The ethical challenges of publishing Twitter data for research dissemination

Helena Webb
Marina Jirotka
University of Oxford, Department of Computer Science
Oxford, United Kingdom
helena.webb@cs.ox.ac.uk
marina.jirotka.cs.ox.ac.uk

Bernd Carsten Stahl
De Montfort University, Department of Informatics
Leicester, United Kingdom
bstahl@dmu.ac.uk

William Housley
Adam Edwards
Matthew Williams
Cardiff University, School of Social Sciences
Cardiff, United Kingdom
housleyw@cardiff.ac.uk
edwardsa2@cardiff.ac.uk
williamsm7@cardiff.ac.uk

Rob Procter*
University of Warwick, Department of Computer Science
Coventry, United Kingdom
rob.procter@warwick.ac.uk
*Corresponding author

Omer Rana
Pete Burnap
Cardiff University, School of Computer Science and Informatics
Cardiff, United Kingdom
ranaof@cardiff.ac.uk
p.burnap@cs.cardiff.ac.uk

ABSTRACT
Empirical research involving the analysis of Internet-based data raises a number of ethical challenges. One instance of this is the analysis of Twitter data, in particular when specific tweets are reproduced for the purposes of dissemination. Although Twitter is an open platform it is possible to question whether this provides a sufficient ethical justification to collect, analyse and reproduce tweets for the purposes of research or whether it is necessary to also undertake specific informed consent procedures. This paper reports on an ethics consultation that formed part of a wider research study and that aimed to identify best practice procedures for the publication of Twitter data in research findings. We focus largely on the UK context and draw on the outcomes of the consultation to highlight the range and depth of ethical issues that arise in this area. We can see Twitter as a case study for a wide number of data sources used in Web Science. This is a highly complex landscape in which questions crystallise around fundamental principles such as informed consent, anonymisation and the minimisation of harm. Furthermore, tensions exist between commercial, regulatory and academic practices, and there are also circumstances in which good ethical practice might compromise academic integrity. There is an absence of consensus in Web science and related fields over how to resolve these issues and we argue that constructive debate is necessary in order to take a proactive approach towards good practice.

WebSci '17, June 25-28, 2017, Troy, NY, USA

ACM ISBN 978-1-4503-4896-6/17/06.
http://dx.doi.org/10.1145/3091478.3091489

Categories and Subject Descriptors
K.4 [**Computers and Society**]: Public and Policy Issues – *abuse and crime involving computers, ethics, regulation*

General Terms
Management, Human Factors, Legal Aspects.

Keywords
Research ethics, Twitter, social media, informed consent

1. INTRODUCTION
Twitter provides a highly popular data source in Web science research. It is easy to understand why this is the case. Twitter is a widely used social media platform across the world and it is relatively easy for researchers to collect data from it. As an open platform the majority of posts are available to public view and researchers can collect large numbers of tweets in a very short period of time via the platform's Application Programming Interface (API). Much existing work draws upon large scale quantitative approaches which present aggregated findings to discuss, for instance, voting intentions [1], the propagation of content [2], and the expression of sentiment and tension in particular social contexts [3]. Twitter data have also been used for smaller scale qualitative analyses [4], which often use linguistic or discourse analytic approaches to examine how posts are constructed and how the messages within them are conveyed.

Research in the field has already highlighted some of the ethical issues involved in this kind of research – much of it in terms of social media data as an instance of Big Data [5]. These important discussions focus on concerns arising from the use of automated processes to collect large volumes of data and what implications this has for participant recruitment, privacy and identification etc. Discussions have also focused on the appropriate handling and

archiving of such data [6]. Suggestions for what might constitute good ethical practice in these areas vary, often in relation to the different levels of risk associated with different data sources (e.g. private vs public internet platforms) but also as a reflection of different ethical positions [7].

In our own research project, we encountered a number of challenges relating to the responsible handling of Twitter data. Due to the specific nature of our project focus and methodological approach we identified questions that had not yet been addressed in full in the existing literature. These related specifically to questions over the publication of individual tweets in research dissemination and whether, in order to publish them (in an anonymised or non-anonymised form), it was first necessary to contact the original users who posted them to solicit their informed consent. We conducted a consultation to overview current guidance and expert opinion on these matters. In this paper we report and discuss the findings of that consultation. We show that the publication of Twitter data in research raises significant questions for research ethics and that as yet there is an absence of consensus in the field over how to resolve them. We argue that the challenges posed relate to fundamental ethical principles such as the minimisation of harm and the value of informed consent. These challenges also create tensions between commercial (relating in particular to the Twitter platform and the use of tweets by media organisations), regulatory and academic practice. Finally, there are also occasions where suggestions for good practice might be seen to compromise academic integrity. We argue that these tensions identified in our consultation are relevant not only to the publication of Twitter data but also to a range of other ongoing research activities in Web science. It is therefore necessary for open and constructive debate to take place so that researchers can discuss these issues in full and the field of Web science can take a proactive approach towards ethical practice.

2. BACKGROUND: THE DIGITAL WILDFIRE PROJECT

The recently completed 'Digital Wildfire' project[1] was an interdisciplinary research study that sought to identify opportunities for the responsible governance of digital social spaces. The project was funded by the UK's Economic and Social Research Council (ESRC) and was a collaboration between social scientists, computer scientists and computer ethicists. The background to the project lies in the contemporary popularity of social media platforms and the capacity for digital content to spread on a broad and rapid scale [8]. Where rapidly spreading content is in some way inflammatory, antagonistic or provocative – for instance in the form of rumour, hate speech, malicious campaigns etc. – it can risk causing serious harms to individuals, groups and entire communities. The Digital Wildfire project team undertook a range of research activities designed to further understand how these kinds of content propagate across social media, the consequences they have, and how responsible governance strategies might limit the spread of content without impeding freedom of speech [9]. As part of this work we used Twitter data as a case study to investigate the posting and spread of cyber hate on social media.

We collected tweets via the Twitter API and collated conversational 'threads'. Each thread began with a post that might be considered as an instance of cyber hate – that is, an antagonistic post targeted at an individual or group based on personal attributes. These were identified through the accounts of well-known inflammatory posters in the public eye and via sentinel sites such as #YesYou'reRacist and #YesYou'reSexist that serve to collate and expose inappropriate content posted by others. Subsequent posts responding to the opening tweet were also collected and arranged in posting order. This enabled us to observe and analyse how Twitter users respond to each other's posts.

In our analysis we were interested in 1) how users on Twitter construct posts that are treated by others as hateful and 2) how users construct responses to an opening post to perform interactional actions such as expressing agreement or disagreement with it. To give a brief example, Figure 1 shows an opening post tweeted by Katie Hopkins – a public figure in the UK well-known for the expression of inflammatory opinions[2], often targeted towards certain groups of people– and two responses that followed it[3].

Figure 1. Opening tweet and two responses

We can observe that the opening post draws on specific rhetorical devices such as short sentences, the use of emotive categories – e.g. babies – and the presentation of (extreme) opinion as fact – to construct a message that can be seen as inflammatory and likely to provoke responses from others. We can further observe how subsequent posts draw on interactional resources and the functionality of the Twitter platform to produce disagreeing replies that both address Katie Hopkins directly – e.g. through the use of the @handle and terms such as 'what gives you the right' - whilst also producing negative assessments and directives that are designed to be viewable to other users – e.g. #blockKT. This kind of granular, qualitative analysis of individual posts helped us to understand in detail how cyber hate is posted and responded to on

[1] www.digitalwildfire.org

[2] https://www.theguardian.com/media/katie-hopkins

[3] For the purposes of this paper, we have used the following criteria to determine whether consent is needed in order to reproduce Twitter posts: posts made by figures in the public eye who tweet as part of their public role, are reproduced in full without consent being sought. Other users – such as those who posted tweets 2 and 3 were contacted via Twitter and their opt in consent was sought to include their posts in our project publications.

the Twitter platform, and subsequently to conceptualise what forms counter speech – posts serving to push back against the spread of hateful content – might take. This analysis informed statistical modelling work that examined the impact of the occurrence of counter speech on the length of threads that were started after an initial cyber hate post. It also produced valuable insights in its own right that we wanted to disseminate via written publications and conference presentations.

The preference amongst some members of the project team was that this dissemination of our qualitative findings should include the reproduction of certain Twitter posts in order to illustrate the wider patterns found across the dataset. The use of data in this way is standard in the methodological approaches drawn on in the conduct of the analysis – ethnomethodology and conversation analysis (EMCA), membership categorisation analysis (MCA) and interaction analysis [10]. These approaches all emphasise the collection of naturally occurring data for analysis and the (anonymised) reproduction of data to enable audiences to assess for themselves the validity of that analysis. Much work conducted using these approaches is based on the examination of audio or video recorded face-to-face interactions. Anonymisation of this data can be relatively easy to achieve, through the removal of real names in transcripts and the blurring of faces etc., and the capacity for individual participants to be identified is arguably low – particularly in comparison to social media data gathered from a 'public' platform such as Twitter.

As we made plans to disseminate our qualitative analysis, debate arose amongst the team over whether it would be ethically appropriate to reproduce tweets in publications, even if anonymised, and whether we should perhaps contact individual users to seek their informed consent before doing so. Concerns that publication might go against good practice were crystallised into three areas:

1. **Covering up usernames and @handles does not create meaningful anonymisation as it is sometimes possible to enter the main text of a tweet into Twitter's search function, recover the tweet and its associated meta-data, including the username and @handle of the user who posted it. Though Twitter allows users to post under a pseudonym, the meta-data (including picture, location etc.) may enable the poster's identity to be discovered.**
2. **Given that the Digital Wildfire project includes the examination of hate speech, there is concern that users may come to harm if it is possible to identify them and they have been posting content considered to be hateful, inflammatory etc.**
3. **Twitter's User Development policy requires any reproduction of tweets to be done in full, so anonymisation procedures are in breach of that.**

To help us examine these areas more fully, and ideally to find ways to resolve the challenges we faced, we decided to run a consultation exercise to scope existing guidance and practice in this area.

3. ETHICS CONSULTATION

Our project team agreed to undertake a consultation to review the three issues listed above and also to seek answers to a more global question:

In order to publish (anonymised) individual social media posts is it first ethically necessary to contact the user and solicit their informed consent?

In the conduct of the consultation we carried out three strands of activity:

1) *Scoping of current relevant guidance*. We surveyed current regulatory guidance on the use of social media data in research. This included guidance arising from individual academic institutions, research funding bodies and other research and regulatory organisations.
2) *Survey of expert opinion.* We made contact with a number of experts working in the fields of research ethics and computer ethics. We asked them for their thoughts on the questions and challenges covered in the consultation. We also reviewed available relevant literature on the ethics of social media/internet research.
3) *Survey of current practice and opinion* We surveyed the opinion of various individuals working in the field to identify their views on the consultation questions and their current practices regarding the use of social media data. This survey was conducted in a number of ways including – reviews of the content of relevant journals; email conversations with the editors of a number of research journals; email and face-to-face conversations with researchers; and group discussions with researchers at relevant events, such as the Social Media and Society conference in London 2016.

The findings of the consultation were compiled into a dossier which was then used at the basis for further discussion. In this paper we draw on the outcomes of the consultation to discuss the challenges faced by researchers when trying to determine what equates to best practice when working with Twitter data. We found an absence of consensus across the field that reflected the disagreements occurring in our own research team. Perspectives vary amongst individuals but also across institutions. Furthermore, research using Twitter data creates a complex landscape in which various ethical challenges arise. These relate in particular to key ethical criteria such as informed consent, minimising harm and anonymisation. Challenges also arise through tensions between academic, commercial and regulatory practice as well as in contradictory criteria for behaving ethically and upholding academic integrity.

4. THE ETHICAL CHALLENGES OF PUBLISHING TWITTER DATA

4.1 Absence of academic consensus

As noted above it is standard practice in Web science that tweets are collected for analysis via the platform's API. This can be done without users being aware at all that their tweets are being collected. Users are typically not approached directly to solicit their informed consent to take part in research, instead consent is often assumed to have been given by the user's acceptance of Twitter's Terms of Service. These state[4]:

"By submitting, posting or displaying Content on or through the Services, you grant us a worldwide, non-exclusive, royalty-free license (with the right to sublicense) to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute such Content in any and all media or distribution methods (now known or later developed). This license authorizes us to make your

[4] https://twitter.com/tos?lang=en#us

Content available to the rest of the world and to let others do the same".

Regarding the further handling of this data, including the publication of specific posts, there is an absence of consensus. In the US a 2015 amendment to the 'Common Rule' Federal Policy[5] for the Protection of Human Subjects suggests that certain forms of online behaviours can be classed as public behaviour and therefore do not require further ethical review.

"Any research involving standardized testing, surveys, interviews, or observations, including audio and video recording, of public behavior, including behavior online, will be able to proceed without further review".

This amendment is a significant development as the 'Common Rule' tends to determine the decision-making conducted by Institutional Review Boards (IRBs). By contrast in the UK, the ESRC, which funds the Digital Wildfire project, recommends the full ethics review of projects intending to collect social media data, noting potential tensions regarding traceability and what might or might not considered public.

"The potential for identifiabilty of online sources, as well as ethical debates about how privacy is constituted in digital contexts, means that full ethics review may be appropriate for research involving these communities".

Similarly, other ethics guidelines, such as those published by the British Psychological Association [11] and the Association of Internet Researchers [12] – recommend careful consideration of ethical issues when using social media data with particular regard to privacy. These guidelines do not take an overt stance on the matter of consent for publication. On the other hand, a 2016 output produced by the University of Aberdeen [13] and based on project work funded by the ESRC advocates that in the case of sensitive social media content, researchers should either consider the use of paraphrased/composite data instead of reproducing actual posts or use an informed consent approach. Similarly, in 2016 the University of Oxford produced updated guidance[6] on internet-based research. This advises that in the case of Twitter (as a public platform) researchers do not need to solicit consent to collect data but should seek consent to publish individual posts. Alternatively, they can create composite data for the purpose of publication.

"Researchers who wish to display direct quotes and the username and picture of the person in their work (especially if it is published in any way) should normally seek informed consent to do this, especially in cases of very sensitive data (e.g. hate speech). They should contact the participants directly having decided which consent procedure should be followed (e.g. online information sheet, online consent form, click boxes, etc.). If gaining informed consent is not possible, quotes should normally be paraphrased and usernames/pictures de-identified in order to protect the 'participants'".

The lack of consensus in research guidance is mirrored by an absence of consensus in the research community. In our scoping of existing literature and survey of individual researchers we identified a range of opinion based on alternative ethical positions and underlying assumptions about the status of the Twitter platform. This includes different positions regarding whether any consent to publish should be opt-in or opt-out, whether published posts should be anonymised and whether composite data is an acceptable alternative to publishing real posts. We also came across a high number of researchers working in this field and determined to follow good practice who were genuinely uncertain what form that good practice should take. We can see this current absence of consensus as a genuine barrier to the conduct of work in this area with particular obstacles occurring when research collaborations involve team members from institutions whose guidance is not compatible.

4.2 Informed consent

The principle of informed consent is a cornerstone of ethical guidance in contemporary research involving human participants. In its classic form, derived from research in bio-medicine [14], it requires informed consent to be given by participants at the point of data collection. In the case of research involving the collection of Twitter data, informed consent for data collection is typically based on user acceptance of its Terms of Service. Given the large volume of tweets typically collected in a Twitter-based study, it would certainly be time consuming and challenging to attempt to contact all users in a more direct way. However, survey research[7] suggests that social media users are unlikely to read or remember the full Terms and Conditions (T&Cs) of the platforms they sign up to – undermining the assumption that informed consent for data collection has been given. Furthermore, research on public opinion suggests individuals are wary of research that collects and publishes social media data without more overt user consent. For instance, research reported by the UK think tank Demos [15] found a low level of awareness amongst members of the public that their posts might be used for research purposes and a general concern over the implications of research for privacy and the risk of harm. However, in comparison to other platforms, respondents were less worried about the collection of Twitter data for research due to public nature of the platform.

We have found some instances of small scale studies involving Twitter data that overtly seek informed consent from users at the point of data collection. For instance, as part of a study on the responsible collection of social media data Moffat and Koene (2016) [16] made use of a prototype web tool. This tool allowed users to monitor and manage their Twitter interactions whilst also enabling them to determine when their posts were being collected as research data. In order to assess for ourselves how practical a 'full' informed consent procedure might be, the project team agreed to attempt to contact a number of Twitter users and seek their opt-in consent to publish specific tweets. Having identified a data thread of interest we contacted individual users over Twitter via a reply to the specific tweet we wanted to publish. Due to Twitter's character limit, the contact request was split over two tweets and a link was provided to access further information:

[5] http://www.hhs.gov/ohrp/regulations-and-policy/regulations/nprm-home/index.html

[6] http://www.admin.ox.ac.uk/media/global/wwwadminoxacuk/localsites/curec/documents/BPG_06_Internet-Based_Research.pdf

[7] https://www.theguardian.com/commentisfree/2014/apr/24/terms-and-conditions-online-small-print-information

Hi we are researchers studying social media communications. We are looking at 'heated' discussions on Twitter (1 of 2)
We would like to use this tweet in our publications. Please see https://sites.google.com/site/digitalwildfireesrc/home/consent-for-publication and reply if this is OK. (2 of 2)

Over 20 requests to different users were sent out and two responses were received – both giving consent for publication. The process of sending out requests was very quick and it was possible to send many messages in a short period of time. As none of the users involved followed our Twitter account we were unable to send direct (private) messages. This meant our requests were visible to other users on the platform, arguably drawing attention to the post being referred to. When we did not hear a response from users we had contacted, we sent them the same request again around a week later in case they had missed it the first time. We did not feel comfortable sending more than two requests as we did not want our contact to appear intrusive[8].

Our trial with informed consent was based on an opt-in model. An alternative format is an opt-out model, in which users need to specifically state that they do not want to be involved in the research. This model relieves some of the time burden involved in securing consent but could arguably provide a means to ensure good practice – particularly for instance where the social media posts involved are not sensitive or very personal in content. However, many Twitter uses receive a high volume of messages in their feed, or perhaps go for long periods of time without logging on to the platform. So opt-out consent does not provide a guarantee that a request to publish (if sent via a Twitter message) has been seen or understood.

More fundamentally, the often anonymous nature of social media challenges the informed consent model. Twitter does not have a real name policy and users are easily able to set up profiles that do not include their name or any 'real' facts about them. Even with opt-in consent we cannot be totally sure of the identity of the users involved if the only access we have to them is via their public user profiles. In the case of the users we contacted in our own trial, we checked their profiles to identify available details about their age, job etc. We wanted to ensure that they met the criteria of adults capable of giving informed consent. Even though their user profiles indicated that they were, we cannot be totally certain that these details are accurate. As in the case of any informed consent procedure reliant on public profiles and conversations over a social media platform, we cannot necessarily rule out the possibility that a user is under the age of 16 or in some way vulnerable and therefore someone who might typically be regarded as unable to give consent.

4.3 Minimising harm

One of the main causes for concern we identified over the use of individual tweets in research dissemination is that publication of posts might cause harm. Even though Twitter is an open platform, users do not necessarily expect that their tweets will be collected and published to other audiences elsewhere. Publishing tweets can bring users to new or larger forms of attention and if their identities are revealed, they might be at risk from their (potentially unknowing) participation in research.

The minimisation of harm is another key ethical principle in research involving human participants. This was highly relevant to the Digital Wildfire project as our focus was on inflammatory posts such as messages containing cyber hate. It is possible that the victims of a cyber hate post might be identifiable (to themselves and others) when an individual post is reproduced in a publication; this might cause them further harm in addition to the harm they experienced when the content was originally posted. Furthermore, if publication makes identifiable a user who has posted cyber hate messages, this also risks harm. Readers of the publication might seek to retaliate by sending accusatory or 'shaming' messages over social media or even identifying the user offline and seeking to report him/her to the police, contacting his/her employer etc. Although we have not come across any instances in which this kind of harm has occurred as a direct result of research involving social media data, practices of online shaming or 'digilantism' [17] are frequent across contemporary society and there have been numerous cases in which users have been excessively punished for tweets that were perhaps poorly worded or published in the heat of the moment and quickly regretted.

Whilst minimising harm is a central component of good research practice, almost all forms of research carry at least some risk of causing harm. It is therefore necessary to attempt to assess the level of risk involved in individual projects. As part of our consultation exercise we conducted a risk assessment of the Digital Wildfire project's work involving the analysis of social media posts. Whilst noting the inflammatory nature of many of the posts we were analysing we also identified that large numbers of the posts we had collected were non-contentious or mundane in content. We also noted that some users do not include any kind of identifying detail (name, image, place of work etc.) in their profile or tweets. Subsequently we devised an assessment matrix in which risk to users was calculated according to two dimensions. These were: 1) the extent to which the user is identifiable and 2) the extent to which the content of the tweet is antagonistic, provocative etc. This matrix is shown in Figure 2.

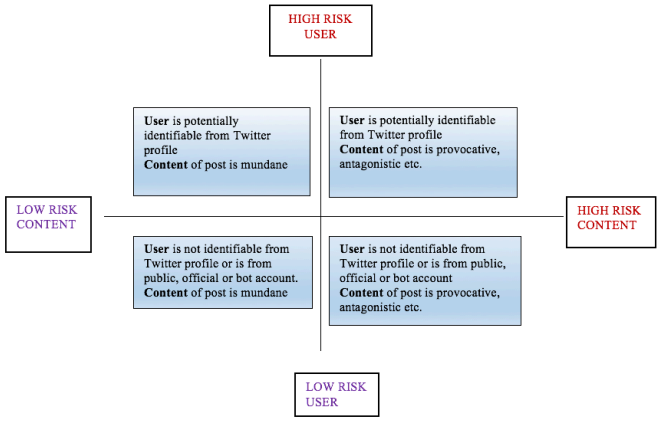

Figure 2. Proposed risk assessment matrix for tweets

At first our risk matrix appeared to provide a useful means to assess the likelihood of harm arising from the publication of individual

[8] One of the anonymous reviewers of this paper also made the helpful suggestion that we could have deleted our tweets sent to users after a certain period of time to further limit their potential obtrusiveness.

twitter posts and we considered adopting alternative consent strategies consent (no consent, opt-out, opt-in) in relation to the different levels of risk identified. However, on further reflection, this model became problematic in some ways. It was noted that even if a user's offline identity is unknown, their online one is likely to be identifiable and this leaves open the possibility that publication can cause them some kind of harm through online messages or 'shaming' etc. Furthermore, it was observed that if readers look up a user's tweets when published, they will also have access to that user's entire profile and (public) posting history – including posts made after publication. Even if a publication quotes only mundane posts from a user, it is possible that readers might find more provocative ones that were not identified in the study or posted after the publication of the study. Similarly, users with non-identifying profiles at the point of publication may change this at some point in future so that their personal details can be seen by readers.

The dynamic nature of Twitter means that it is very difficult to use a static model to assess the extent to which publishing an individual user's posts risks causing them harm. As researchers in this area we are challenged to ask where our responsibilities towards participants begin and end: do we have to protect them from harm both in the present and in the possible future? As discussed next, the principle of minimsing harm is further complicated by the extreme difficulty of ensuring the meaningful anonymisation of individual Twitter posts.

4.4 Anonymisation

It is often standard practice in research to anonymise data for publication. Participants' identities are protected to ensure that they cannot be identified by dissemination audiences and this therefore helps to protect them from harm. Research outputs including individual tweets sometimes attempt to anonymise them by covering or altering the username and @handle of the poster. However, due to the open status of the Twitter platform it can often possible to identify the user with relative ease.

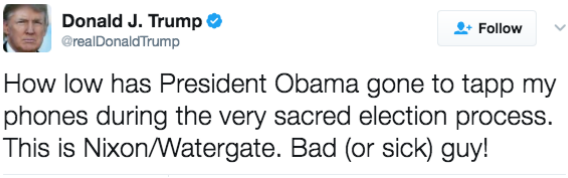

Figure 3: recent tweet posted by Donald Trump

We can take as an example a recent tweet posted by Donald Trump – see Figure 3. Even if we do not know the user name or @handle of the poster, we can put the content of the post into Twitter's search function and it will show us the original post as a result – revealing the user's online identity. To overcome this, we might try to make small amendments to the content of the post. Correcting spelling errors or removing or substituting words might (arguably) not affect the integrity of the analysis but provide a means to shroud the identity of the poster. So, we might amend the post in question to:

How low has Obama gone to tap my phones during the very special election process. This is Nixon/Watergate. Bad or sick guy!

However, putting this content into Twitter's Advanced Search produces a small number of results, including the original tweet and tweets that quote it. This also occurs when entire sentences are removed from the original post. Once again the identity of the user can be easily deduced. Although we have used a high profile user as demonstration, we have carried out this exercise with some of the content posted by unknown users in our datasets and achieved similar results. Twitter's Advanced Search function can consistently identify the profile of a user or narrow it down to a very small subset and further investigation can quickly reveal the relevant user.

4.5 Commercial vs regulatory vs academic practice

Incompatibilities between commercial, regulatory and academic practices present a further challenge to determining best practice in the use of Twitter data. As has already been mentioned, Twitter's own Terms of Service are often relied on in academic work as providing informed consent for the collection and analysis of tweets. For researchers working within the European Union this reliance is clouded by Data Protection[9] legislation which states, broadly, that *data should not be used for purposes other than which it was created*. Posting content to communicate on a platform is an objectively different purpose to providing data to be included in research. The notion that users provide informed consent for their tweets to be used in research is further undermined by evidence suggesting that users do not read the Terms and Conditions of a platform in full or understand their implications. If we accept that informed consent occurs at the point of data collection, then we are required to question even standard, apparently non-controversial, practices of collecting tweets via the platform's API. A more fundamental question is to what extent should our own research obligations to ensure good ethical practice be passed over to, and rely on, the actions of commercial agents such as Twitter?

A further form of incompatibility lies in the requirement of Twitter's User Development Policy[10] that tweets are reproduced in full in publications. Any alternations – such as to remove or change a username – would be considered a breach. This requirement challenges standard academic practices of anonymisation (although as already noted, this is hard to achieve with tweets in any case) and also conflicts with European Union[11] and (in the UK) Information Commissioner's Office[12] regulations regarding the handling of data.

On account of Twitter's policy, it is standard to see tweets reproduced in full in media articles. This creates an interesting scenario in research dissemination in which we can link to a news item showing tweets in full but cannot be assured that showing those same tweets that we have perhaps collected via the API would be an acceptable practice. Twitter's policy regarding the reproduction of tweets without alteration is in fact often breached in academic publications. We have seen various instances of published journal articles that cover up usernames and @handles [18] and have heard that journals will sometimes insist on this

[9] http://ec.europa.eu/justice/data-protection/

[10] https://dev.twitter.com/overview/terms/agreement-and-policy

[11] http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:31995L0046

[12] https://ico.org.uk/for-organisations/guide-to-data-protection/

practice where the content of the tweets is in some way sensitive. To the best of our knowledge, Twitter have not taken any action in response to these breaches.

4.6 Research integrity

Although ethical practice should always take priority over academic findings, our consultation highlighted a number of questions over the importance of academic integrity in research using social media data and the ways in which this can be challenged by procedures designed to minimise harm. One key example is in the suggestion that qualitative analysis should present composite or paraphrased data rather than publish actual tweets. This has been suggested in academic publications (in relation to various data contexts) [19] as well as some of the guidance we discussed earlier. It can be seen to present an ideal solution as findings are represented in a global way without any threat to individual identities being revealed, providing a qualitative equivalent of aggregated data.

However, whilst this approach appears acceptable to some it is very problematic to others, in particular to those in research communities where analysis and publication emphasises the value of naturally occurring data. As noted earlier in the paper, the interactional approach followed in our project is one form of analysis that emphasises this approach. It is typically necessary to illustrate findings using data fragments and these fragments are routinely based on naturally occurring data. We contacted the editors of journals we would typically seek to publish in and asked if composite data would be accepted in their publication. We included the suggestion that the actual data could be made available to reviewers to help them evaluate the robustness of the analysis. Their response was that they would reject any articles that used this approach. One editor told us:

Most journals would be reluctant to contemplate fabricated data. There are two reasons for this. The first is that it is not a good proxy for the real thing... The second is that it runs across all the current concerns about research integrity and the potential for falsifying findings. Making data available to reviewers but not publishing it would also run into problems with the open data movement... We accept that qualitative materials may be minimally edited for anonymity - see above- but we cannot contemplate publishing papers that offer no opportunity for readers independently to evaluate the interpretations and analyses offered by their authors.

As this respondent indicates, the reasons to be wary of composite data on the basis of academic integrity are both specific to the type of analysis being undertaken and broadly applicable to research in general. In specific terms the interactional analysis of tweets depends on the detailed and precise analysis of content. To return to the Katie Hopkins post in Figure 1, we consider that it is the particular combination of descriptive terms, use of grammar, sentence construction etc. that makes this post inflammatory and likely to provoke others. The precise content of the responses following this post is also dependent on these particular features. Due to the highly detailed focus of our analysis we are dependent on showing the exact posts to convey our argument. It is virtually impossible for us to accurately illustrate our argument with anything other than the original data as the different wording of a composite post would suggest a different analytic reading of it.

In broader terms we need to ask whether it is ever acceptable to fabricate data in the publication of research. Even if done with genuine intentions this practice contradicts standard understandings of academic integrity. For instance, the Universities UK Concordat to Support Research Integrity[13] states on page 17 that:

"Research misconduct can take many forms, including:
• fabrication: making up results or other outputs (eg, artefacts) and presenting them as if they were real
• falsification: manipulating research processes or changing or omitting data without good cause"

The creation of composite data therefore creates a risk of being seen as research misconduct. We might counter this by stating that the fabrication of posts would be done in a robust and systematic way – but how would we be able to judge this and who would evaluate it?

In our consultation other arguments have been raised around research integrity. As researchers we have a duty to examine difficult topics and social problems. The spread of hate speech on social media can be seen as an instance of this. The prevalence and impact of hateful content is frequently reported in the news and the major platforms are often criticised for not doing enough to stop it[14]. We might therefore argue that it is a priority for us to conduct research into these matters and share our results in a suitable way. This can help to identify solutions to the social problem and perhaps hold platforms to account for their lack of action. However, it might be reasonable to assume that users posting extreme forms of hate speech would not give their consent to be involved in research – for purposes of data collection and publication. Does this mean we cannot carry out the research? It is possible that requirements for informed consent might form a barrier to research being undertaken and lead to researchers shying away from these important issues in favour of topics that are 'easier' to deal with in ethical terms. Another point is that justice for participants is also an ethical principle. It can be argued that those who have been made victims of abuse on a public social media platform have a right for that abuse to be exposed in its original form. If we attempt to paraphrase the content of the abuse or hide the identity of the user who posted it, could it be argued that we are doing a dis-service to the victim of that abuse? Academic integrity is itself an ethical issue and poses significant questions for work in this area.

5. DISCUSSION

This paper has drawn on the outcomes of an ethics consultation to illustrate the challenges that arise when trying to determine what constitutes good practice in research involving the collection, analysis and publication of Twitter data. We have used our own project research as a starting point to investigate the ethical issues that require consideration when working in this area, particularly in the case of publishing specific tweets when disseminating research findings. As illustrated in the sections above, scoping the field has revealed a highly complex landscape that presents major challenges for ongoing research.

[13] http://www.universitiesuk.ac.uk/policy-and-analysis/reports/Documents/2012/the-concordat-to-support-research-integrity.pdf

[14] https://www.theguardian.com/media/2017/mar/14/face-off-mps-and-social-media-giants-online-hate-speech-facebook-twitter

We have identified a lack of consensus in the field over the appropriate ways to collect and handle social media data and procedures through which to disseminate it. These differences exist across institutions, academic publishers and individuals. Reliance on Twitter's Terms of Service to assume informed consent to collect data is standard in many studies but is problematic. Direct approaches to users to seek consent (for instance to publish specific posts) can be achieved but are time consuming and, more fundamentally, do not necessarily guarantee that a user is capable of giving consent. Research using Twitter data – in particular when the content of posts is in some way sensitive – carries a risk of harm to participants. It is therefore necessary to assess this risk when determining procedures for data handling and publication. Challenges arise when we consider that Twitter is a dynamic platform: users will continue to post after we have collected the posts that particularly interest us. If publishing drives audiences towards a particular user, their risk of harm relates to all their posts rather than just the posts we have published. Does that mean our risk assessment needs to attend to previous posts and possible future ones? As an open platform, standard practices of anonymisation do not map well onto Twitter posts as it can be very easy to identify users by putting the content of their posts into the platform's search function. Furthermore, attempts at anonymisation are a breach of Twitter's User Development policy. Twitter's requirements for data handing are also potentially incompatible with regulatory requirements and media practices for publishing posts frequently differ from academic ones. Finally, the use of composite data in place of publishing actual tweets is incompatible with certain analytic approaches and genuine questions arise over how ethical practice can be balanced with the demands of academic integrity. In particular, it could be argued that as researchers we have a duty to study difficult social problems such as the spread of hate speech on social media and that the creation of barriers to accessing and publishing data might lead us to neglect this duty in favour of 'easier' topics.

Our ethics consultation was not able to draw firm conclusions about best practice with regard to publishing tweets and in relation to the use of Twitter data more generally. Instead it highlighted division of opinion across the field that reflected differences of perspective within our own project team. These divisions are frequently founded in different conceptions of what it means to collect data from an open platform and to what extent the status of data as publicly available does or does not alter standard ethical obligations relating to informed consent, anonymisation and the minimisation of harm. They can also be seen to reflect a more fundamental stalemate between two ethical positions. On the one hand a universalist approach might state that if consent cannot be gained and anonymity cannot be assured, then in any and every case research cannot be published. On the other hand, a situated approach might operate on a case-by-case basis, taking into account factors such as the nature of the content being studied and the status of the users involved (as in the public eye or easy to identify etc.). Without taking a stance on either side of this debate, we note that researchers working in the fields of Web science and social science are subject to different and constantly negotiated ethical positions and interpretations of what constitutes harm etc. The implementation of general rules does not map well onto these negotiations and interpretations, and carries a real risk of leading to the censorship of academic work.

The current complexities surrounding good practice and Twitter research create significant challenges for ongoing work in the field. How can researchers collaborate successfully if they are following different guidelines and competing ethical standpoints? How can the results of qualitative research be published if informed consent is necessary but impossible to achieve and if composite data is regarded as unacceptable fabrication? These issues are not specific to Twitter data alone and apply to other forms of Web science research such as those using blogs and forums as data sources in addition to other social media platforms. Furthermore, the challenges we have highlighted have also been discussed in relation to other uses of digital data [20]. The fact that they remain unresolved demonstrates how intractable they are. Our discussions over the ethical handling of Twitter data serve as a case study for a broader range of data sources in Web science and highlight the challenges that the field needs to address.

It is our position that having identified these various tensions around the ethical use of Twitter data in research, it is our obligation to attend to them and reflect seriously on how we can pursue genuine good practice in our work. Furthermore, we also argue that these tensions create real dilemmas for Web science and that it is crucial that attempts are made to address them to enable the field to move forwards. We suggest that open and constructive debate should take place in relation to all of the issues we have highlighted here. Whilst it is unlikely that full consensus could ever be reached, it is possible – and vital – to find a shared pathway that researchers can follow. This debate should include the broad examination of key ethical criteria. For instance, fields such as ubiquitous computing [21] have made attempts to move beyond traditional concepts of informed consent and we can ask whether alternative strategies could acceptably be applied to social media research. Similarly, we need to pose questions such as: in the age of Big Data, to what extent is it possible for informed consent to occur at the point of data collection? To what extent are existing (often static) models of anonymity and risk compatible with public and dynamic sources of data? Where do our responsibilities to minimise harm begin and end in relation to these kinds of data and how can we assess them? How can we balance the risks of conducting and publishing research against the potential risks of not conducting research? Whilst these questions are undoubtedly challenging and the route towards a shared pathway for good practice might be a rocky one, these issues create an exciting opportunity for genuine debate to occur and to effect meaningful change.

6. ACKNOWLEDGMENTS

Our thanks to all those who participated in the ethics consultation reported on in this paper. Plus, thanks to the anonymous reviewers for their comments on the submission. We are also grateful to the Economic and Social Research Council (ESRC) for funding the project: 'Digital Wildfire: (Mis)information flows, propagation and responsible governance' ref ES/L0133981.

7. REFERENCES

[1] For overview see: Gayo-Avello, D. 2013. A meta-analysis of state of the art electoral prediction from Twitter data. *Social Science Computer Review* 31, 6 (Aug. 2013) 649-679. DOI=10.1177/0894439313493979

[2] For overview see: Webb, H., Burnap, P., Procter, R. Rana, O., Stahl, B., Williams, A., Housley, W., Edwards, A., and Jirotka, M. 2016 Digital Wildfires? Propagation, Verification, Regulation and Responsible Innovation *ACM Transactions on Information Systems.*34, 3 (April 2016). DOI= 10.1145/2893478

[3] For example see: Burnap, P. and Williams, M.L. 2016 Us and them.: indentifying cyberhate on Twitter across multiple protected characteristics. *EJP Data Science* 5, 11. Available online at http://orca.cf.ac.uk/88072/. And: Pak, A., and Paroubek, P. 2010 Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *LREc*. Vol. 10. No. 2010.

[4] For overview see: Marwick, A. 2013 Ethnographic and Qualitative Research on Twitter. In Weller, K., Bruns, A., Puschmann, C., Burgess, J. and Mahrt, M. (eds.) *Twitter and Society* New York: Peter Lang, 109-122.

[5] For example, see: Stahl, B. C. Heersmink, R. Goujon, P. Flick, C. van den Hoven, J. Wakunuma, K. Veikko Ikonen, V.T.T. and Rader, M. 2010. Identifying the ethics of Emerging Information and Communication Technologies. *International Journal of Technoethics* 1,4,20-38; boyd, d and Crawford, K. 2012 Critical questions for big data *Information, Communication and Society* 15,6. 662-679 DOI= 10.1080/1369118X.2012.678878; Zimmer, M. 2010. "But the data is already public" on the ethics of research in Facebook. *Ethics and Information Technology* 12,4, 313-325; Zwitter, A.J. 2014. Big Data Ethics *Big Data and Society* 1,2 Nov 2014 doi/10.1177/2053951714559253; Schroeder, R. 2014. Big Data and the brave new world of social media research *Big Data and Society* 1,2 Dec 2014. DOI=10.1177/2053951714563194.

[6] For example see: Neuhas, F. and Webmoor, 2012. Agile ethics for massified research and visualization. *Information, Communication and Society* 15,1, 43-65 DOI= 10.1080/1369118X.2011.616519. And: McNeily, M Hutton, L. and Henderson, T. 2013. Understanding ethical concerns in social media privacy studies. In *Proceedings of the ACM CSCW Workshop on measuring networked social privacy: Qualitative and Quantitative approaches.*

[7] For overview see: Zimmer, M. and Proferes, N.J. 2014. A topology of Twitter research: disciplines, methods and ethics. *Aslib Journal of Information Management* 66,3, 250-261. , 1-10. DOI= 10.1108/AJIM-09-2013-0083.

[8] World Economic Forum, 2013 *Digital Wildfires in a hyperconnected world*. Global Risks Report. World Economic Forum (Feb. 2013). Available online at: .http://reports.weforum.org/global-risks-2013/risk-case-1/digital-wildfires-in-a-hyperconnected-world/

[9] Webb, H., Jirotka, M., Carsten Stahl, B., Housley, W., Edwards, A., Williams, M., Procter, R., Rana, O. and Burnap, P. (2015). Digital wildfires: hyper-connectivity, havoc and a global ethos to govern social media. *Computers and Society 45(3), 193-201.* DOI=10.1145/2874239.2874267

[10] For further details of these approaches see: Housley, W., Webb, H., Williams, M. et. al. (forthcoming) Social Media, Interaction and Categorisation: The Case of Twitter Campaigns. *Social Media and Society* special issue on Social Movements and Social Media. And: Tolmie, P., Procter, R., Rouncefield, M., Liakata, M. and Zubiaga, A. Microblog Analysis as a programme of work. Submitted to *ACM Transactions on Social Computing.*

[11] British Psychological Society, 2013. *Ethics Guidelines for Internet-mediated Research*. INF206/1.2013. Leicester: British Psychological Society. Available online at: www.bps.org.uk/publications/policy-andguidelines/research-guidelines-policydocuments/research-guidelines-poli

[12] Markham, A. and Buchanan, E, 2012. *Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0).* AoIR. Available online at: https://aoir.org/reports/ethics2.pdf

[13] Townsend, L. and Wallace, C. *Social Media Research: A guide to ethics.* Available online at: http://www.dotrural.ac.uk/socialmediaresearchethics.pdf

[14] Dingwall, R. and Murphy, E. 2003. *Qualitative Methods and Health Policy Research.* New York: Aldine de Gruyter.

[15] Evans, H., Ginnis, S. and Bartlett, J 2015. *#social Ethics: a guide to embedding ethics in social media research.* Demos. Available online at https://www.ipsos-mori.com/Assets/Docs/Publications/im-demos-social-ethics-in-social-media-research-summary.pdf

[16] Moffat, A. and Koene, A. 2016. Public Outreach Evaluation Tool (TOOL). Poster presented at the *Social Media and Society Conference* July 2016.

[17] For example, see: Ronson, J. *2015 So You've been publicly shamed.* London: Penguin Publishing group. And see also: http://www2.warwick.ac.uk/fac/soc/pais/people/sorell/tom_sorrell_-_digilantism.mp4

[18] For example, Awan, I. 2014. Islamophobia on Twitter: A Typology of Online Hate Against Muslims on Social Media *Policy & Internet* 6, 2, 133-150.

[19] Markham, A. 2005. "Go ugly early": Fragmented narrative and bricolage as interpretive method. *Qualitative Inquiry 11,16, 813-819* DOI=10.1177/1077800405280662.

[20] Carus, A., and Jirotka, M. 2009 From Data Archive to Ethical Labyrinth *Qualitative Research* 9, 3, 285- 299

[21] See for example: Moran, S., Luger, E. and Rodden, T. 2014. An emerging toolkit for attaining informed consent in UbiComp. *UbiComp '14 Adjunct* Sep 2014 Available online at: http://ubicomp.org/ubicomp2014/proceedings/ubicomp_adjunct/workshops/Consent/p635-moran.pdf