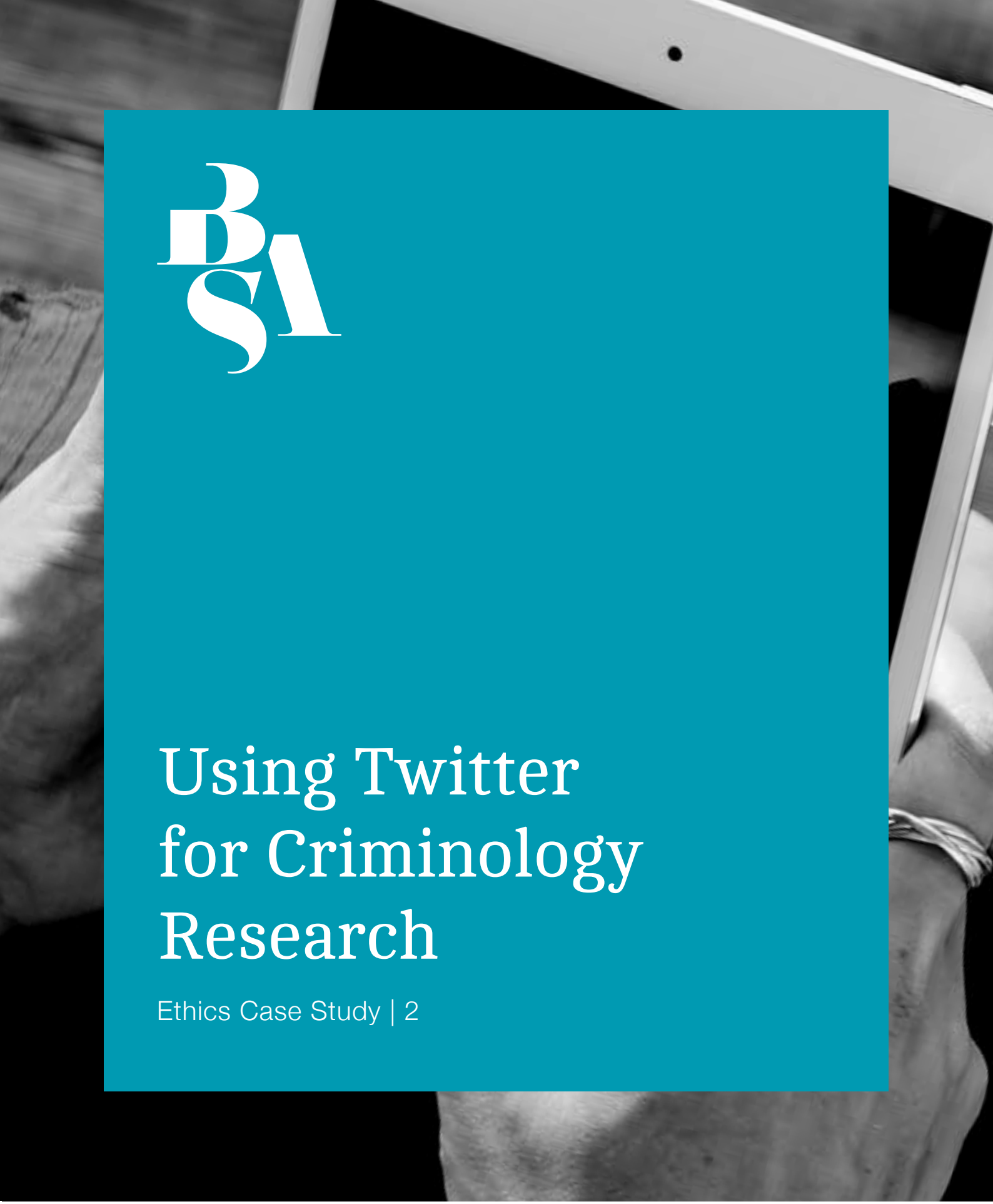

Using Twitter for Criminology Research

Ethics Case Study | 2

www.britsoc.co.uk

BRITISH SOCIOLOGICAL ASSOCIATION

Using Twitter for Criminology Research

Matthew L Williams and Pete Burnap
Directors, Social Data Science Lab, Cardiff University

A principal ethical consideration in most learned society guidelines on digital social research is to ensure the maximum benefit from findings whilst minimizing the risk of actual or potential harm (interpreted as physical or psychological harm, including discomfort, stress and reputational risk).

All groups involved in the research, including social media users, commercial platforms and researchers, should be protected throughout the lifecycle of the project, from inception to data archiving. Users are often the primary concern given their vulnerability in the process. Potential for harm in social media research increases when sensitive data are collected.

These data include personal demographic information (such as ethnicity and sexual orientation), information on associations (such as memberships to particular groups or links to other individuals known to belong to such groups) and communications of an overly personal or harmful nature (such as details on morally ambiguous or illegal activity and expressions of extreme opinion). These forms of sensitive information abound on social media networks. In some cases such information is knowingly placed online (whether or not the user is fully aware of who has access to this information).

In other cases sensitive information is not knowingly created by users – this can often occur in cases of association between users (not everything can be known about another user before connecting, nor can changes in affiliation be monitored on a routine basis). This information can come to light through the process of analysis, visualization (of networks) and representation of social media data by researchers (Rupert 2015).

Most social media research projects are likely to encounter only the first type of sensitive information. This is certainly the case where topics focus on mundane social activities online. However, those projects that take as their focus behaviours that have been deemed problematic risk encountering multiple forms of sensitive information.

Projects that take as their focus behaviours that have been deemed problematic risk encountering multiple forms of sensitive information.

Recent RCUK and government funded projects that have taken as their focus cyberhate following terrorist events (Burnap et al. 2014, Williams & Burnap 2015, Burnap & Williams 2015, Burnap & Williams 2016), the spread of racial tension online (Burnap et al 2015), the estimating offline crime patterns using online signals (Williams & Burnap 2016) and suicidal ideation (Scourfield et al. 2016) have encountered all forms of sensitive information outlined above. Here we take the example of cyberhate (Burnap et al. 2014, Williams & Burnap 2015, Burnap & Williams 2015, Burnap & Williams 2016) and provide an overview of our ethical decision-making process in sensitive social media research. The motivation for the ESRC and Google funded project stemmed from the increasing use of social media to communicate highly emotive reactions to events, such as terrorist attacks.

The project's objectives were to i) monitor hateful responses on social media following a series of events; ii) profile hateful social media networks; iii) link hateful content with other data, such as Google search terms and offline press; iv) model hateful information flows to identify enabling and inhibiting factors; and v) study forms of counter speech. The project drew upon both computational and social science research techniques. We used the COSMOS platform[1] to collect and visualise Twitter reactions to the murder of Lee Rigby in Woolwich.

Our first ethical dilemma was therefore related to consent: (i) as researchers should we obtain consent from all users in the social media dataset? As our intention was to conduct only quantitative analysis and aggregate level visualization that retained the anonymity of users we were satisfied that the consent provided to Twitter in their Terms of Service satisfied our criteria for minimizing harm (see final paragraph for discussion of consent in qualitative social media research).

The next stage of the project required the use of machine learning algorithms to classify hateful content and to build networks of users. Automated text classification of social media content performs well when conducted on datasets around specific events. However, their accuracy decreases beyond the events around which they were developed due to changes in language use (Burnap & Williams 2015).

Social network graph algorithms operate differently from classification algorithms, but they are also open to misrepresentation if there are data quality issues (such as missing data due to poor operationalization of collection search terms). Reliance on algorithms presented the second ethical dilemma: (ii) how should researchers develop, use and reuse algorithm driven text classification and social network graph processes that have the consequence of labeling content and users as hateful (and in some cases potentially criminal)? Where text classification techniques are necessitated by the scale and speed of the data (e.g. classification can be performed as the data are collected in real-time), researchers must ensure the algorithm performs well (i.e. minimizing the number of false positives) for the event under study in terms of established text classification standards. [2] Furthermore, researchers have a responsibility to ensure the continuing effectiveness of the classification algorithm if there is an intention to use it beyond the event that led to its design.

High-profile failures of big data, such as the inability to predict the US housing bubble in 2008 and the spread of influenza across the United States using Google search terms, have resulted in many questioning the power and longevity of algorithms (Lazer et al. 2014). Algorithms therefore need to be routinely tested for effectiveness and may need to be 'refreshed' with new human input and training data if false positives are to be minimized, avoiding the mislabeling of content and users. Where social network graphs indicate users are associated with particular groups, which if made public may cause distress or reputational risk, researchers must question the quality of the data used to generate the association (as would be expected in all scientific reporting) and make careful decisions on whether to publish such content. Where such information is published, every effort must be made to maintain the anonymity of users in the graph, including efforts to reduce the likelihood of deductive disclosure (Stewart and Williams 2005).

Twitter Terms of Service forbid the anonymization of tweet content

Following on from text classification, statistical model building was utilized to predict hateful information propagation around the Woolwich terrorist attack. These models identified which factors, such as type of user, network capital, and type of language used (such as counter-speech) enabled and inhibited hateful information flows. This presented the third ethical dilemma: (iii) is the process of identifying factors that stem the spread of online hate speech a universally accepted goal? This may seem like a redundant question to citizens of many European countries, where some forms hate and antagonistic speech are criminalised, including the UK. However, in the US hate speech is not criminalized, and online communications are protected by the first amendment. Therefore, project funders that are located in the US (such as Google) may not wish to be associated with research that infringes upon such protections. The researcher therefore must use their moral compass to balance these jurisdictional prerogatives with the pursuit of scientific objectivity.

Representation of our findings presented the fourth ethical dilemma: (iv) is it possible to present the content of hateful and counter speech in tweets in publication? Anonymous publication of actual examples of hateful tweets is precluded under Twitter Terms of Service. 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 in most big data projects 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. Furthermore, Twitter Terms of Service also requires that authors honour any future changes to user content, including deletion. As academic papers cannot be edited continuously post publication, this condition further complicates direct quotation (needless to mention the burden of checking content changes on a regular basis). However, researchers should not conclude that conventional representation of qualitative data in social media research is precluded due to these Terms of Service.

As in conventional qualitative research, researchers can make efforts to gain informed consent from a limited number of posters if verbatim examples of text are required (although posters must understand that anonymity is not possible in these cases given tweet text is searchable). In cases where consent is not provided, Markham (2012) suggests some innovative methods for protecting privacy in qualitative social media research.

Acknowledging that traditional methods for protecting privacy by hiding or anonymizing data no longer suffice in digital settings that are archived and searchable, Markham advocates bricolage-style reconfiguration of original data that represents the intended meaning of interactions. While this may be suitable for general thematic analysis, it may not satisfy the needs of more fine-grained approaches, such as conversation and discourse analysis.

Social Data Science Lab Risk Assessment and Ethical Principles

Social research ethics are at the core of the Social Data Science Lab's programme of work. Recent work shows how users of social media platforms are uneasy about their posts being collected without their explicit consent (NatCen 2014, Williams 2015). However, many social media terms of service specifically state that users' data that are public will be made available to third parties, and by accepting these terms users legally consent to this. In the Lab's research programme we interpret and engage with these terms of service through the lens of social science research which often implies a higher ethical standard than provided in legal accounts of the permissible use of these kinds of data. The topic of ethics in social media research has been a key focus of ours and formed a primary research question in our first ESRC Digital Social Research Demonstrator Grant. Ethics as a topic continues to be embedded in our follow-on grants and we are continuously reflecting upon our practice as social and computational researchers. We are acutely aware of the key ethical issues of harm, informed consent, the invasion of privacy and deception as they relate to the collection, analysis, visualization and dissemination of social media data. Below we detail our risk assessment and ethical principles that have been adopted by several social science several research ethics committees in the UK.

Risk Assessment

Low Risk

Tweet is from official/institutional account:
Publish without seeking consent in most cases.

High Risk

Tweets are from individual users and contain mundane or sensitive information (overly personal, abusive etc.).
Must contact the user (direct message/@mention/email) and ask their permission to publish.
Only publish if consent is received.

High Risk

Tweet has been deleted precluding publication under Twitter Developer Agreement/Policy.

High Risk

Tweet is from a deleted account meaning it has been deleted precluding publication under Twitter Developer Agreement/Policy.

Ethical Principles

- We abide by the Economic and Social Research Council's Framework for Research Ethics
- All projects undergo Research Ethics Committee Review
- Any significant changes to research design following Research Ethics Review approval are reported back to the Committee for re-approval
- We abide by Twitter's Developer Policy and Developer Agreement
- We abide by the UK Data Protection Act 1998
- We only use social media data for academic research purposes
- We keep all information gathered on individual Twitter users confidential on secure password protected servers
- We maintain the anonymity of all individual Twitter users in our research
- We only publish in research outputs aggregate information based on data derived legally and ethically from the Twitter APIs
- In research outputs we never directly quote individual Twitter users without their informed consent. Where informed consent cannot be obtained, we represent the content of tweets in aggregate form (e.g. topic clustering, wordclouds) and themes (decontextualised examples and descriptions of the meaning or tone of tweet content). These forms of representation preclude the identification of individual Twitter users, preserving anonymity and confidentiality
- In research outputs we do directly quote from Twitter accounts maintained by public organisations (e.g. government departments, law enforcement, local authorities) without seeking prior informed consent
- We never share data gathered from Twitter APIs for our research outside of the COSMOS project team
- We destroy all personal data if it is no longer to be used for research purposes

Funding: This work was supported by five Economic and Social Research Council grants: 'Digital Social Research Tools, Tension Indicators and Safer Communities: a Demonstration of the Cardiff Online Social Media ObServatory (COSMOS)', Digital Social Research Demonstrator Programme (Grant Reference: ES/J009903/1), 'Hate Speech and Social Media: Understanding Users, Networks and Information Flows', Google Data Analytics Research Programme (Grant Reference: ES/K008013/1), 'Social Media and Prediction: Crime Sensing, Data Integration and Statistical Modeling', National Centre for Research Methods (Grant Reference: ES/F035098/1/512589112), 'Digital Wildfire: (Mis) information Flows, Propagation and Responsible Governance', Global Uncertainties Ethics and Rights in Security Programme (Grant Reference: ES/L013398/1), and 'Public Perceptions of the UK Food System: Public Understanding and Engagement, and the Impact of Crises and Scares', Understanding the Challenges of the Food System Programme (Grant Reference: ES/M003329/1).

References

Burnap, P. Williams, M. L. & Sloan, L. (2014) Tweeting the Terror: Modelling the Social Media Reaction to the Woolwich Terrorist Attack, *Social Network Analysis and Mining*, 4: 206.

Burnap, P. & 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(2):* 223–242.

Burnap, P. & Williams, M. L. (2016) Us and Them: Identifying Cyber Hate on Twitter Across Multiple Protected Characteristics', *EPJ Data Science* 5(11).

Burnap, P., Williams, M. L. Rana, O., Edwards, A., et al. (2013) Detecting Tension in Online Communities with Computational Twitter Analysis, *Technological Forecasting & Social Change, 95:* 96-108.

Lazer, D., Kennedy, R., King, G. & Vespignani, A. (2014), The Parable of Google Flu: Traps in Big Data Analysis, *Science*, 343: 1203–5.

Markham, A. (2012) Fabrication as Ethical Practice: Qualitative Inquiry in Ambiguous Internet Contexts, *Information, Communication and Society*, 15(3): 334-353.

NatCen (2014) *Research Using Social Media: Users' Views*, London: Natcen.

Ruppert, E. (2015) Who Owns Big Data?, *Discover Society*, 23.

Scourfield, J.B., Colombo, G., Burnap, P., Jacob, N.K., et al. (2016) The Response in Twitter to an Assisted Suicide in a Television Soap Opera. *Crisis: The Journal of Crisis Intervention and Suicide Prevention,* 37(5): 392-395.

Stewart, K. F. & Williams, M. L. (2005) Researching Online Populations: The Use of Online Focus Groups for Social Research, *Qualitative Research* 5(4): 395-416.

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

Williams, M.L, Burnap, P. & Sloan, L. (2017) Crime Sensing With Big Data: The Affordances and Limitations of Using Open-source Communications to Estimate Crime Patterns, *British Journal of Criminology,* 57(2): 320–340.

Williams, M. L. (2015), Towards an Ethical Framework for Using Social Media Data in Social Research, presented at Social Research Association Workshop, Institute of Education, UCL, 15 June 2015.

Williams, M. L. and Burnap, P. (2015) Cyberhate on Social Media in the Aftermath of Woolwich: A Case Study in Computational Criminology and Big Data *British Journal of Criminology* 56(2): 211-238.

Williams, M. L., Edwards, A., Housley, W., Burnap, P., et al. (2013) Policing cyber-neighbourhoods: Tension monitoring and social media networks, *Policing and Society* 23(4): 461-481.

[1] http://socialdatalab.net/software

[2] Established measures include: precision (the fraction of retrieved tweets that are relevant to the search – i.e. for each class how many of the retrieved tweets were of that class); recall (fraction of tweets that are relevant to the search that are successfully retrieved – i.e. for each class how many tweets coded as that class were retrieved); F-Measure (a harmonized mean of precision and recall); and Accuracy (the total correctly classified tweets normalized by the total number of tweets). Results of 0.75 and above (on a scale of 0-1)s in each measure are considered outstanding (van Rijsbergen, 1979).