Understanding Society
Working Paper Series

No. 2020 - 04

March 2020

Linking Survey and Social Media Data

Tarek Al Baghal (University of Essex)

Understanding Society

Economic and Social Research Council

Non-technical summary

Linking social media and survey provides an opportunity to improve research currently being done with both data sources, through supplementing additional measures of interest or improving existing methodologies.

Access to many social media data sources is limited, with Twitter being the most accessible and direct to collect, link, and use alongside survey data.

Consent is required to link these data sources, due both to ethical and logistical considerations. Few studies have explored consent to linkage, with those that have finding some variation in consent rates, possibly due to methodological factors.

The few studies that have linked survey and Twitter data have focused initially on political research, but there are both substantive and methodological research that can be done using linked data. Methodologically this includes improving classification tools designed for social media research and for survey research this may assist with intrawave data collection and data from units dropping out of the survey.

Archiving these linked data for use by a wider audience requires additional consideration beyond normal data archiving practices. Consideration is needed because of the disclosive nature of social media data and balancing the need for anonymity while maintaining the usefulness of the data.

Linking Survey and Social Media Data

Tarek Al Baghal (University of Essex)

Abstract: In light of issues such as increasing unit non-response in surveys, several studies argue that social media sources can be used as a viable alternative. However, there are also a number of shortcomings with social media data such as questions about its representativeness of the wider population and the inability to validate whose data you are collecting. A useful way forward could be to link survey and social media data to supplement and improve both. **This briefing will explore various facets and issues of linking survey and social media data**. These facets and issues include: the current context of surveys and social media, particularly in the UK; availability of data from various social media sites; gaining consent for linkage; linking and collecting the data; usage of the linked data; and archiving and re-use of the linked data. Work done as part of a research group exploring these linkages, including a study conducted in the tenth wave of the *Understanding Society Innovation Panel* (IP) (University of Essex 2018) has been among the first to be published on these topics. As such, this brief incorporate the ongoing work done by the research group, as well as exploring what has been done in limited number of published studies exploring such specific linkages where possible.

Keywords: data linkage, social media, response rates

Acknowledgements: The author acknowledges the contribution of Luke Sloan, Nina Di Cara, and Oliver Davis for information on current state of research in the field.

Understanding Society is an initiative funded by the Economic and Social Research Council and various Government Departments, with scientific leadership by the Institute for Social and Economic Research, University of Essex, and survey delivery by NatCen Social Research and Kantar Public. The research data are distributed by the UK Data Service. This methodological briefing has been funded by the Economic and Social Research Council UK Population Lab Innovation Development Grant (ES/S016651/1).

Data note: University of Essex. Institute for Social and Economic Research, NatCen Social Research, Kantar Public. (2018). *Understanding Society*: Waves 1-8, 2009-2017 and Harmonised BHPS: Waves 1-18, 1991-2009. [data collection]. 11th Edition. UK Data Service. SN: 6614, http://doi.org/10.5255/UKDA-SN-6614-12

Corresponding author: Tarek Al Baghal, Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ, talbag@essex.ac.uk

Context

Coinciding with declining survey response rates has been the increasing interaction of people with social media and other technologies, generating a massive amount of new data. In regards to social media, usage in the UK of sites like Facebook, Twitter, and Instagram continues to grow. In 2011, 45% of Britons said they used the internet to access social networking sites, with 65% reporting so in 2018 (ONS 2018). Usage is highest among younger people, with 93% of 16-24 and 88% of 25-34 reporting usage (ONS 2018). These differences in usage are important for possible application to panel studies: in *Understanding Society* younger people were less likely to initially respond and more likely to drop-out (Lynn et al. 2012).

In light of issues surveys face such as nonresponse, social media may be a useful supplementary source of data to study social phenomena. Social media has been used in studying areas of social networks, health, economics, politics and crime patterns (Ellison et al. 2007; Steinfield et al. 2009; Scanfeld et al. 2010; DiGrazia et al. 2013; Papaioannou et al. 2013; Williams et al. 2016). The limited work to date using linked survey and social media data has focused on political research (Karlsen and Enjolras 2016; Eady et al. 2019), but clearly there are additional variables that may be obtained through analysis of social media data. Further, social media data may be able to improve survey methodologies such as improving unit and item nonresponse adjustment and testing measurement properties of existing questions (Al Baghal et al. 2019). While several studies argue that sources such as Twitter can possibly substitute for surveys (Cummings et al. 2011; O'Connor et al. 2011; DiGrazia et al. 2013). a more recent study found that Twitter and survey data would only yield similar conclusions under particular conditions (Pasek et al. 2018). This latter study's findings suggest correspondence is better if examined over longer periods of time, for example through linkage with existing longitudinal studies

Part of the problem with social media use is that it is not universal, and it is likely that social media users are a self-selected and possibly not representative part of the population. The usage of these sites/applications (and hence data produced) within users also varies greatly. The last statistics put out on UK usage by Facebook and Twitter come from 2013. Facebook reported 33 million monthly-active users in the UK in 2013 (Halliday 2013), with 15 million reported by Twitter (Curtis 2013). A recent survey conducted for Ofcom found that Facebook remains the most popular social media site, with 88% of internet users saying they have an account they still use (Ofcom 2019). However, this is a decline from 91% in 2017. WhatsApp is used by 61%, Instagram used by 38%, Twitter 25%, and LinkedIn 16%. The finding for usage of Twitter is similar to results in the IP (21.6%) and the NatCen Panel (25.6%), both conducted in 2017.

Further, social media is 'data-light' (Gayo-Avello 2012) with many of the key demographic variables used in social scientific analysis such as age, gender, occupation and class missing. Social media contains a large amount data, but knowledge of who is producing it is significantly limited (Sloan 2017a). In response to this specific challenge, researchers have endeavoured to derive important demographic data from the content and metadata of Twitter. This data includes information on location and language (Graham, Hale & Gaffney 2014), gender (Sloan et al. 2013), occupation, social class and age (Sloan et al. 2015). However, research has shown that these proxy approaches can be inaccurate (Sloan 2017b). Comparatively, survey data contains good measures of these demographics, and social media data is not needed to provide this information. Rather, linked social media can add new supplementary data to surveys while also providing an opportunity to verify the accuracy of demographic categorisers and to understand where proxy indicators are going wrong.

Given the issues facing both the use of survey and social media data, a useful way forward could be to combine survey and social media data to supplement and improve both. To do so, consent within a survey is first needed. In the case of Twitter, data is openly available; however, consent is still necessitated for not only ethical reasons, but also logistic as handles need to be collected to correctly link the data.

Availability of Social Media Data

There has been limited research in linking survey and social media data. When research has attempted to link social media to survey data, Twitter has been the most frequently used social media source, for several reasons. Most important is accessibility. Twitter allows access to its data through the site's Application Programming Interface (API), which is easily accessible through a number of freely available programs, including COSMOS (http://socialdatalab.net/) and several packages in R and Python coding languages.
Conversely, several other social media applications are near impossible to access. In particular, Facebook has increased security and access to its data, partly due to the events surrounding Cambridge Analytica. Some linkage research was done using Facebook prior to the introduction of these new restrictions, but future research appears limited. Facebook also owns WhatsApp and Instagram and has added similar security to those applications as well, limiting possible data access. To access these data, researchers would need to go through the company directly. Snapchat, another major social media application, also does not make user data available. LinkedIn allows access to its API, but its terms of service seems to forbid data collection/scraping, although there is some ambiguity that requires further clarification from the company.

An additional reason for the focus on Twitter data is the nature of the data available. Posts to Twitter, "tweets", are largely textual, and there are a variety of methodologies and tools to analyze text, both qualitative and quantitative in nature. The number of tweets from any given user that can be accessed at any one time is the 3200 most recent – and in a longitudinal study, these requests could be made over numerous time points. Twitter data also contains metadata, such as number of followers a user has, how many they follow, how many tweets they have posted, retweeted, liked, etc. Metadata include any user-provided details such as location data, although these are frequently missing, due to the user not providing information or disabling features.

It is possible to access YouTube data, but data is about individual channels within Youtube, including information on posted videos and comments made. Many, if not most respondents, may not participate in posting videos and/or making comments. To the extent they were, data would need to be collected about behaviour using the site, what channels are being used, and user name in order to collect data.

Consent to Data Linkage

It is important from an ethical perspective to inform respondents about what their involvement in the study entails so they can agree to participate, however, recent changes in law has implications for gaining consent (Sloan et al. 2019). Depending on how the data will be processed for a given study, GDPR regulation may require consent as the legal basis for collecting data.[1] However, GDPR does open up the possibility of studies such as *Understanding Society* using public task as the lawful basis for processing data. Currently, the Study team's view is that in order to maintain constructive relations with our participants we will continue to seek informed consent for linkage.

Recent qualitative work done by Davis and Haworth (https://dynamicgenetics.org) as part of CLOSER work package 21 with ALSPAC (unpublished) explored the views respondents (both youth cohort and parents) have towards possibly linking their social media and survey data. The work suggested that the both younger and older participants felt that the linkage of these data were acceptable. This view of acceptability seemed to be related to the trust panel members have in the study and understanding of the usefulness of this data to enhance the research. However, the acceptance to link data varied over the type of data, with participants less willing to share photos than text and less willing to share info from friends/followers. The main differences between the

[1] https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/consent/

generations were on their views of data privacy – older participants seemed less concerned about the prospect of sharing their information in general than the younger generation. ALSPAC is currently contacting all participants in the young people and parent generations to request consent to link Twitter data. These acceptability findings were consistent with the researchers' previous findings in focus groups with the Twins Early Development Study (TEDS). During Twitter linkage carried out as part of an MRC Centenary Award to Davis and Haworth, 6000 participants identified themselves as having a Twitter account and 2400 provided consent to link Twitter data.

Although not specifically about data linkage, in its fifth main-stage wave *Understanding Society* asked respondents "May we send you a message through any social networking web site such as Facebook or LinkedIn if we are unable to reach you through mail, telephone or regular e-mail?". Facebook would have allowed the study to search for sample members and then contact them using the *Understanding Society* Facebook account under the terms and conditions at the time the question was initially placed into the survey. However, during fieldwork changes in Facebook's terms and conditions made it against their terms to use a 'business' page (which ISER was classified as) to contact people or for a 'personal' page to contact people on behalf of a business. Hence no further work was carried out on this project. It is noteworthy, though, as only 18.8% of social media users (73.7% of the total observed sample) agreed to be contacted in this way.

A small number of studies have reported consent rates to link social media and survey data, although only one published study explored consent quantitatively as the outcome of interest. One study took place in Norway, where researchers conducted a survey of political candidates running for parliament in 2013 (Karlsen & Enjolras 2016). In this survey, they asked responding candidates to consent to link their Twitter data to survey responses in order to study campaign use of social media. Of the 41% of candidates using Twitter as part of their campaign, 49% consented to data linkage.

Three additional studies occurred in the United States, all using online panel survey samples (but only using data from one wave). Murphy et al. (2013) found that 19% of respondents used Twitter, and of these users, 27% consented to link their survey and Twitter data. A more recent study used an online panel sample that had already consented to link their survey data to their Twitter account data.[2] However, the study also asked respondents to consent to linkage of their survey and Facebook data.[3] Among Facebook users, 45% consented to this linkage - even though all of these respondents had previously consented to link their Twitter data (Guess et al. 2018). Most recently, the Pew Research Center obtained a much higher consent rate (90%) among Twitter users to link their data to survey responses (Wojcik & Hughes 2019).

One published study has explored consent to link Twitter and survey data as the outcome of interest, comparing consent rates in three studies in the UK, including the tenth wave of the IP (Al Baghal et al. 2019). The consent rates varied somewhat across the studies. The British Social Attitudes 2015 (BSA) found 18% of respondents were Twitter users, and 37% of these consented to linkage. Almost 26% of the NatCen Panel (NCP) respondents (in July 2017) were Twitter users, and 27% of these consented to link their data. The IP (2017) obtained a 31% consent rate among the 22% of the respondents identifying as Twitter users. Further analyses found that in two of these surveys (BSA, NCP), younger respondents consented at a lower rate than older, with women less likely to consent in one study (NCP) and no difference found in the other two. Importantly, two of these studies used mixed-mode designs (NCP, IP).[4] Consent rates varied by survey mode, with respondents being asked to consent directly by an interviewer having higher consent rates than among respondents answering via the web. It is argued that this difference could be due to the interviewer providing a reassuring face to the study and being able to answer questions, improve trust and reduce concerns about privacy.

The Pew Research Center consent rate stands out compared to these UK findings. Although the reason for these differences have not been definitely ascertained, one cause may be the sample – Pew targeted Twitter users on

[2] Consent obtained by YouGov, these consent rates are not available
[3] This occurred in 2016 before much of the increased data limitations
[4] The BSA was face-to-face only

an online panel, whereas the BSA, NCP, and IP are nationally representative probability samples recruited through face-to-face methods. Another possible reason may be the question used to ask for consent. The UK studies all first had to ask if a respondent used Twitter (not needed for Pew) and then ask for consent. These consent questions were lengthier and more informative than that used by Pew, including several additional help screens (see Appendix A).

The additional information provided, particularly in the NCP and IP, were concerned with communicating: (1) why the data was being collected; (2) what was planned for the data; (3) what information would be collected; (4) that the data would held securely; (5) that there will not be identifiable published information. The Pew question, while straightforward, was more general and potentially vague about some of these, particularly (2) and (3). However, further research is needed to understand the impact of different wordings on consent, while maintaining fully informed consent.

Collecting the data

Once consent has been obtained, social media identifiers, such as Twitter handles, need to be recorded in a follow-up question. Even among those consenting, linkage is not ensured. A small pilot test conducted in Mexico (Al Baghal et al. 2015) asked for Twitter handles from identified users. Sixty-eight respondents, 40.5% of users, provided a handle. However, most were suspected to be invalid, with only 28% (n=19) seeming to have a reasonable likelihood of belonging to respondents. The remaining respondents gave handles that were invalid (i.e. not able to be matched), were matched but didn't appear to actually be the respondent's account (e.g. different languages or listed gender), while others were linked but were private and not directly accessible.

Even though everyone in the Guess et al. (2018) sample consented to link their Twitter and survey data, only 62% provided a response when asked for a handle, and not all of these were valid. The analytic sample was 52% (n=1,816) of the original sample of consenters. This study's linkage to Facebook required signing in directly to the site through a study-designed third-party application, so error in identifiers was not an issue. Again, the Pew Research Center (Wojcik & Hughes 2019) study did better, though it still had some loss due to errors with 84% of consenters providing a valid Twitter handle. There are some possible explanations for these issues: accounts could be removed; changes can be made to usernames between the consent request and linkage attempt; respondents may confuse login details and handles; usernames could be mistyped; and some respondents might submit intentionally misleading responses.

Once the user identifiers have been collected and verified, several programs have been developed to access the social media site's API to collect available data, such as tuber, which collects YouTube data using the program R (Sood 2018). For Twitter data, there is the COSMOS web application, which has been designed to allow non-technical users access to this rich and voluminous data source.[5] CLOSER work package 21 has worked with CLOSER cohort leaders and participants to develop an open source Twitter linkage software framework specifically designed for linking Twitter data in UK cohorts. It is designed to be simple to install and run on a cohort's own servers, ensuring that identifiable information does not leave the cohort's data safe haven (https://dynamicgenetics.org). In Python there are wrappers such as tweepy, TwitterSearch, and twython, among others.[6] In R, there are packages such as twitteR (Gentry 2014) and streamR (Barbera 2018). As noted by Sloan et al. (2019: 4) the "simplicity of such tools belies the complexity of the data that is being collected and the sheer amount of information that is associated with a single tweet, which can come with over 150 associated 'attributes'[7] or, in the language of the social sciences, 'variables'." Every single tweet collected can contain up to 150 columns of data, including the tweet, but also language, colour of the profile, likes, etc. However, this large amount of data allows for an extensive resource for researchers to use.

[5] http://socialdatalab.net/COSMOS
[6] https://stackabuse.com/accessing-the-twitter-api-with-python/
[7] https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json.html

Usage of Data

Few studies have actually analyzed linked survey and social media. The work using the IP (and NCP) is in the beginning stages of this work, with findings forthcoming. Early work used both human and machine coding of tweets to try and classify respondent on various demographics captured in the survey, with mixed success (Murphy et al. 2013), mirroring some of the classification errors identified in other studies (Sloan 2017b). More recently, in their study of Norwegian parliamentary candidates, Karlsen and Enjolras (2016) identified two styles of social media campaigning, one party-centered and one individual-centered. Those using an individual style were more likely to be active on Twitter. The linked data show that the most influential (those with most mentions in others' tweets) candidates on Twitter were male, younger, and have a higher list position (part of the Norwegian party electoral system) than those less influential on Twitter.

Another study using linked data found that respondents' self-reports of social media use was correlated with actual usage, but there were substantial discrepancies (both over- and under-reporting) as well (Guess et al. 2018). To improve these issues, the authors recommend asking for a wider range of social media behaviours and about specific details. Analysis of this same linked data showed that respondents on Twitter did not only follow news that mirrors their political views (Eady et al. 2019). Rather, there is substantial overlap in news sources followed on Twitter between those on both the left and right of the political spectrum.

Beyond the specific research questions these studies explored, linked data can begin to address methodological issues that affect myriad research projects using social media and survey data. In regard to social media research, there are concerns about the accuracy of methods to understand who is providing data. Linked data has the potential to improve existing tools for social media analysis by testing against a gold-standard (i.e. survey data), build improved classifiers and in turn be more confident (or more realistic) in understanding the power of algorithms to make important social scientific distinctions between groups. An example could be that looking at occupational terms NS-SEC (i.e. social class) allocation is enhanced by the consideration of verbs rather than nouns, such as the word 'lecturing' in a social media profile may be sufficient to classify a user as a 'lecturer'. Further, using machine learning techniques may increase accuracy by drawing on literature on how language use is associated with gender and age (Argamon et al. 2006), particularly for social media (Schwartz et al. 2013).

Additionally, the way indicators for attitudes and behaviours are generated by social media data may be improved through data linkage. For example, when predicting elections Burnap et al. (2016) identified tweets containing the names of political party leaders that had been classified as 'positive' through sentiment analysis. Using the number of for positive individual leader mentions as a vote distribution, the authors computed this as a national swing. This is a simple approach to finding the "signal in the noise", these positive mentions to do not directly indicate party affiliation or voting intentions. Survey data that has these types of measures (e.g. party affiliation, intention to vote) can be linked to social media data, comparing if the same constructs can be measured through the social media activity of respondents.

Mirroring this potential improvement in social media data, measurement may also be improved in surveys, especially in a longitudinal context, through linkage of the two data sources. Using programs such as the tidyverse package in R (Silge & Robinson 2017), text analysis can generate indicators to match the domains measured in the survey. Sentiment analysis can provide indicators of the valence of social content. A program such as Linguistic Inquiry and Word Count (LIWC) can also be used, as it mines the text and outputs 80 variables, relating to, among others, linguistic, psychological, social and biological processes, beliefs, and socio-economic issues (Pennebaker et al. 2015). Given the dynamic and fluid nature of social media, these data can be generated between panel waves adding further information to the nature of change which is of particular interest to longitudinal studies. The indicators for social media can be used as new measures of interest or can be used to further improve other aspects of the survey method.

These new indicators may add to the richness of the data available, and also to use in improving measures and methods for non-response adjustment. For example, much of the observed change across panel waves is likely spurious (Jäckle 2009), and collected and derived intra-wave data may help identify the nature of events leading to reported change. Measurement models can also be used to estimate differences in measurement and error between data sources, and can also incorporate longitudinal aspects of the data (Biemer 2011). Unit nonresponse may be improved in the instance where a respondent who agrees to link their social media data drops out of the study (other than explicit withdrawal or for reasons making them ineligible), and passive data collection from social media could be continued. Item nonresponse adjustment methods can be developed by using data from linked social media accounts in methods used currently in *Understanding Society* including regression models, predictive mean matching, and hot-deck imputation (Knies 2018). These methods are necessarily limited to those respondents having social media and consenting to link these data, however. This limited sample set is of concern, as nonresponse in surveys can limit the overall observed sample and those using social media platforms such as Twitter can be a relatively small portion of the population. However, the potential remains and further work is warranted to explore the extent these data sources can mutually improve the other.

Archiving and re-use of data[8]

The archiving and sharing of research data are important elements of the research process and often a requirement for funding organisations or research publications. The processes for archiving linked social media and survey data and making it available for re-use should, in principle, build on established processes for secure data linkage. Archiving social media data in isolation is not without its challenges (although see Kinder-Kurlanda et al. 2017 for an example of how to archive geotagged Twitter data), however, so it follows that archiving linked survey and Twitter data is even more complex. Variables that are derived by a research team from identifying data (e.g. tweets) could be released as linked data, but these variables may not always match objectives for other researchers. For example, raw tweet content may be coded as supportive of political parties such as "pro-Labour" or "anti-Labour", and this coded linked data could be archived with low risk of disclosing identities. Such variables would be useful for many users, but not every possible variable needed can be known ahead of archiving.

In order to ensure that archiving and re-use is possible and done correctly, consent questions should be worded to ensure participants are aware that data may be archived. Further, security should be maintained and risk of harm minimised through controlling access and data reduction, and following secure data management and deletion protocols appropriate to the nature of the data being archived or accessed. However, there are potentially additional complications within this context. In regards to Twitter (the most used social media data source), the terms of use prevent the sharing of datasets larger than 50,000 Tweets beyond the user (or their research team) who initially access the data. For studies that fall into this category, this would likely mean that raw Twitter data would not be able to be legally archived and shared. However, it is possible to share and archive tweet and user IDs. These can act as 'dehydrated' forms of the data, which can be used by researchers to query the Twitter API and access the raw data, 'rehydrating' it. Indeed, Twitter make special provisions regarding sharing tweet IDs for academics conducting non-commercial research.[9]

One consequence of this approach is that should a user delete their account or a tweet that was part of any initial analysis, it will not be included in the 'rehydrated' dataset. Such a deletion may be seen as a withdrawal of consent, and these cases should be excluded from the dataset. However, for the purposes of replication it means that researchers re-accessing the data may not be working with the same information that the original analysis was based on.

[8]This section comes from work contained in Sloan, L., Jessop, C., Al Baghal, T., & Williams, M. (2019). Linking Survey and Twitter Data: Informed Consent, Disclosure, Security, and Archiving. *Journal of Empirical Research on Human Research Ethics*. https://doi.org/10.1177/1556264619853447

[9] https://developer.twitter.com/en/developer-terms/agreement-and-policy.html

Enabling access to any data requires some level of work for those responsible for curating it, particularly where those data include identifiable information (for example removing data where consent is withdrawn, or setting up access in a secure environment and reviewing outputs taken out of a secure environment for disclosure risk). However, the nature of Twitter (and more generally, social media) data and its analysis creates novel challenges. Depending on the context, the data analysis may require specific software, and many social media analysis tools are web-based. Even if this were not the case, the 'rehydration' of Tweet IDs would require internet access to query the Twitter API. This access to the internet may itself bring security into question. As with all the other aspects discussed herein, there is definite need for further exploration and understanding. However, this report also outlines the possible benefits of linking social media and survey, making any such exploration worthwhile.

References

Al Baghal, T., Sloan, L., Jessop, C., Williams, M., & Burnap, P. (2019). Linking Twitter and Survey Data: The Impact of Survey Mode and Demographics on Consent Rates Across Three UK Studies. Online first at *Social Science Computer Review* DOI: 10.1177/0894439319828011

Al Baghal, T., Nares, Y. & Wenz, A. "Linking Social Media to Survey Responses: Possible Issues and Potential Uses" Presented at the European Survey Research Association, July 2015

Argamon S., Koppel M., Fine J, & Shimoni A.R. (2006) Gender, genre, and writing style in formal written texts. *Text—Interdisciplinary Journal for the Study of Discourse* 23, 321–346.

Barbera, P. (2018). Package 'streamR' for R Statistical Package. https://cran.r-project.org/web/packages/streamR/index.html

Biemer, P. (2011) *Latent Class Analysis of Survey Error*. New York: John Wiley & Sons

Burnap P., Gibson R., Sloan L., Southern R., & Williams M. (2016). 140 characters to victory? Using Twitter to predict the UK 2015 General Election. *Electoral Studies,* 41, 230–233.

Cummings D., Oh H., & Wang N. (2010). Who needs polls? Gauging public opinion from Twitter data. available at http://nlp.stanford.edu/courses/cs224n/2011/reports/nwang6-davidjc-harukioh.pdf, accessed 17/06/2019

Curtis, S. (2013) Twitter claims 15m active users in the UK. *The Telegraph*, 06/09/2013, available at http://www.telegraph.co.uk/technology/twitter/10291360/Twitter-claims-15m-active-users-in-the-UK.html, accessed 17/06/2019

DiGrazia, J., McKelvey, K., Bollen, J., & Rojas, F. (2013) More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior. *PLoS One,* 8(11).

Eady, G., Nagler, J., Guess, A., Zilinsky, J. & Tucker, J.A. (2019) How Many People Live in Political Bubbles on Social Media? Evidence From Linked Survey and Twitter Data. *SAGE Open* January-March 2019: 1–21

Ellison, N.B., Steinfield, C., and Lampe, C. (2007) The Benefits of Facebook 'Friends:' Social Capital and College Students' Use of Online Social Network Sites, J*ournal of Computer-Mediated Communication*, 12, 1143-1168.

Gayo-Avello, D. (2012) I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper: A Balanced Survey on Election Prediction using Twitter Data. Department of Computer Science, University of Oviedo (Spain) Arxiv: http://arxiv.org/pdf/1204.6441.pdf, accessed 17/06/2019

Gentry, J. (2014). Package 'twitteR' for R Statistical Package. http://cran.r-project.org/web/packages/twitteR/index.html

Graham, M., Hale, S.A. & Gaffney, D. (2014) Where in the World Are You? Geolocation and Language Identification in Twitter., *The Professional Geographer*, 66:4, 568-578

Guess, A., Munger, K., Nagler, J., & Tucker, J.A. (2018). How Accurate Are Survey Responses on Social Media and Politics? *Political Communication*, 36, 241-258

Halliday, J. (2013). Facebook: four out of five daily users log on via smartphone or tablet. *The Guardian*, 14/08/2013, available at http://www.theguardian.com/technology/2013/aug/14/facebook-users-smartphone-tablet, accessed 17/06/2019

Jäckle, A. (2009) Dependent interviewing: A framework and application to current research. In *Methodology of Longitudinal Surveys*, in P. Lynn (ed.), 93–112. Chichester, England: John Wiley.

Karlsen, R. & Enjolras, B. (2016) Styles of Social Media Campaigning and Influence in a Hybrid Political Communication System: Linking Candidate Survey Data with Twitter Data. *The International Journal of Press/Politics,* 21, 338–357

Kinder-Kurlanda, K., Weller, K., Zenk-Möltgen, W., Pfeffer, J. & Morstatter, F. (2017). Archiving Information from Geotagged Tweets to Promote Reproducibility and Comparability in Social Media Research. *Big Data & Society*, 1-14.

Knies , G. (ed.) (2018) *Understanding Society*: The UK Household Longitudinal Study Waves 1-8,User Manual. Institute for Social and Economic Research, University of Essex, Colchester.
Lynn, P., Burton, J., Kaminska, O., Knies, G., & Nandi, A. (2012) An Initial Look at Non-Response and Attrition in *Understanding Society Understanding Society Working Paper Series*, No. 2012-02

Murphy, J., Landwehr, J., & Richards, A. (2013). "Using Twitter to Predict Survey Responses." Paper presented at the Midwest Association of Public Opinion Research conference, Nov. 2013

O'Connor, B., Balasubramanyan, R., Routledge, B.R., & Smith, N.A. (2011) From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 122-129

Office of Communications (Ofcom) (2019) *Adults' Media Use and Attitudes Report 2019.* Research Document.

Office of National Statistics (2018) *Internet Access – Households and Individuals 2018*. Statistical Bulletin.

Papaioannou, P., Russo, L., Papaioannou, G., & Siettos, C. (2013). Can social microblogging be used to forecast intraday exchange rates? *NETNOMICS: Economic Research and Electronic Networking*, 14, 47–68

Pasek, J.,Yan, H. Y., Conrad, F. G., Newport, F., & Marken, S. (2018). The Stability of Economic Correlations Over Time: Identifying Conditions Under Which Survey Tracking Polls and Twitter Sentiment Yield Similar Conclusions. *Public Opinion Quarterly,* 82, 470-492

Pennebaker, J.W., Booth, R.J., Boyd, R.L. & Francis, M.E. (2015) Linguistic Inquiry and Word Count: LIWC2015. Austin, TX: Pennebaker Conglomerate

Scanfeld, D., Scanfeld, V., & Larson, E.L. (2010) Dissemination of health information through social networks: Twitter and antibiotics*. American Journal of Infection Control*, 38, 182-188.

Schwartz. H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. & Ungar, L. (2013) Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. PLoS ONE 8:9 e73791

Silge, J. & Robinson, D. (2017) *Text Mining with R: A Tidy Approach*. O'Reilly Media, Inc: Sebastpol, California

Sloan, L. (2017a). Social science 'Lite'? Deriving demographic proxies from Twitter. In: Sloan, L. & Quan-Haase, A. eds. *The SAGE Handbook of Social Media Research Methods*. SAGE, 90-104.

Sloan, L. (2017b). Who tweets in the United Kingdom? Profiling the Twitter population using the British social attitudes survey 2015. *Social Media and Society* 3, 1-11

Sloan, L., Jessop, C., Al Baghal, T., & Williams, M. (2019). Linking Survey and Twitter Data: Informed Consent, Disclosure, Security, and Archiving. *Journal of Empirical Research on Human Research Ethics*. https://doi.org/10.1177/1556264619853447

Sloan, L. Morgan, J, Housley, W., Williams, M.L., Edwards, A.M., Burnap, P. & Rana, O.F. (2013). Knowing the Tweeters: Deriving sociologically relevant demographics from Twitter. *Sociological Research Online,* 18(3) 7.

Sloan, L., Morgan, J, Williams, M.L., Edwards, A.M., & Burnap, P. (2015). Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *Plos One* 10 (3), e0115545.

Sood, G. (2018) Package 'tuber' for R Statistical Package. http://cran.r-project.org/web/packages/tuber/index.html

Steinfield, C., DiMicco, J. M., Ellison, N. B., & Lampe, C. (2009). Bowling online: Social networking and social capital within the organization. *Proceedings of the Fourth International Conference on Communities and Technologies*, 245-254

University of Essex. Institute for Social and Economic Research. (2018). Understanding Society: Innovation Panel, Waves 1-10, 2008-2017. [data collection]. 9th Edition. UK Data Service. SN: 6849 http://doi.org/10.5255/UKDA-SN-6849-10

Williams, M. L., Burnap, P. & Sloan, L. (2016). Crime sensing with big data: the affordances and limitations of using open source communications to estimate crime patterns. *British Journal of Criminology*, 57, 320-340

Wojcik, S. & Hughes, A. (2019). Sizing Up Twitter Users. Pew Research Center Report, April 2019.

Appendix A: Question Wordings for Consent

BSA (2015)

Do you have a personal Twitter account?

Yes
No

IF Yes
We are interested in being able to link people's answers to this survey to the ways in which they use Twitter. We would also like to know who uses Twitter. A research project about who and how people use Twitter is being conducted by a team of researchers at Cardiff University. Are you willing to tell me the name of your personal Twitter account and for this to be passed to researchers at Cardiff University, along with your answers to this survey? Your Twitter name would not be published.

Yes
No

IF Yes
INTERVIEWER: Please enter the respondent's Twitter name here
Open Question (Maximum of 100 characters)

NatCen Panel (July 2017)

Do you have a personal Twitter account?

Yes
No

IF Yes
As social media plays an increasing role in society, we would like to know who uses Twitter, and how people use it. We are also interested in being able to add people's, and specifically, your answers to this survey to publicly available information from your Twitter account such as your profile information, tweets in the past and in future, and information about how you use your account. Your Twitter information will be treated as confidential and given the same protections as your interview data. Your Twitter username, and any information that would allow you to be identified, will not be published without your explicit permission.

Are you willing to tell me your personal Twitter username and for your Twitter information to be added to your answers to this survey?

Yes
No

HELP SCREENS AVAILABLE
HELP SCREEN: <u>What information will you collect from my Twitter account?</u>
We will only collect information from your Twitter account that is publicly available. This will include information from your account (such as your profile description, who you follow, and who follows you), the content of your tweets (including text, images, videos and web links), and background information about your tweets (such as when you tweeted, what type of device you tweeted from, and the location the tweet was sent from).We will

collect information from your past tweets (up to the last 3,000) and will update this with information from more recent tweets on a regular basis.

HELP SCREEN: What will the information be used for?
The information will be used for social research purposes only. Adding your Twitter information and your survey answers will allow researchers from universities, charities and government to better understand your experiences and opinions. For example, using extra information from your Twitter account, researchers can start to:

- Understand who uses Twitter and how they use it
- See what Twitter information can tell us about people, and how accurate it is
- Know what people in the UK are saying about things we don't ask in our survey
- Look at additional information related to questions asked in the survey

HELP SCREEN: Who will be able to access the information?
Matched data which includes both your survey answers and Twitter information will be made available for social research purposes only. Researchers who want to use your matched Twitter and survey information must apply to access it and present a strong scientific case to ensure that the information is used responsibly and safely. Matched statistical information from your Twitter account which you cannot be identified from (e.g. how often you Tweet, or whether you follow any politicians) will have the same access controls as your other survey answers. At no point will any information that would allow you to be identified be made available to the public

HELP SCREEN: What will you do to keep my information safe?
All information we collect will be held in accordance with the Data Protection Act 1998. Because Twitter information is public data that anyone can search, it is impossible to anonymise completely. To keep your information safe, researchers will only be able to access the matched survey answers and detailed Twitter information in a secure environment set up to protect this type of data. Only approved researchers who have gone through special training may access this information, and they will have to apply to do so. Statistical information from your Twitter account which you cannot be identified from (e.g. how often you Tweet, or whether you follow any politicians) will have the same level of protection as your other survey answers.

HELP SCREEN: What if I change my mind?
This information will be collected and stored for as long as they are useful for research purposes, or until you contact us to withdraw your permission. You can do this at any time by emailing us at panel@natcen.ac.uk or calling 0800 652 4569, and do not have to give a reason.
{END OF HELP SCREENS}

IF Yes
What is your Twitter username?

SOFTCHECK: "Twitter usernames must begin with an @ character, followed a maximum of 15 characters (A-Z, a-z, 0-9, underscore), no word spaces. Please check and amend."

Innovation Panel wave 10 (2017)

Do you have a personal Twitter account?

Yes
No

IF Yes
We would like to know who uses Twitter, and how people use it. We are also interested in being able to add people's answers to this survey to publically available information from your Twitter account such as your profile

information, tweet content, and information about how you use your account. Your Twitter information will be treated as confidential and given the same protections as your interview data. Your Twitter username, and any information that would allow you to be identified, will not be published without your explicit permission. Are you willing to tell me the name of your personal Twitter account and for your Twitter information to be linked with your answers to this survey?

Yes
No

HELP SCREENS AVAILABLE
HELP SCREEN: What information will you collect from my Twitter account?
We will only collect information from your Twitter account that is publically available. This will include information from your account (such as your profile description, who you follow, and who follows you), the content of your tweets (including text, images, videos and web links), and background information about your tweets (such as when you tweeted, what type of device you tweeted from, and the location the tweet was sent from). We will collect information from your past tweets (up to the last 3,000) and will update this with information from more recent tweets on a regular basis. This information will be collected and stored for as long as they are useful for research purposes, or until you contact us to withdraw your permission. You can do this at any time, and do not have to give a reason.

HELP SCREEN: What will the information be used for?
The information will be used for social research purposes only. Adding your Twitter information and your survey answers will allow researchers from universities, charities and government to better understand your experiences and opinions. For example, using extra information from your Twitter account, researchers can start to:
* Understand who uses Twitter and how they use it
* See what Twitter information can tell us about people, and how accurate it is
* Know what people in the UK are saying about things we don't ask in our survey
* Look at additional information related to questions asked in the survey

HELP SCREEN: Who will be able to access the information?
Researchers who want to use matched Twitter and survey information must apply to access it and present a strong scientific case to ensure that the information is used responsibly and safely. Matched statistical information from your Twitter account which you cannot be identified from (e.g. how often you Tweet, or whether you follow any politicians) will have the same access controls as your other survey answers.

HELP SCREEN: What will you do to keep my information safe?
Matched statistical information from your Twitter account which you cannot be identified from (e.g. how often you Tweet, or whether you follow any politicians) will have the same level of protection as your other survey answers.

IF Yes

What is your Twitter username (e.g. @usociety)?

Soft check: Twitter username does not being with '@' or contains spaces "Please check and amend. Twitter usernames should begin with an @ character and should not contain any spaces."

Pew Research Center (2018 survey)

We would like to better understand the role of Twitter in society. In order to do that it would be very helpful if you would share your Twitter handle with us. The handle is the username you have selected for your Twitter

account. Handles will be used for research purposes only. We won't use it to contact you and we won't share it with anyone for marketing purposes.

Please list your Twitter handle in the box below.

[TEXT BOX WITH @ IN FRONT OF THE TEXT BOX]