Understanding Society
THE UK HOUSEHOLD LONGITUDINAL STUDY

Archiving and Using Linked Social Media and Survey Data

Tarek Al Baghal

University of Essex

ISER

UKRI Economic and Social Research Council

An initiative by the Economic and Social Research Council, with scientific leadership by the Institute for Social and Economic Research, University of Essex, and survey delivery by NatCen Social Research and Kantar Public

Acknowledgments

ESRC grant: “Understanding [Offline/Online] Society: Linking Surveys with Twitter Data”

- Luke Sloan – University of Cardiff
- Curtis Jessop – NatCen for Social Research
- Matthew Williams – University of Cardiff

National Centre for Social Research

CARDIFF UNIVERSITY

University of Essex

What are we trying to do, and why?

- Link survey participants' answers to publicly available information from their Twitter accounts
- Allows survey data to benefit from real-time, 'natural' behavioural and attitudinal data
- Adds the 'who' to Twitter data – creates a sample frame, and allows for the analysis of different groups
- Complement, not contrast

Social Media (in the UK)

2011: 45% access Internet to use social media

2023: 89% access Internet to use social media

- 99% of 16-24; 98% of 25-34; 95% of 35-44; 92% 45-54

- 69% Facebook
- 51% Instagram
- 36% TikTok
- 29% Twitter
- 26% Snapchat
- 18% LinkedIn

Archiving and Sharing

- Archiving and sharing of data is important:
- Replication of results
- Maximise value of data

- Particular issues:
- Who is responsible for maintaining the data?
- Deleted Tweets/withdrawn consent
  - Multiple consent requests in longitudinal survey?
- Legal issues of sharing Twitter datasets

Data Used

Innovation Panel (IP) Wave 10

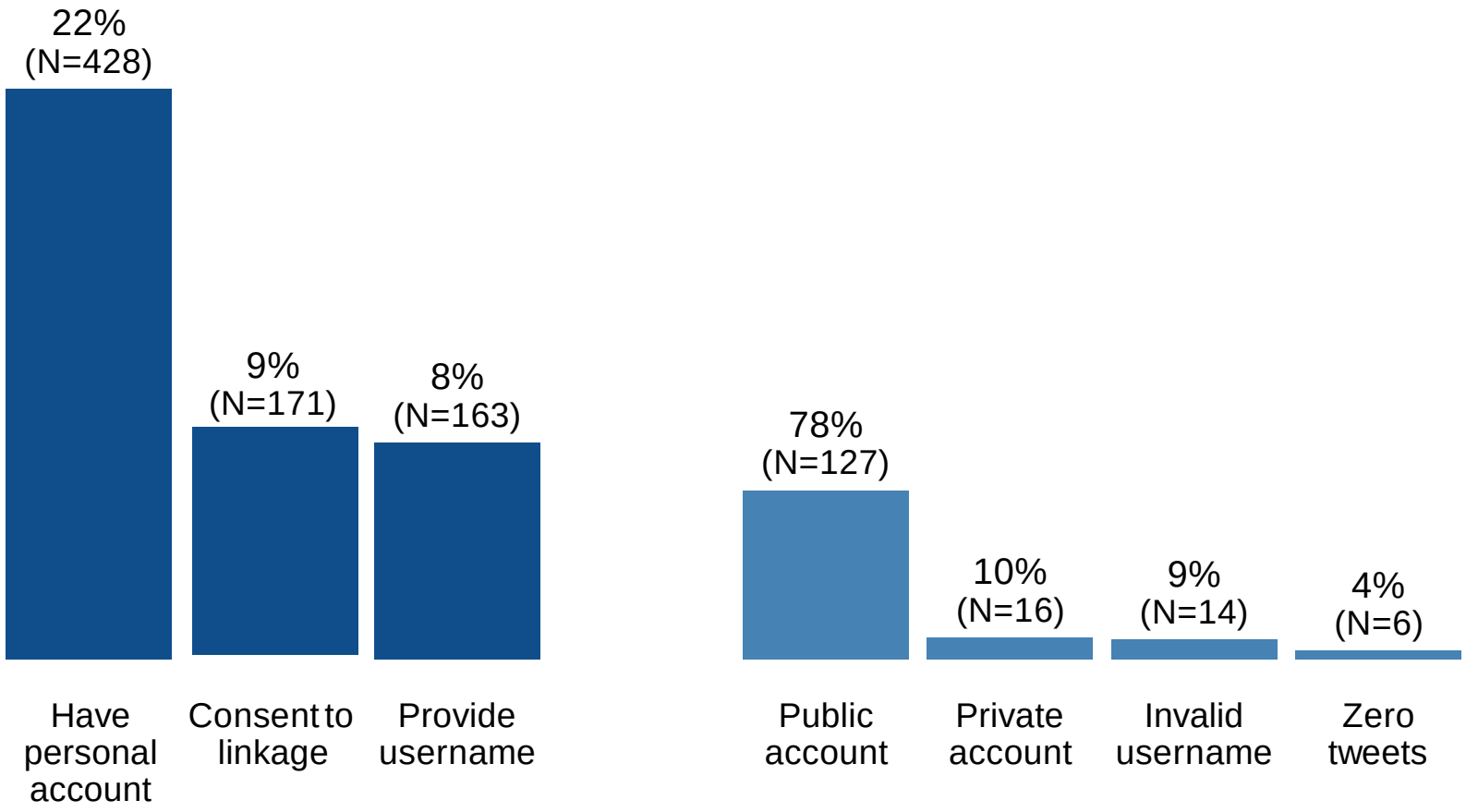
- Part of Understanding Society
- Annual probability panel, focus on experiments
- Fielded Summer/Autumn 2017
- N= 1945
- RR = 52.4%

Tweets collected from June 2007 – February 2023

Part of larger study – linkage asked in 6 other surveys/waves

- Only IP10 used for deposit

Respondent linkage IP10



Total Respondents: N=1,945.

Impact of Data Quantity

- What amount of Twitter data can be collected from respondents in a longitudinal survey?
- Amount can impact capture of signal in the noise
- Increase in variance, reduction in information
- Is there potential bias in substantive analyses?

Amount and respondent characteristics

Regression of total number of tweets (log)

- **Female ⬇**
- **A-level or professional degree ⬆**
- **Number of Twitter followers ⬆**

- Number of Twitter accounts followed ↔
- Frequency of Internet use ↔
- Age ↔
- Ethnicity ↔
- Marital status ↔
- HH income ↔
- Employment status ↔

Two datasets

Platform-based behavior (raw and derived metrics from user-level metadata) [**30 variables**]

Tweet data (raw and derived metrics from tweet-level metadata) [**135 variables**]:

- Tweet raw metadata
- Sentiment Analysis
- Syntactic and Lexical Features
- Readability
- Lexical Diversity
- Complex content: Part-of-Speech tagging

API Provided User Metrics

following - number of accounts the user was following

followers - number of followers of the user's account.

public_list – number of public lists account belongs to

tweets – total number of tweets posted

Tweet Derived Metrics

count_reply - number of replies to a tweet by another user.

count_quote – number of quote of tweets posted by the user.

count_original - number of original content tweets (excludes quoted tweets).

count_retweets - count of retweets by the user.

likes -How many times user's tweet was liked

retweet- How many times user's tweet was retweeted

tweets_prop_activedays - Proportion of days respondent was active on Twitter

User Metrics

Variable	*N*	*Mean*	*Std Dev*
Tweets	146	2512.01	6314.32
Followers	146	228.24	508.49
Following	146	382.58	682.06
Public Lists	146	4.79	17.22

Tweet Derived Metrics

Variable	*N*	*Mean*	*Std Dev*
Likes	127	1753.39	5121.93
Retweets	127	327.50	1079.09
Count Original	127	784.02	3191.11
Count Quote	127	57.42	215.96
Count Reply	127	842.50	1990.78
Count Retweet	127	727.92	2375.46
Prop Active Days	127	0.21	0.26

Respondent Data

Variable	*N*	*Mean*	*Std Dev*
Age	146	37.63	14.67
Female	146	0.52	0.50
University	144	0.53	0.50
Income	146	2290.83	1931.43
Married/Cohabit	145	0.60	0.49
Employed	146	0.80	0.40

Analysis of Linked Data

- Attrition at next wave (IP11), of 146:
  - 115 responded (75.6%) 27 attritted (17.8%) 10 ineligible (6.6%)
- GHQ Wellbeing scale 0-36 (higher = worse) (IP10)
  - N= 144 Mean = 11.3 SD= 5.4
- Use square root of all Twitter count metrics
- And respondent demographics

Results 1

Logistic Regression on Attrition (n=121):

- Nothing significant (at $p<0.05$)!
- Possibly due to small n (100/21 split)
- Partially evidence by lack of significance from demographics

Results 2

GLM on GHQ Wellbeing score (n=123):

- **Number of following** ⬆
- **Number of user retweets** ⬆
- **Female** ⬆

*Higher = Worse on GHQ Scale

- Number of followers ↔
- Number of public lists↔
- Number of original tweets ↔
- Number of quotes ↔
- Number of replies ↔
- Retweets ↔
- Likes ↔
- Days of Activity ↔
- Age ↔
- Education ↔
- Income ** ↔
- Marital status ↔

Deposit

- Reviewed by data security experts to ensure minimized risks
- Created code book on how to use
- Data processed using Understanding Society procedures
- *Understanding Society: Innovation Panel Twitter Study, 2007-2023*
- https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=9208
- Open access to researchers to link to the longitudinal data

Conclusion

- Some evidence for social media data impact
  - Perhaps more use on measurement side?
- This is a framework/jumping off point
- Expand to new social media (working on LinkedIn)
- Twitter (X) now limits/charges but:
  - Can still get some variables for free:
    - followers, following, tweet count, twitter creation time, twitter bio information, geolocation for account, whether account protected/suspended/exist, display name.
  - Using tweepy (or similar) on free API

Tweet-level Sentiment Analysis

Sentiment Analysis	
sentimentr_jockers_rinker_b	Average sentiment score for sentences in the tweet using the combined and augmented version of Jockers (2017) & Rinker'saugmented Hu & Liu (2004) positive/negative word list as sentiment lookup values, ie dictionary of positive/negative word list.
sentimentr_jockers_b	Average sentiment score for sentences in the tweet using a modified version of Jockers (2017) sentiment lookup table used in szuhet R package. Sentiment values ranging between -1 and 1.
sentimentr_huliu_b	Average sentiment score for sentences in the tweet using an augmented version of Hu & Liu's (2004) positive/negative wordlist as sentiment lookup values. Sentiment values ranging between -2 and +1.

Tweet-level Lexical analysis

Syntactic and Lexical Features	
chars	Count of characters per tweet.
sents	Count of sentences in the tweet.
tokens	Count of tokens (words) per tweet.

Lexical Diversity	
C	Herdan's C (Herdan, 1960, as cited in Tweedie & Baayen, 1998; sometimes referred to as LogTTR)
R	Guiraud's Root TTR (Guiraud, 1954, as cited in Tweedie & Baayen, 1998)
TTR	The ordinary Type-Token Ratio

References

Al Baghal, T., Wenz, A., Serodio, P., Liu, S., Jessop, C., and Sloan, L. (2024). Linking Survey and LinkedIn Data: Understanding Usage and Consent Patterns, *Journal of Survey Statistics and Methodology*, smae029.

Liu, S., Sloan, L., Al Baghal, T., Williams, M., Serôdio, P. and Jessop, C. (2024). Examining household effects on individual Twitter adoption: A multilevel analysis based on U.K. household survey data. *PLoS One*

Liu, S., Sloan, L., Al Baghal, T., Williams, M., Jessop, C., and Serôdio, P.(2024). Linking Survey with Twitter Data: Examining Associations among Smartphone Usage, Privacy Concern and Twitter Linkage Consent. *International Journal of Social Research Methodology*

Tanner, A.R., Di Cara, N.H., Maggio, V., Thomas, R., Boyd, A., Sloan, L., Al Baghal, T., MacLeod, J., Haworth, CMA, Davis, OSP (2023). Epicosm—a framework for linking online social media in epidemiological cohorts. *International Journal of Epidemiology:* 952–957

Sloan, L. Al Baghal, T, and Jessop, C. (2022) Linking Survey and Twitter Data: Informed Consent, Disclosure and Data Security. In L. Sloan, and A. Quan-Haase (eds.) *SAGE Handbook of Social Media Research Methods (2nd Ed)* p. 691-702.

References

Al Baghal, T., Wenz, A., Sloan, L., and Jessop, C. (2021). Linking Twitter and Survey Data: Quantity and Possible Biases. *EPJ Data Science*, 10:32.

Breuer, J., Al Baghal, T., Sloan, L., Bishop, L. Kondyli, D., and Linardis, A. (2021). Informed consent for linking survey and social media data: Differences between platforms and data types. *IASSIST Quarterly*, 45(1),

Di Cara, N.H., Boyd, A., Tanner, A.R., Al Baghal, T., Calderwood, L., Sloan, L.S., Davis, O.S.P., and Haworth, C.M.A. (2020). Views on social media and its linkage to longitudinal data from two generations of a UK cohort study. *Wellcome Open Research*, 5:44

Sloan, L., Jessop, C., Al Baghal, T., and Williams, M. (2020). Linking Survey and Twitter Data: Informed Consent, Disclosure, Security, and Archiving. *Journal of Empirical Research on Human Research Ethics*, 15:63-76

Al Baghal, T., Sloan, L., Jessop, C., Williams, M., and Burnap, P. (2020). Linking Twitter and Survey Data: The Impact of Survey Mode and Demographics on Consent Rates Across Three UK Studies. *Social Science Computer Review*, 38: 517-532