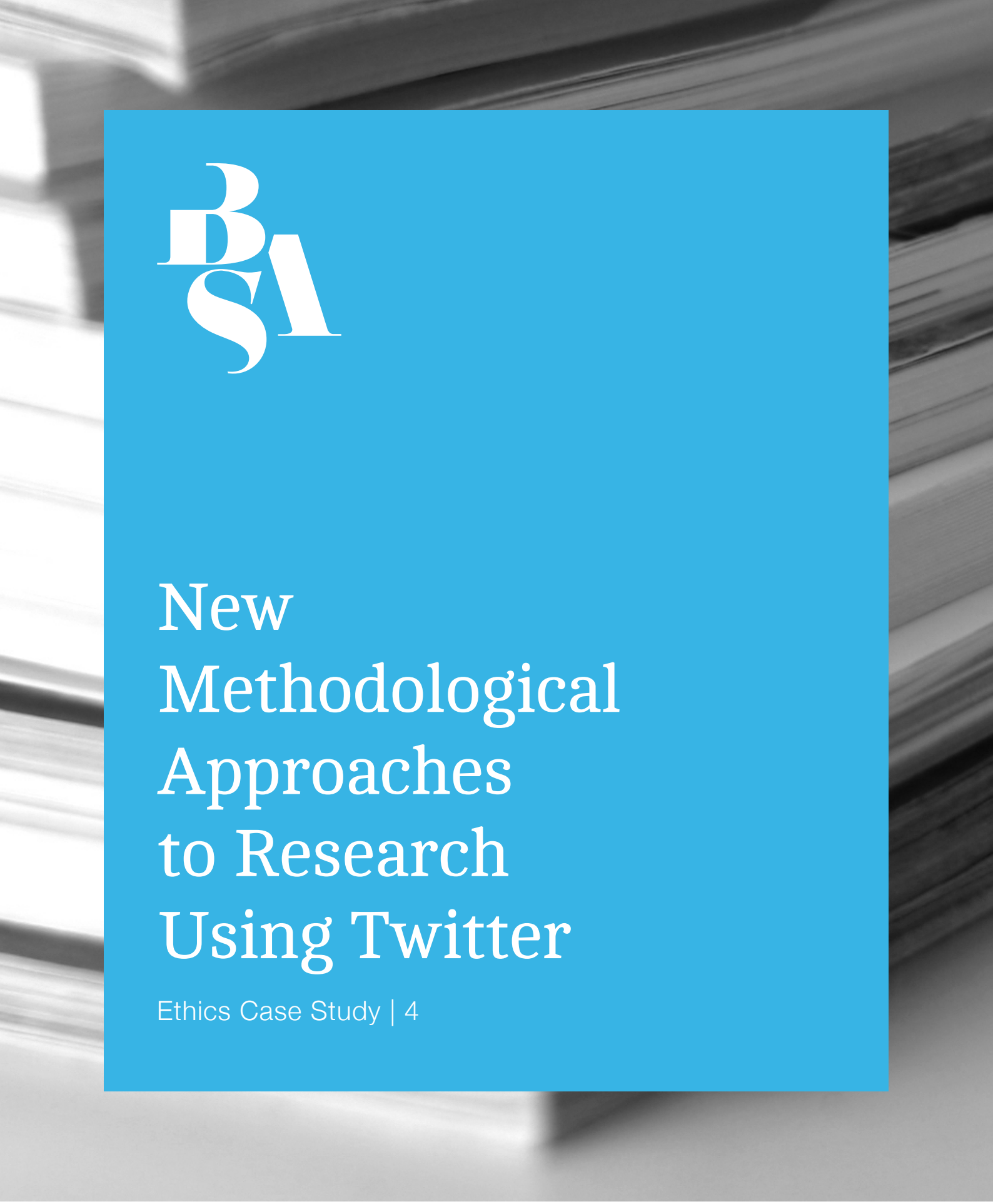

New Methodological Approaches to Research Using Twitter

Ethics Case Study | 4

www.britsoc.co.uk

BRITISH SOCIOLOGICAL ASSOCIATION

New Methodological Approaches to Research Using Twitter

Anonymous

This is an application form to an ethics committee for research into new methodological approaches to Big Data, specifically the Twitter micro-blogging service.

Describe the rationale, study aims and the relevant research questions of your study.

This study aims to develop a new methodological approach to research on social media, specifically the Twitter micro-blogging service. Whilst there is already considerable research interest in the 'influence' of social media, studies of Twitter have – to date – only considered the extent and circulation of 'retweets' (where users pass on an original post to their own followers) as the measure of 'influence'.

The aim of the research proposed here is to extend this measure of influence by adding contextual information about followers and users' Twitter network; and by exploring the relationship between tweets, retweets and the network of followers that surround a user. Collecting this additional information will allow assessment of the relationship between the users' activity and the wider Twitter context within which this activity takes place. For example, it would be possible to see how a change in followers affects tweets; or how retweets might generate new followers.

This is an inter-disciplinary research project which draws together perspectives from the social and computational sciences.

From sociology, the project takes theories of social activity and digital transformations in the information age, as well as epistemological debate about the relations between theory and method. From computer science, the specific interest is in the technical opportunities and constraints of data harvesting. Overall, this perspective, the research is mainly focused on the process of data collection, from the conceptualisation of what constitutes important information among the available data, to the implementation of the collection *per se.*

The focus of the project is not on particular participants or the activities associated with individuals, but on (i) the conceptualisation and process of data collection and (ii) the aggregate analysis of the patterns and relationships in the digital data.

This case study was originally published in draft form on the British Sociological Association Digital Sociology Study Group blog (2016)

Describe the design of your study

Access to the data

This data is publicly available via two Twitter Application Programming Interfaces (API).

The Stream API allows researchers to have access to the tweets and the REST API provides access to the users' profile information. Profile information is to be used only for checking the accuracy of the harvesting process (see below).

The terms of use for these APIs are set up by Twitter and the research complies with these rules. These rules can be found at the following address: https://developer.twitter.com/en/developer-terms/agreement-and-policy.html

Selection of the population

The research will explore the activity of 200 initial Twitter users in each of 3 different groups (total 600 initial users). These groups have been chosen purposively, to explore different types of online activity:

1. An online group with a clearly existing offline community.

2. A group formed around a Hashtag (a word following the symbol # to allow user to participate in the same discussion) to study the specificity of a community developed around a specific and temporal interest.

3. A group of randomly chosen users.

Information collected

In each case, the study will collect profile information, network information and Twitter activity.

Profile information

The profile information will contain the screen name (the name displayed on the account user), the location (if it is set up by the user), the language (if it is set up by the user) and its id_str (a unique identifier created for each account on Twitter to identify them even if they change their screen name).

The screen name will only be used to perform manual checks on the data to validate the accurate operation of the script. All linking of screen name to data collected will be deleted as soon as this data collection is done. The language and the location will be retained to allow later analysis of shared/divergent characteristics in the user networks. The id_str which takes the form of an integer is retained as a unique identifier to collect information and to ensure the same user is tracked. This will only be used within the data collection process, not in any subsequent analysis or publication.

Network information

Two types of inter-user links can be found on Twitter, the Followers and the Friends. The 'Followers' are those accounts following the user. The 'Friends' are the accounts the user is following.

A 'snapshot' of Followers and Friends will be taken for each of the original 200 users each time profile information is collected (as above). In order to trace the emergent networks of users, the research proposed here will take regular snapshots of each of the 200 users, Followers and Friends.

This will generate a second list of users to be included in the research. The profile information for the second list will be collected according to the same principles outlined above. Regular snapshots will also be taken of these users' network of Followers.

The regularity of the snapshot in both cases will depend on the number of users included in the second list and the REST API limitation (180 calls every 15 minutes). The fact that the second list of users, is generated by the activity of the original 200 users makes impossible an a priori estimation of the number of API calls that will be needed. Therefore, the time interval will be dependent on the number of users included and the number of Followers and Friends every user has. However, a lower limit of one snapshot per day is set up in the script to ensure a regularity. This lower limit will then act as a limit of users in the second list and can change between two different datasets.

The size of the second list cannot be predicted either, as it is based on the activity of the primary users. However, a total limit of 5000 users (the addition of the first list and the second list for each group) is needed as it is a limit from the Stream API or the limit of one snapshot per day, which is a limit from the REST API.

This number of 5000 users does not means it will be only 5000 users, but only 5000 users at the same time. After a defined period of time (to be confirmed if the list reaches the limit of 5000) during which there has been no activity between the second and first users, second users will be dropped from the list. Therefore a user can be in the second list, then being removed, then being in the list again, depending on activity.

Every time the script collects information about a user, the list of current Followers and Friends is updated as well as any change in comparison to the previous list. This list itself contains only the id_str of the friends and followers. The id_str will be used to access collect Profile Information on Friends and Followers, but not the screen name.

Twitter activity

The Twitter activity is the tweet posted by the users on their public timeline.

I have access to this information from both the Stream API and REST APIs.

The information collected is the text itself with a time stamp.

The text may contain URLs (links to other websites), hashtags (linking the tweet to other tweets) and direct mentions of other users. Whilst the URLs and hastags will be used for later analysis, any direct mentions will only be used to build the sample.

Stream and REST API

The Stream API is used to collect the tweet in real time while the REST API is used to collect past tweets.

If a user from the first or the second list has a network activity with a new user, the REST API is used in order to collect the last 3500 tweets of the new user. Then it is added to the list of users screened by the Stream API.

The Stream API is used for two reasons: first, as a second access channel to Twitter data, to overcome the limited number of calls to the API; and second, to provide real-time information about users' activity.

The users from the first list are added in the Stream API search terms. Every time they tweet something or retweet or are mentioned in someone else tweet, the API collects this information.

The tweet is therefore stored and if the tweet mentions a user, this user is included in the second list. In this way, the activity centred around the publication of the message is in real-time and does not use too much API call.

Storing data

The data collection does not involve any analysis, or even observation of the data, beyond some basic checking to ensure that the data harvesting is proceeding as planned. The entire process is automated through a script developed for this purpose.

The data are stored on a NOSQL database (using MongoDB) in scheme to facilitate the retrieval of information but does not add other personal information than the ones retrieved from Twitter (I can communicate the template if needed).

User consent

During the data collection, I will use a specific Twitter account created for the research, to contact each individual for whom data has been harvested to ask if they have any objection to the anonymized analysis of their data.

Users will be contacted via the Direct Message system of Twitter to give a link to a website (removed for anonymity) (the URL is temporary; for the actual respondents, university address will be used to ensure a better credibility). This website provides more information about the study, the harvesting of the data, the process of anonymization and the contact information for any enquiry.

At the bottom of the web page, an opt-out form can give them the opportunity to be removed from the dataset (for the reason of an opt-out system, see the point 13 and 18) if they wish to.

It is planned to send a first message to the people added to the first level of the list right after the data collection, and one week, then three weeks after. These messages contain the same URL as the page with the option to withdraw, and a short message (140 characters maximum) to describe the link.

Message: *"I am a PhD student researching Twitter use. For this, I have collected some of the publicly available information from your Twitter account. Information on how to withdraw if you wish are given in the following link"*

Second and third messages will only be sent to users who have not already replied.

After four weeks, if the participant has not expressed any wish to be removed from the dataset, I will consider that I can use the data for the analysis purpose.

This Twitter account can be found here: [REMOVED]

Data analysis

The purpose of the thesis is to develop an improved method for analysing Twitter data. The main focus of the thesis is on the development of the method and its theoretical implications rather than information about usage per se.

However, to test the hypothesis about the influence of context, networks and activity, some particular metrics will be used to analyse these data:

- The evolution of number of Followers and Friends
- The link shared within the tweets
- The number of mentions and retweets a user sent and received.

To conduct the analysis, the dataset will be completely anonymized, removing any information which could lead to the identification of the user (see point 20 and 21).

Who are the research participants?

There are three related (or 'nested') lists of participants:

Level 1: The main users

200 users will be selected from each of three groups (see 9.1 above). Consent will be sought prior to any data collection, as removing these 'primary' users later would cause significant disruption to the overall data set.

The information collected about these participants is Profile Information – Network Information and Twitter Activity.

Level 2: The activity users

The second list of participants is dynamically created. It depends on who the participants from the first list interact with. Ifa participant on the first list mentions, retweets, adds or removes a user, this user is added the second list.

The information collected about these participants is Profile Information – Network Information and Twitter Activity. Further information collected differs from Level 1 users in two ways:

(i) if there is no further interaction after one week, this user is dropped from the list and no further information is collected, unless an interaction is again detected with a user from the primary list.

(ii) if these 'second level' users interact with other users, this interaction is not used to gather more people, it is only information kept to know their activity.

Level 3: The contextual users

The third list is created by each user presented in the Followers and/or Friends networks from the primary and the secondary list. The amount of information is only limited to the id_str, and the information about the number of friends and followers they have, as well as the number of statuses published (containing the tweets they originally posted but also the retweet they published), the number of friends and followers. No more information is collected.

No consent will be asked for this list of users as they are representing a social context. No information about their followers and friends list is retained. It is only used to be able to draw a network graph and see the overlapping interaction between the users from the first and the second list. Also, the size of this dataset makes it impossible to individually contact the participants without being considered an abusive behaviour under the Twitter contract of API use.

If you are going to analyse secondary data, from where are you obtaining it?

N/A

Will participants be taking part in your study without their knowledge and consent at the time (e.g. covert observation of people)? If yes, please explain why this is necessary.

The data collection will take place without participants' knowledge. However, the data will not be analysed before consent has been given.

The first reason is the nature of Twitter itself. An account does not necessarily imply a real person behind it. It could be an organisation, a group of people or a robot and in this context, asking for prior consent can lead to remove some participants that will not pose any ethical issues, while they will offer valuable insight for the study. For instance, an organisation will have different behavioural patterns than an individual, and it is expected that the study will show these differences.

The second reason is the possibility that users will not see the message in time. Some accounts are not active or very active, therefore the user associated to the account can miss the message before the collection of data. By sending several messages with a sufficient time lapse between them, it is possible to collect the data (automatically), and ensure the maximum visibility of it. By waiting for a prior consent, the relevance of the information is not possible. For instance, it is impossible to collect tweets older than a week. If the user doesn't reply in this interval, the information about the context is lost.

The final reason, specific to the third list, is the number of people included in it and the limited amount of information collected. It is practically impossible to send a message to all people collected through the Network list, it involves thousands of users and the Twitter service will not allow any account to follow that many people in a short period of time and send them a message. It will be considered by Twitter as spam and the account will be suspended.

For all these reasons, it is not possible to ask prior consent for users and it is why the opt-out system is adopted as a more efficient method.

If you answered 'no' to the questions above, how will you obtain the consent of participants?

N/A

Is there any reason to believe participants may not be able to give full informed consent? If yes, what steps do you propose to take to safeguard their interests?

N/A

If participants are under the responsibility or care of others (such as parents/carers, teachers or medical staff) what plans do you have to obtain permission to approach the participants to take part in the study?

N/A

Describe what participation in your study will involve for study participants. Please attach copies of any questionnaires and/or interview schedules and/or observation topic list to be used.

Only observation, no interaction or questions to the participants.

How will you make it clear to participants that they may withdraw consent to participate at any point during the research without penalty?

During the collection there is no consent, but for the analysis, a private message will be sent with a Twitter account created for this purpose. This account will give details about the research and contact details (see point 9 above).

The message will give a link to a web page describing the purpose of the research, the respect of the anonymity and the possibility to remove the data from the dataset.

Detail any possible distress, discomfort, inconvenience or other adverse effects the participants may experience, including after the study, and how you will deal with this.

N/A

How will you maintain participant anonymity and confidentiality in collecting, analysing and writing up your data?

The collected data will not be anonymized at the first stage of the collection, as the identity (represented by the id_str as a unique identifier used by Twitter) is important to ensure the quality of the dataset. The screen name is only present to do some manual checks, mainly to make the reading of the script's logs easier, and will not be used for any other purpose. The screen name will be dropped prior to any analysis and from this point on users will be labelled with an anonymized random number, automatically generated and separated from the user ID. Prior to any analysis, every id_str will be matched in a separate database created for the purposes of this research with a random number. It is this number that will be used to conduct analysis. Information about profile location and language will still be stored, but the information published will not be directly associated to one user. That will give only valuable information on the dataset for the analysis purpose (as knowing the spread of the dataset over the world and the different language spoken).

The destruction of the database containing the concordance between random number and Twitter ID will be done only at the very end of the research. It is to allow me to be able to remove people if they are asking for it, even after I start to analyze the data (a possibility is to keep this database as long as the dataset is available, if this latter option is better, then the database will be encrypted and stored on a different server than the server hosting the dataset).

How will you store your data securely during and after the study?

The data will be stored in a virtual machine hosted on the University server. The only person who has access to it is me. The table having the correspondence between the Twitter ID and the random number will be stored on a password-protected laptop behind the University fire wall.

Describe any plans you have for feeding back the findings of the study to participants.

The feedback about the findings will use the same methods as for contacting the users before and during the analysis, described in 18. The feedback will contains information about the PhD itself, the main result found and the assurance that none of the data are identifiable. In case of publication, the reference of the publication will be given as well.

What are the main ethical issues raised by your research and how do you intend to manage these?

The research is based on publicly available information. However, whilst users have posted information publicly, and indeed the purpose of Twitter is to tell the world 'what's on your mind', we cannot assume that users are aware of the possibilities for analysis of the data that they are posting online. For this reason, all data will be anonymized following emergent practice in the field of Twitter research. Furthermore, users included in the study will be given the opportunity to opt out prior to data analysis.

The opt out approach is chosen due to the very nature of Twitter. It is impossible to know if the user is a human, a bot, or a company. Therefore, it is only if the user actively expresses a desire not to be included in the analysis that all data about him/her, will be removed.