proclivity to produce cyberhate (these agents produced 95 per cent of the cyberhate in the data set). The majority of Twitter users in this category were members of the 'digital public'. Of the alternative agents identified in the data set, only Far Right Political Agents emerged as significantly associated with the production of cyberhate, supporting hypothesis H2. Political organizations (e.g. BNP, EDL) and party members made up the minority of these agents, with more general far right identifying groups and individuals making up the majority. Proportionally, compared to Other Agents, the odds of producing cyberhate on Twitter were three times larger for these agents following the terrorist event. A closer inspection of the tweets produced by these agents reveals that it was the more general far right identifying groups and individuals who produced cyberhate and mostly at a moderate level (e.g. 'send them home'). The odds for including hashtags were higher for cyberhate

Table 2 *Generalized ordered logit regression predicting production and extremity of cyberhate*

	Coef.	SE	OR
Content factors			
Sentiment	0.139	0.032	1.149
Hashtag	0.688*	0.048	1.990
URL	−0.893*	0.049	0.409
Social factors			
News Agent	0.134	0.112	1.144
Police Agent	−3.479	0.34	0.000
Political Agent	0.009	0.337	1.009
Far Right Political Agent	1.333*	0.458	3.793
Ref: Other Agent			
Tweet Count	−0.007	0.038	0.993
External factors			
Press Headlines	0.003*	0.000	1.003
Google Searches	−0.007	0.002	0.993
Control factors			
Commute Morning	0.597*	0.093	1.817
Work	0.321*	0.081	1.378
Commute Evening	0.041	0.097	1.041
Evening	0.516*	0.083	1.675
Ref: Night			
Sunday	−0.409*	0.120	0.664
Monday	0.054	0.116	1.056
Tuesday	0.222	0.121	1.248
Thursday	−0.190	0.146	0.827
Friday	−0.983*	0.150	0.374
Saturday	−0.058	0.112	0.944
Ref: Wednesday (day of attack)			
Model fit			
Log likelihood			−10,823
Chi-square			945.46
Sig.			0.00
Pseudo R^2			0.04
N^a			210,807

[a]Reduction due to removal of retweets, leaving only original tweets.

*$p < 0.01$.

tweets, while they were lower for containing URLs. This may suggest those wishing to promote hateful content online and sensitize others to minority groups in society via contagion use hashtags to enhance the discoverability of their content. Conversely, URLs are possibly less common in hateful tweets given linked content (most often a popular media source) is unlikely to corroborate racist opinion and biased speculative rumours. The positive association between a rise in news headlines about the event and cyberhate tweets evidences the link (albeit relatively weak) between old and new media, supporting hypothesis H3. The odds of the production of extreme cyberhate increased by a magnitude of 1.3 for every 100 additional news headlines produced. During early stages following the event (e.g. impact and inventory), tweeters may be fuelled by coverage in the press who have a role in 'setting the agenda' and 'transmitting the images' (Cohen 2002: xxiii), especially those who wish to spread hate, biased rumours and speculation.

Information propagation size model

Table 3 reports the results from the information propagation *size* model. The most novel and salient finding was that tweets containing cyberhate were negatively associated with the size of information flows emanating from the Woolwich terrorist event. None of these tweets were statistically likely to form *large* information flows following the event. Tweets containing hate terms were 45 per cent less likely to be retweeted as compared to tweets not containing such content (IRR 0.55). Figure 4 shows that original and retweeted cyberhate peaked during the early stages following the event (impact stage) and sharply declined over the four days following (inventory stage), providing evidence in support of hypothesis H4.[11] Given that terrorist events have been shown to increase levels of anti-immigrant sentiment (Legewie 2013) and hate crimes and incidents offline (Hanes and Machin 2014; King and Sutton 2014), it is surprising to find a lack of cyberhate *propagation* in terms of size following the Woolwich attack. In line with this finding, we found that positive sentiment increased the retweet IRR by a factor of 1.38. Therefore, tweets containing positive words and phrases (e.g. 'warm wishes to the family of Lee Rigby', 'brave family', 'respect for armed forces', etc.) as opposed to negative words or phrases were 38 per cent more likely to form large information flows.

Finally, type of agent emerged as significant, further supporting hypothesis H6. Compared to the reference category Other Agent, tweets emanating from News Agents were more likely to be retweeted by a factor of 4.3, providing evidence to support the notion that traditional media messages maintain their role in '*setting the agenda*' and '*transmitting the images*' (Cohen 2002: xxiii) in the age of social media. Police Agents were also more likely to be retweeted by a factor of 5.7. The first finding is novel and shows for the first time that Twitter users propagate police tweets following terrorist events in the United Kingdom. It is evident that users are sharing police requests for information (e.g. metpoliceuk: 'We are appealing for anyone who may have witnessed the incident in #Woolwich to contact us via the Anti-Terrorist Hotline'), case updates (e.g. metpoliceuk: 'Two men aged 22 and 28 arrested on suspicion of murder

[11] The peak in moderate cyberhate six days into the study window is associated with the EDL march in London.

Table 3 *Zero-inflated negative binomial regression predicting counts of retweets (size model)*

	Coef.	SE	IRR
Content factors			
Cyberhate	−0.604**	0.017	1.721
Sentiment	0.322**	0.018	1.380
Hashtag	0.217**	0.025	1.242
URL	0.438**	0.026	1.551
Social factors			
News Agent	1.460**	0.044	4.304
Police Agent	1.742**	0.408	5.708
Political Agent	0.670**	0.150	1.954
Far Right Political Agent	0.632*	0.327	1.882
Ref: Other Agent			
Tweet Count	−0.215**	0.00	1.000
External factors			
Press Headlines	0.000*	0.000	1.000
Google Searches	0.005**	0.001	1.005
Control factors			
TimeLagRT5	0.000**	0.000	1.000
Commute Morning	0.030	0.048	1.030
Work	0.014	0.038	1.015
Commute Evening	0.074	0.045	1.077
Evening	−0.010	0.040	0.990
Ref: Night			
Sunday	−0.133**	0.041	0.875
Monday	−0.302**	0.047	0.739
Tuesday	−0.344**	0.051	0.709
Thursday	−0.365**	0.044	0.733
Friday	−0.311**	0.044	0.733
Saturday	−0.122**	0.052	0.853
Ref: Wednesday (day of attack)			
Binomial model (Inflation/Excess Zeros)			
Number of Followers	−0.899**	0.017	
Constant	4.586	0.063	
Model fit			
Log likelihood			−92,196.36
Chi-square			2,594.57
Sig.			$p = 0.00$
LRT for alpha = 0			$p = 0.00$
Vuong			$Z = 45.00$, $p = 0.00$
N[a]			210,807

[a]Reduction due to removal of retweets, leaving only original tweets.

*$p < 0.05$; **$p < 0.01$.

remain in hospital in a stable condition #woolwich') and general commentary (e.g. PoliceFedICC: 'EDL marches on Newcastle as attacks on Muslims increase tenfold in the wake of Woolwich machete attack').[12] Evidence from the United States suggests this pattern of Twitter user behaviour was also evident following the Boston Marathon terrorist attack (Davis et al. 2014). The finding that Far Right Political Agents were the least likely to have content retweeted is consistent with the previous finding in relation to cyberhate.

[12] Usernames and text reproduced here as the tweet accounts belong to public organizations, i.e. the Metropolitan Police and the Police Federation.

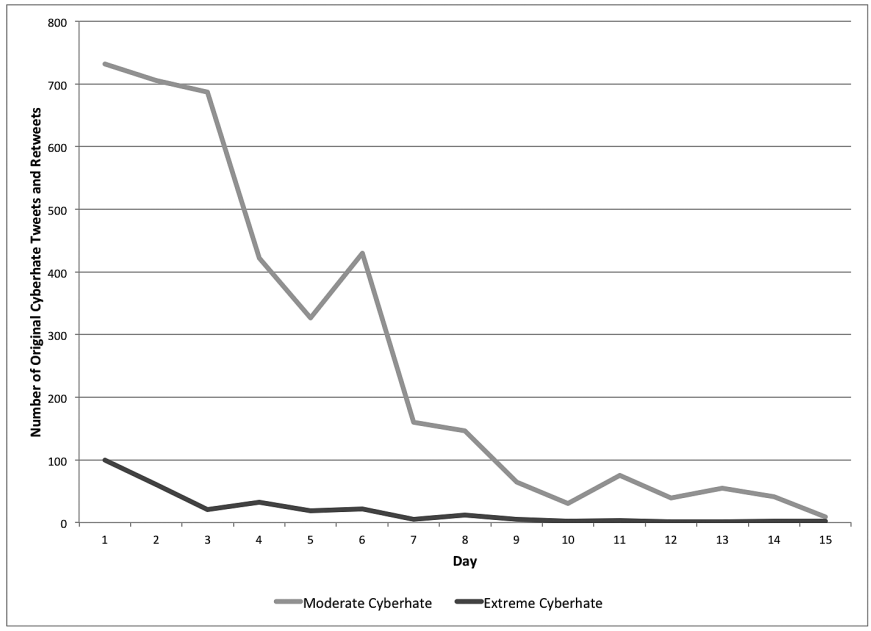

FIG. 4 Frequency of cyberhate during the analysis window.

Survival model

A positive estimate in the Cox regression model (Table 4) is interpreted as increasing hazards to survival and therefore reduces the duration of the information flow. In support of hypothesis H5, cyberhate is negatively associated with long-lasting information flows, emerging as having the highest positive hazard ratio (1.19) of the variables of interest. Supporting hypothesis H6, News Agents emerged as significantly negatively associated with hazards to survival, indicating tweets from such agents were likely to last longer in the study period. Counter intuitively, Far Right Political Agents emerged as having the second highest negative hazard ratio, after Police Agents, indicating that information flows emanating from these types of agents were likely to outlast those emanating from other agents at some point in the 15-day analysis window. To better aid interpretation, we used Kaplan–Meier (KM) survival estimation to plot the survival functions of cyberhate and Agent Type. Figure 5 illustrates the comparative survival rates of cyberhate information flows, showing that flows containing extreme BME or religious hate terms die out rapidly, between 20 and 24 hours following the event. Tweets containing moderate BME or religious hate terms last a little longer, between 36 and 42 hours before dying out. Tweets containing no cyberhate show a longer survival curve. This evidence confirms that extreme cyberhate was propagated in social media networks following this event in the immediate impact stage, which was then replaced with the propagation of moderate cyberhate in the inventory stage, and finally little or no propagation of cyberhate in the early reaction phase (Cohen 1972). This sharp de-escalation resonates with the work

Table 4 *Cox regression predicting hazards to tweet survival (survival model)*

	Coef.	SE	Hazard ratio
Content factors			
Cyberhate	0.171**	0.073	1.186
Sentiment	−0.041**	0.011	0.960
Hashtag	−0.049**	0.016	0.952
URL	−0.437**	0.017	0.646
Social factors			
Number of Followers	0.000**	0.000	1.000
News Agent	−0.082**	0.030	0.922
Police Agent	−0.572*	0.268	0.565
Political Agent	−0.112	0.088	0.894
Far Right Political Agent	−0.417**	0.148	0.659
Ref: Other Agent			
Tweet Count	0.062**	0.013	1.064
External factors			
Press Headlines	0.001**	0.000	1.001
Google Searches	−0.001*	0.001	0.999
Control factors			
TimeLagRT5	0.000**	0.000	1.000
Commute Morning	0.163**	0.032	1.117
Work	0.210**	0.025	1.233
Commute Evening	0.252**	0.029	1.287
Evening	0.201**	0.027	1.105
Ref: Night			
Sunday	0.252**	0.045	1.287
Monday	0.216**	0.048	1.242
Tuesday	0.060	0.050	1.062
Thursday	0.213**	0.055	1.237
Friday	0.056	0.052	1.057
Saturday	0.099*	0.045	1.105
Ref: Wednesday (day of attack)			
Model fit			
Log likelihood			−152,650.40
Chi-square			1,473.47
Sig.			$p = 0.00$
N^a			210,807

[a]Reduction due to removal of retweets, leaving only original tweets.
*$p < 0.05$; **$p < 0.01$.

of Legewie (2013) and King and Sutton (2014) who postulate that the increase in offline anti-immigration sentiment and hate crimes and incidents following terrorist events has a half-life. It seems likely that this offline pattern is replicated online in social media networks. Figure 6 represents the KM survival estimates plot for Agent Type. It is evident that information flows emanating from Far Right Political Agents outlast all other agent types up to 36–42 hours after the event, at which point they lose ground to News Agents, whose information flows last the longest (excusing Other Agent). The survivability of tweets emanating from Police Agents in the first 24-hour window and their subsequent demise at around 36 hours is a novel and policy-relevant finding. Why the Far Right and Police are so dominant in terms of information flow survival in the early stages of the reaction to this terrorist event is explored further in the next section.

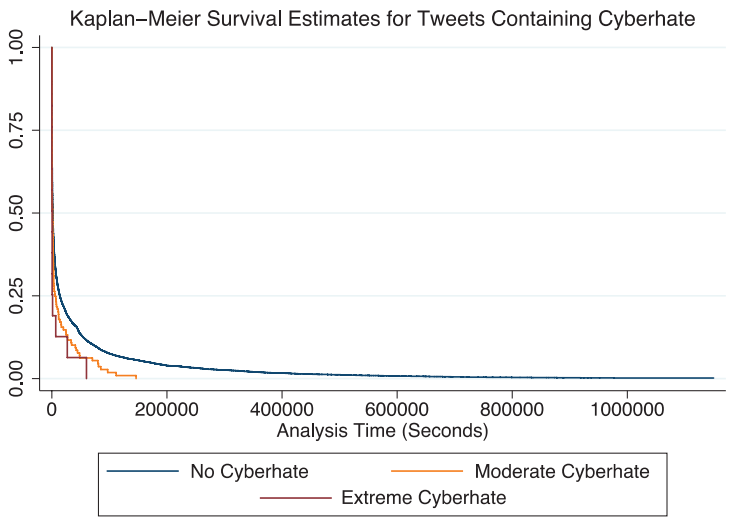

FIG. 5 Survival of Twitter information flows by cyberhate content.

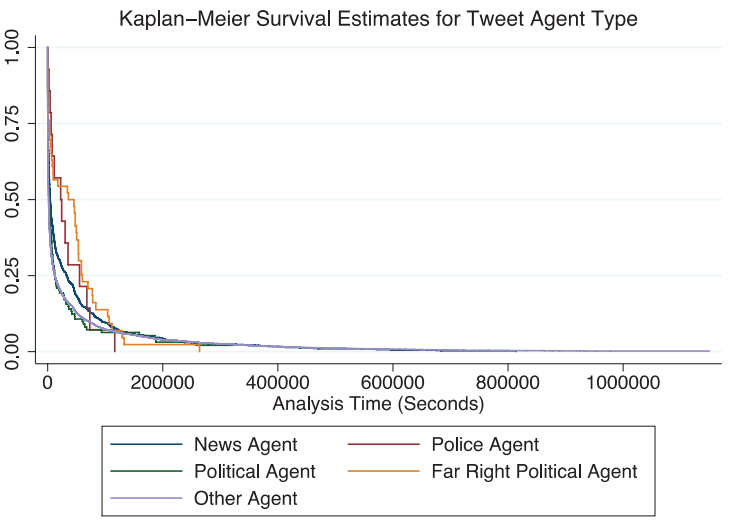

FIG. 6 Survival of Twitter information flows by tweeter type.

Discussion

In this paper, we evidenced how a fine-grained temporal analysis of locomotive social media data, supported by computational criminological methods, can reveal how an antecedent trigger terrorist event is related to an observable public social media reaction. We interpreted these social media communications via an application of Cohen's four-stage process of action and reaction. In support of hypotheses H1 and H2, via our bespoke hate speech supervised machine classifier (see Burnap and Williams 2015), we found evidence of cyberhate originating from individual Twitter users, in particular those identifying with right wing political groups, that was directly related to the trigger event. The correlation between the prevalence of far right political groups and individuals within social media networks and the production of cyberhate in the early stages post event, while unsurprising, is a novel empirical finding. These findings extend work that shows terrorists events result in a spike in *offline* hate crimes and incidents, by demonstrating cyberhate also spiked in the immediate aftermath of the Woolwich attack (Legewie 2013; Hanes and Machin 2014; King and Sutton 2014). The rapid spike in racial and religious cyberhate in the immediate aftermath of Woolwich is not surprising. In the second edition of moral panics, Cohen (1987) introduced the notion of 'symbols of trouble', and in the third edition, Cohen (2002) outlined clusters of social identity that he considered predictable symbols, including ethnic and religious minorities. He shows through socio-linguistic analysis that the press conflates these minorities with violence. Therefore, we might argue immigration, race and religion have become a *warning sign* for the real, much deeper threat of terrorism. The fanning of the flames on social media by individuals identifying with right-wing political groups, bolstered by traditional press coverage (we saw the correlation of news headlines and cyberhate in our model), further promotes these spurious connections and the propagation of cyberhate. The religious and racial dimension of the targeted cyberhate is potentially problematic if we are to assume, as Cohen did, that the social response to events is partly responsible for deepening its impact or causing new events through the cyclical process of action, reaction and amplification. Given the force-amplifier effect of social media and the role it may play in solidifying the crowd and validating sentiments and tensions, opening up a potential space for the galvanizing and spread of hostile beliefs, we next explored its possible contagion effect fuelled by 'rumours and the milling process' (1972: 19).

Having established that cyberhate did emerge in social media in response to the Woolwich event, we set out to investigate if it propagated via hypotheses H4 and H5. Our size and survival models allowed us to determine the escalation, duration, diffusion and de-escalation of cyberhate and non-cyberhate information flows during the impact, inventory and reaction stages following the terrorist trigger event. Our analysis revealed that while information flows containing cyberhate peaked in the impact stage following the event, a sharp decline was evident during the inventory stage. Our size and survival models confirmed that information flows containing cyberhate were significantly less likely to grow large and survive for long periods during the study window. These results lend support to the notion that terrorist events act as triggers for the production of cyberhate, possibly facilitated by 'rumours and the milling processs' (Cohen 1972: 19) that characterize the impact stage and that this type of event-specific cyberhate is relatively short term and conditional upon certain factors. The 'half-life' of

cyberhate mirrors research into *offline* event-related hate crimes and incidents (Legewie 2013; King and Sutton 2014). Perhaps the more salient of the conditional factors is the correlation between offline newspaper headlines and the propagation of information shown in all three models, supporting hypothesis H3. The number of newspaper headlines was predictive of the production of cyberhate, evidencing that tweeters posting hateful content may be fuelled by coverage in the press in the early impact stage, lending support to Cohen's argument above. Number of headlines was also predictive of the size of information flow, while News Agent type was predictive of survival of information flow (outlasting all but Other Agent in the analysis window). This study is the first to evidence the *prima facie* plausible relationship between new and old media in the context of social reactions to criminal events, and it lends support to the classic criminological notion that old media retains a significant role in 'setting the agenda' and 'transmitting the images' (Cohen 2002: xxiii) following crisis situations.

The dominance of actor type in all three models warrants further attention, particularly the ways in which information flows from these actors shape social reaction online. Despite the positive correlation between the number of press headlines and cyberhate in the early impact phase, we might postulate that the dominance of the traditional media effect in the inventory and reaction phases, as shown in the Model 3 and the KM on Agent Type estimation (Figure 6), is partly responsible for the reduction of cyberhate beyond the impact stage, counteracting its association with the production of cyberhate in the impact stage. It is clear from the WordCloud in Figure 6 that the two media stories on the Cameron and Blair speeches dominated Twitter activity during the latter stages of the study window, moving information flows onto wider contextual issues, in keeping with Cohen's reaction phase. It is therefore plausible that traditional media narratives in the inventory stage shape the preliminary picture of what has happened, closing down the possible spaces for the contagion of hateful and antagonistic sentiment fuelled by the unorganized response to death, speculation and rumour indicative of the impact stage. However, it is important to note that the media alone cannot be responsible for the reduction in cyberhate and that other agents are likely to play a part. In particular, information flows emanating from Police Agents within the first 24 hours of the study window showed high survival rates, above News Agents. As shown earlier, a portion of these police information flows contained case updates, helping dispel rumour and speculation. Social media affords Police Agents with a direct line of communication with citizens, bypassing the usual mass media filter—opening up possibilities to influence public opinion around events (Williams et al. 2013). However, the dominance of traditional media and police information flows during the impact and early inventory stage was accompanied by small (in terms of size as determined by retweeting volume) but sustained information flows emanating from far right political groups and individuals. The small but sustained nature of these flows indicates that there is limited endorsement of these twitter narratives, but where there is support it emanates from core group who seek out each other's messages over time. Therefore, contagion of cyberhate information flows is contained and unlikely to spread widely beyond such groups. Furthermore, preliminary analysis not covered here evidences the presence of counter-cyberhate speech following the terrorist event. These narratives from Twitter users either directly or indirectly challenge cyberhate. It remains to be seen if such self-regulation in social

media networks represents a form of responsibilization that can lighten the burden of policing the 'cyber-streets' (Williams et al. 2013; Giannasi 2014).

Conclusion

The connection between events and the production and propagation of cyberhate is mostly anecdotal. The rapid uptake of social media has resulted in a massive distributed social sensor net that affords criminologists with the opportunity to identify, monitor and trace social reactions to events to the second in real time. This unprecedented level of detail in data affords hate crime researchers with the ability to detect cyberhate in the aftermath of antecedent trigger events, such as the Woolwich terror attack much more rapidly than is achievable with conventional data. In this paper, we have shown the temporal variation in cyberhate that relates to concepts at the core of much criminological theory, such as the escalation, duration, diffusion and de-escalation of crime. If, as Cohen suggests, the social response to events is partly responsible for deepening its impact or causing new events through the cyclical process of action, reaction and amplification, then social media has the potential to act as a force amplifier. With respect to escalation, we showed how social media opens up a new digital 'public agora' for the mass production, consumption and spread of social reaction in relation to trigger events. Part of this event triggered 'amplified' social reaction is deviant in nature, and this study showed for the first time that the production of cyberhate is evident within the first few hours following the terrorist event. Drawing on Cohen's process of action, reaction and amplification, we identified this period as the 'impact stage', characterized by the unorganized response to death, rumour and speculation. These characteristics, accompanied by a 'terrestrial' paced police response to dispel speculation, open up a space for the initial production of unfettered prejudice and hate towards groups that are assumed to share similar characteristics with the suspected perpetrators. Our study of diffusion and duration evidenced that the spread of cyberhate was inhibited beyond the impact stage, with any sustenance likely due to the actions of a core group of far right political actors in the social media network. As in research into the temporal dimension of offline hate crime (Legewie 2013; King and Sutton 2014), it was apparent that cyberhate also had a 'half-life', evidenced by the rapid de-escalation post impact, and a near absence in the reaction stage. We postulated that dominance of traditional media and police information flows during the inventory and reaction stages, during which speculation gives way to facts about the case and finally to a focus on the wider issues, may be partly responsible for this half-life of cyberhate in social media networks. We also found initial evidence of counter-cyberhate speech in the data set, suggesting a form of responsibilization. However, confirming either of these postulations was beyond the scope of the data and this paper and future research should seek to explore these potential relationships further. If such causal associations do exists then the relationship between the media and the public during the course of reactions to events may be more variable than previously theorized.

Given the recent criminal justice response to cyberhate and incitement cases, our findings have several potential operational and policy implications. First, the ability to observe a large portion of the population[13] in near real time via social media networks provides those responsible for ensuring the safety of the public a new window onto mass social reactions. Evidence from this paper shows that deviant reactions, in the form of cyberhate in our case, can form part of a social reaction in relation to a trigger event. Therefore, these technologies may act as early warning systems for the amplification of deviance beyond the event itself. Second, the 'half-life' of cyberhate and its rapid de-escalation following the first 24 hours of the antecedent event suggests practitioners need to focus their interventions within this impact stage to increase the rate of de-escalation further. Third, the dominance of traditional media and police information flows in social media indicates these are likely effective channels for the countering of rumour, speculation and hate.

We end this paper with a methodological note. The majority of those currently under 20 years of age in the Western world were 'born digital' and will not recall a time without access to the Internet. Combined with the migration of the 'born analogue' generation onto the Internet, fuelled by the rise of social media, we have seen the exponential growth of online spaces for the mass sharing of opinions and sentiments, many of them characterized by volatility and a lack of regulation. These online spaces represent a socio-technical assemblage that creates a new public sphere enabling digital citizenship (Mossberger 2008) through which aspects of civil society are played out. No study of contemporary society can ignore this dimension of social life. Social media presents researchers with a rich new form of data from which criminologists, assisted by computational methods, can extract meaningful insights into contemporary social processes at unprecedented scale and speed. How we marshal these new forms of data is a key challenge for the social sciences. 'Computational criminology' may remain on the fringes of our discipline for some time yet, but we believe it is now worth considering the addition of 'computer science' to the list of disciplines with which criminology liaises to address contemporary forms of crime such as cyberhate.

Funding

This work was supported by the Economic and Social Research Council and Google Data Analytics Research Grant: 'Hate Speech and Social Media: Understanding Users, Networks and Information Flows' (grant number: ES/K008013/1) and the ESRC National Centre for Research Methods Grant: 'Social Media and Prediction: Crime Sensing, Data Integration and Statistical Modeling' (grant number: 512589112).

References

BBC (2013), Woolwich Murder Sparks Anti-Muslim Backlash, available online at http://www.bbc.co.uk/news/uk-22664835.

Braga, A., Papachristos, A. and Hureau, D. (2012), *Hot Spots Policing Effects on Crime.* Campbell Systematic Reviews.

[13] At the last count there were 15 million Twitter accounts in the United Kingdom (Rose 2014).

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*, 4: 223–42.

Burnap, P., Williams, M. L., Rana, O., Edwards, A., Avis, N., Morgan, J., Housley, W. and Sloan, L. (2015), 'Detecting Tension in Online Communities with Computational Twitter Analysis', *Technological Forecasting & Social Change*, 7: 96–108.

——. (2014), 'COSMOS: Towards an Integrated and Scalable Service for Analysing Social Media on Demand', *International Journal of Parallel, Emergent and Distributed Systems*.

Burnap, P., Williams, M. L. and Sloan, L. (2014), 'Tweeting the Terror: Modelling the Social Media Reaction to the Woolwich Terrorist Attack', *Social Network Analysis and Mining*, 4: 206, available online at http://link.springer.com/article/10.1007%2Fs13278-014-0206-4.

Cabinet Office (2013), European Council and Woolwich Incident: Prime Minister's Statement, Cabinet Office, available online at https://www.gov.uk/government/speeches/european-council-and-woolwich-prime-ministers-statement.

Cohen, S. (1972), *Folk Devils and Moral Panics*. MacGibbon and Kee Ltd.

——. (1987), *Folk Devils and Moral Panics*. MacGibbon and Kee Ltd.

——. (2002), *Folk Devils and Moral Panics*. MacGibbon and Kee Ltd.

Cox D. (1972), 'Regression Models and Life Tables', *Journal of Royal Statistical Society B*, 34: 187–220.

Davis, E. F., III, Alves, A. A. and Sklansky, D. A. (2014), 'Social Media and Police Leadership: Lessons from Boston', in *New Perspectives in Policing Bulletin*, 1–20. U.S. Department of Justice, National Institute of Justice.

Edwards, A., Housley, W., Williams, M. L. and Sloan, L. (2013), 'Digital Social Research, Social Media and the Sociological Imagination: Surrogacy, Augmentation and Re-orientation', *International Journal of Social Research Methodology* 16: 245–60.

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.

Feldman, M. and Littler, M. (2014), *Tell MAMA Reporting 2013/14 Anti-Muslim Overview, Analysis and 'Cumulative Extremism'*. Teesside University.

Giannasi, P. (2014), Hate on the Internet: Progress on the UK Government's Hate Crime Action Plan, presented at the All Wales Hate Crime Criminal Justice Board, British Transport Police, Cardiff, 23 July 2014.

Greenawalt, K. (1989), *Speech Crime & the Uses of Language*. Oxford University Press.

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.

King, R. D. and Sutton, G. M. (2014), 'High Times for Hate Crimes: Explaining the Temporal Clustering of Hate Motivated Offending', *Criminology*, 51: 871–94.

Leets, L. (2001), 'Responses to Internet Hate Sites: Is Speech Too Free in Cyberspace?', *Communication Law and Policy*, 6: 287–317.

Legewie, J. (2013), 'Terrorist Events and Attitudes Toward Immigrants: A Natural Experiment', *American Journal of Sociology*, 118: 1199–245.

Levin, B. (2002), 'Cyberhate: A Legal and Historical Analysis of Extremists' Use of Computer Networks in America', *American Behavioral Scientist*, 45: 958–88.

Marneffe, M., MacCartney, B. and Manning, C. D. (2006), 'Generating Typed Dependency Parses FromPhrase Structure Parses', paper presented at the International Conference on Language Resources and Evaluation (LREC), 24–26 May, Genoa, Italy.

Morell, G., Scott, S., McNeish, D. and Webster, S. (2011), *The August Riots in England*. NatCen.

Mossberger, K., Tolbert, C. J. and McNeal, R. S. (2008), *Digital Citizenship: The Internet, Society and Participation*. MIT Press.

Oksanen, A., Hawdon, J., Holkeri, E., Nasi, M. and Rasanen, P. (2014), 'Exposure to Online Hate among Young Social Media Users', in M. Nicole Warehime, ed., *Soul of Society: A Focus on the Lives of Children & Youth*, 253–73. Emerald.

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

Rose, K. (2014), The UK Social Media Landscape for 2014, available online at http://www.rosemcgrory.co.uk/2014/01/06/uk-social-media-statistics-for-2014/.

Sakaki, T., Okazaki, M. and Matsuo, Y. (2010), 'Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors', in *Proceedings of the 19th International Conference on World Wide Web*, 851–60. ACM Press.

Savage, M. and Burrows, R. (2007), 'The Coming Crisis of Empirical Sociology', *Sociology*, 41: 885–99.

Sloan, L., Morgan, J., Burnap, P. and Williams, M. L. (2015), 'Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data', *PLoS One*, available online at http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0115545.

Sloan, L., Morgan, J., Williams, M. L., Housley, W., Edwards, A., Burnap, P. and Rana, O. (2013), 'Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter', *Sociological Research Online*, 18: 7.

Smith, D. (2014), How Many People Use the Top Social Media? Digital Market Ramblings, available online at http://expandedramblings.com/index.php/r esource-how-many-people-use-the-top-soci al-media/. Accessed 14 August 2014.

Thelwall, M., Buckley, K., Paltogou, G., Cai, D. and Kappas, A. (2010), 'Sentiment Strength Detection in Short Informal Text, *Journal of the American Society for Information Science and Technology*, 61: 2544–58.

Thompson, D. (2013), 'There's a Problem within Islam, Says Tony Blair. He Should Know. He Helped Create It', *Telegraph*, 3 June 2015, available online at http://blogs.telegraph.co.uk/news/damianthompson/100219806/theres-a-problem-within-islam-says-tony-blair-he-should-know-he-helped-create-it/.

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

Wall, D. S. and Williams, M. (2007), 'Policing Diversity in the Digital Age: Maintaining Order in Virtual Communities', *Criminology and Criminal Justice*, 7: 391–415.

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

Williams, M. L., Burnap, P., Rana, O. F., Edwards, A., Housley, W. and Avis, N. (2011), *'Digital Social Research, Tension Indicators and Safer Communities: Introducing the Cardiff Online Social Media Observatory (COSMOS)'*. ESRC Digital Social Research Programme.

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

Williams, M. L. 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.

——. (2013), 'Cybercrime', in C. Hale, K. Hayward, A. Wahidin and E. Wincup, eds, *Criminology*, 3rd edn. Oxford University Press.

Yang, J. and Counts, S. (2010), 'Predicting the Speed, Scale, and Range of Information Diffusion in Twitter', International Conference on Weblogs and Social Media (ICWSM).

Zarrella, D. (2009), The Science of Retweets, available online at http://danzarrella.com/science-of-retweets.pdf.

Appendix: Computational Methods Used to Derive the Extremity of Cyberhate Variable

To classify the extremity of cyberhate from the collected Twitter data, we built a supervised machine learning classifier in the Weka tool to distinguish between hateful and/or antagonistic responses with a focus on race, ethnicity, religion and more general responses, following the event. To validate the machine classifier, we established a gold standard data set of human coded annotations. We sampled 2,000 tweets to be human coded and coders were provided with each tweet and the question: 'is this text offensive or antagonistic in terms of race ethnicity or religion?' They were presented with a ternary set of classes—yes, no, undecided. We utilized the CrowdFlower online service that allows for Human Intelligence Tasks, such as coding text into classes, to be distributed over multiple workers. We implemented the Stanford Lexical Parser (Marneffe et al. 2006), along with a context-free lexical parsing model, to extract typed dependencies within the tweet text. Typed dependencies provide a representation of syntactic grammatical relationships in a sentence (or tweet in this case) that can be used as features for classification. A ten-fold cross-validation approach was used to train and test the supervised machine learning method. It functions by iteratively training the classifier with features from 90 per cent of the human coded data set and classifying the remaining 10 per cent as 'unseen' data, based on the features evident in the cases it has encountered in the training data. Validation results suggested that overall the most efficient features for classifying cyberhate were n-gram typed dependencies combined with n-gram hateful and antagonistic terms. In fact, the hateful terms alone achieved the same precision performance but had a lower performance for recall. The number of false negative results (missed instances of cyber hate) was 7 per cent higher when using hateful terms alone. This is an interesting result as it provides evidence to suggest that human annotators identify hateful or antagonistic content on Twitter that does not necessarily contain hateful or antagonistic terms and requires a more nuanced representation of what is deemed cyber hate when aiming to classify tweets. For further details on the machine classification results and a full evaluation of the process, please see the open access article here: http://onlinelibrary.wiley.com/doi/10.1002/poi3.85/epdf.