Understanding Society

THE UK HOUSEHOLD LONGITUDINAL STUDY

Using social media metrics and linked survey data to understand survey behaviors

Tarek Al Baghal- University of Essex
Paulo Serodio – University of Essex
Shujun Liu – Cardiff University
Luke Sloan – Cardiff University
Curtis Jessop – NatCen Social Research

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

Social Media (in the UK)

2011: 45% access Internet to use social media

2020: 70% access Internet to use social media

- 97% of 16-24; 91% of 25-34; 90% of 35-44

- ~90% Facebook
- ~65% Whatsapp
- ~40% Instagram
- ~25% Twitter
- ~15-25% LinkedIn

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

Understanding Survey Outcomes

Continual, ongoing past attrition

Can we use to trace or weight?

Understanding survey measurements

Either methodological or substantive

But limited to specific subgroup

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- Deleted Tweets/withdrawn consent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

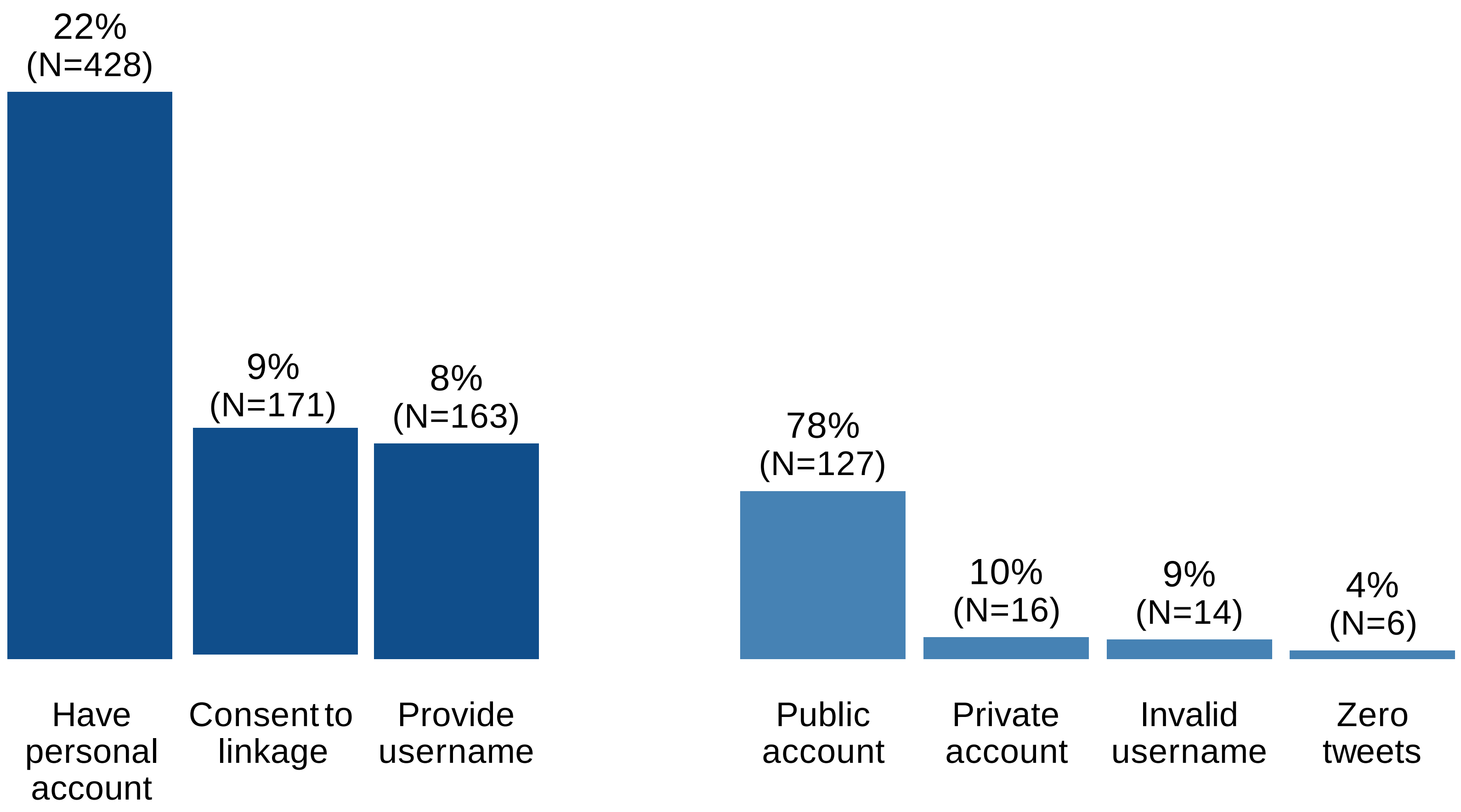
- Part of Understanding Society
- Annual probability panel, focus on experiments
- Fielded Summer/Autumn 2017
- N= 1945
- RR = 52.4%

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.

Two datasets

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

Tweet metadata (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.

Tweet Derived Metrics (2)

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

- Attrition at next wave (IP11), of 146:
  - 115 responded (75.6%)
  - 27 attritted (17.8%)
  - 10 ineligible (6.6%)
- Use square root of all Twitter count metrics
- And respondent demographics

Attrition Results

Logistic Regression on Attrition (n=121):

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

Analysis of Linked Data -Wellbeing

- 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

Well-being Results

GLM on GHQ Wellbeing score (n=123):

*Higher = Worse on GHQ Scale

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

- 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
- Deposit to the UK Data Archive (Study 9208)

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
- 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