Investigating Demographic Representation on Twitter

The webinar will begin at 3pm

- You now have a menu in the top right corner of your screen.
- The red button with a white arrow allows you to expand and contract the webinar menu, in which you can write questions/comments.
- We won't have time to answer questions while we are presenting, but will answer them at the end
- You will be on mute throughout – we can't hear you.

UK Data Service

Investigating Demographic Representation on Twitter

Webinar

6 September 2016

Luke Sloan and Margherita Ceraolo

UK Data Service

Can you hear us?

UK Data Service

Can you hear us?

- If Not:
  - Check your volume, and that your speaker/headset is plugged in.
  - Your invitation also included a phone number, you can call that to listen in.
    - UK +44 (0) 20 3713 5012
    - US +1 (415) 930-5229
  - We are recording this webinar, so you can always listen to it later.

SOCIAL DATA SCIENCE LAB.

Social Science 'Lite'? Understanding Who Uses Twitter

Luke Sloan (@DrLukeSloan)
Deputy Director
Social Data Science Lab

Cardiff School of Social Sciences

Cardiff University

tweet: @socdatalab
web: socialdatalab.net

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

About Me

- My research focuses on Twitter and how social media data can be used to understanding social phenomenon…
  - Predicting the UK General Election 2015 (Burnap, P.et al. 2016. 140 characters to victory?: Using Twitter to predict the UK 2015 General Election. Electoral Studies (10.1016/j.electstud.2015.11.017)
  - Exploring the relationship between reported crime, Census data and Twitter mentions of low-level disorder in London (Co-Investigator, NCRM Methodological Innovation Project)
  - Investigating information propagation on Twitter following the Horsemeat food scandal (Co-Investigator, ESRC/FSA)
  - Understanding voter and candidate behaviour on Twitter for the Welsh Assembly Election 2016 (Co-Investigator)
- However, all of these projects could be enhanced by knowing who uses Twitter…

SOCIAL DATA SCIENCE LAB.

Demographics

- Social scientists are interested in group differences (gender, age, ethnicity etc)
- Comparative method (groups relative to each other)… but how to identify these groups on social media?
- User generated content can be ‘data light’ (Mislove et al. 2011, Gayo-Avello 2012)
- Facebook is different because it stores baseline demographic information (Schwartz et al. 2013)
- Twitter has signatures, but nothing systematic (Edwards et al. 2013)
- When the data is not available we develop proxies, so why not for Twitter?

SOCIAL DATA SCIENCE LAB.

Demographics

- What insights do demographic proxies offer for behaviour on Twitter?
- Does Twitter behaviour differ by demographic groups?
- Do real-world demographic differences manifest in the virtual world?

SOCIAL DATA SCIENCE LAB.

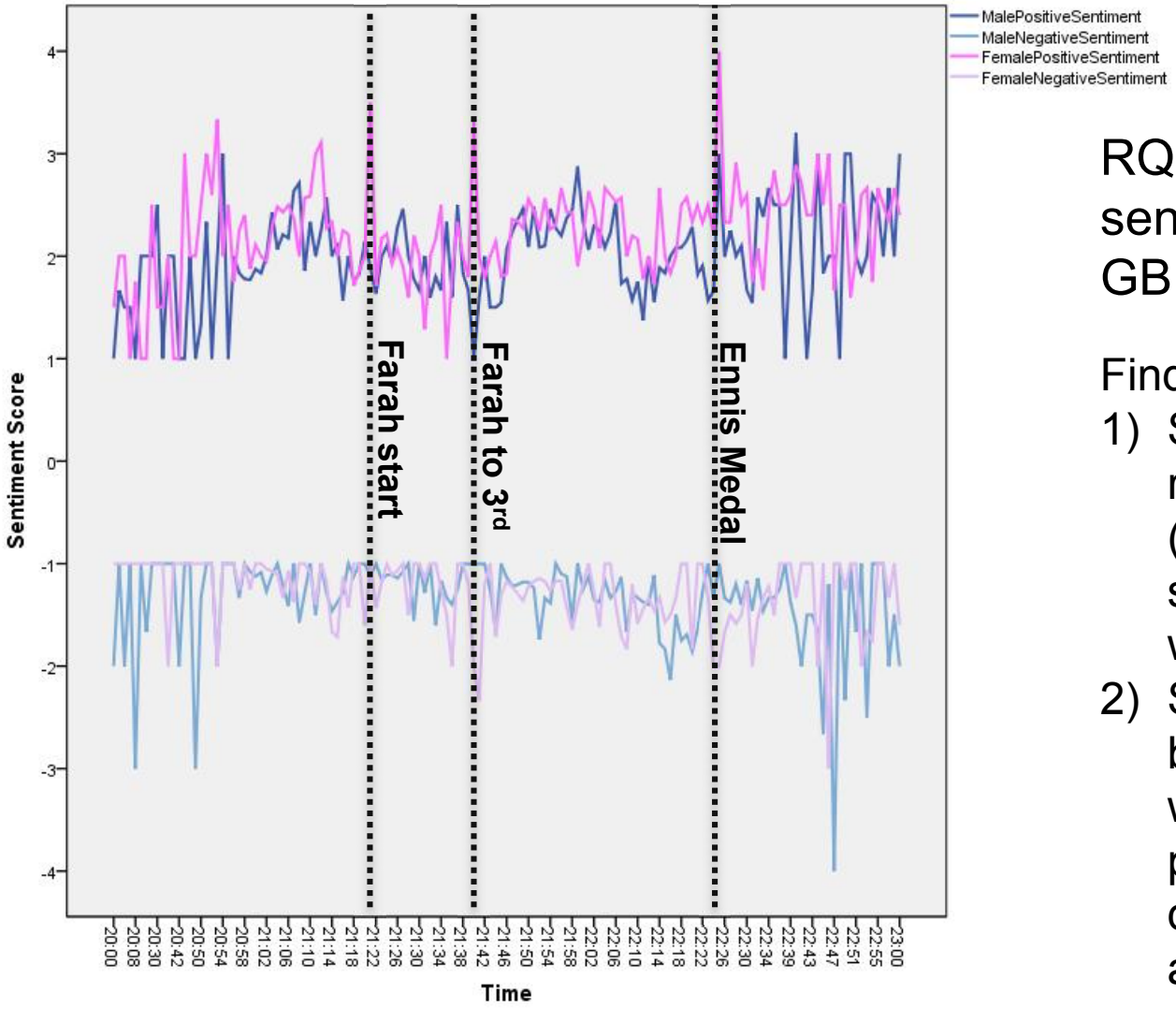
Demographics: Gender

- Use the name field of the Twitter profile (for UK users)
- Clean the data to extract a first name and compare against a large database of first names
- Important to categorise ‘unisex’ and ‘unknown’
- Of those we could identify: 48.8% male and 51.2% female… very close to 49.1% and 50.9% split in the 2011 Census

SOCIAL DATA SCIENCE LAB.

Sloan, L. et al. 2013. Knowing the Tweeters: Deriving sociologically relevant demographics from Twitter. Sociological Research Online 18(3), article number: 7. (10.5153/sro.3001)

Demographics: Gender



RQ: How does sentiment towards Team GB differ by gender?

Findings:

1) Sentiment peaks reflect real world events (relationship between social media and real world)
2) Sentiment differs between men and women (difference is so pronounced that gender detection method appears to work)

Demographics: Location

- Three primary sources of location:
  - User profile information
  - Content of tweets (inc. ‘mundane geography’)
  - Geo-tagged tweets

- Geo-tagged tweets are the gold standard

- Allows us to locate people at the time they tweeted in existing geographies (output area level!)

- RQ: do people tweet about crime in high crime areas? See: Williams, Burnap & Sloan 2016

SOCIAL DATA SCIENCE LAB.

Sloan, L. et al. 2013. Knowing the Tweeters: Deriving sociologically relevant demographics from Twitter. Sociological Research Online 18(3), article number: 7. (10.5153/sro.3001)

Demographics: Location

SOCIAL DATA SCIENCE LAB.

Source: Sloan et al. 2013

Demographics: Location

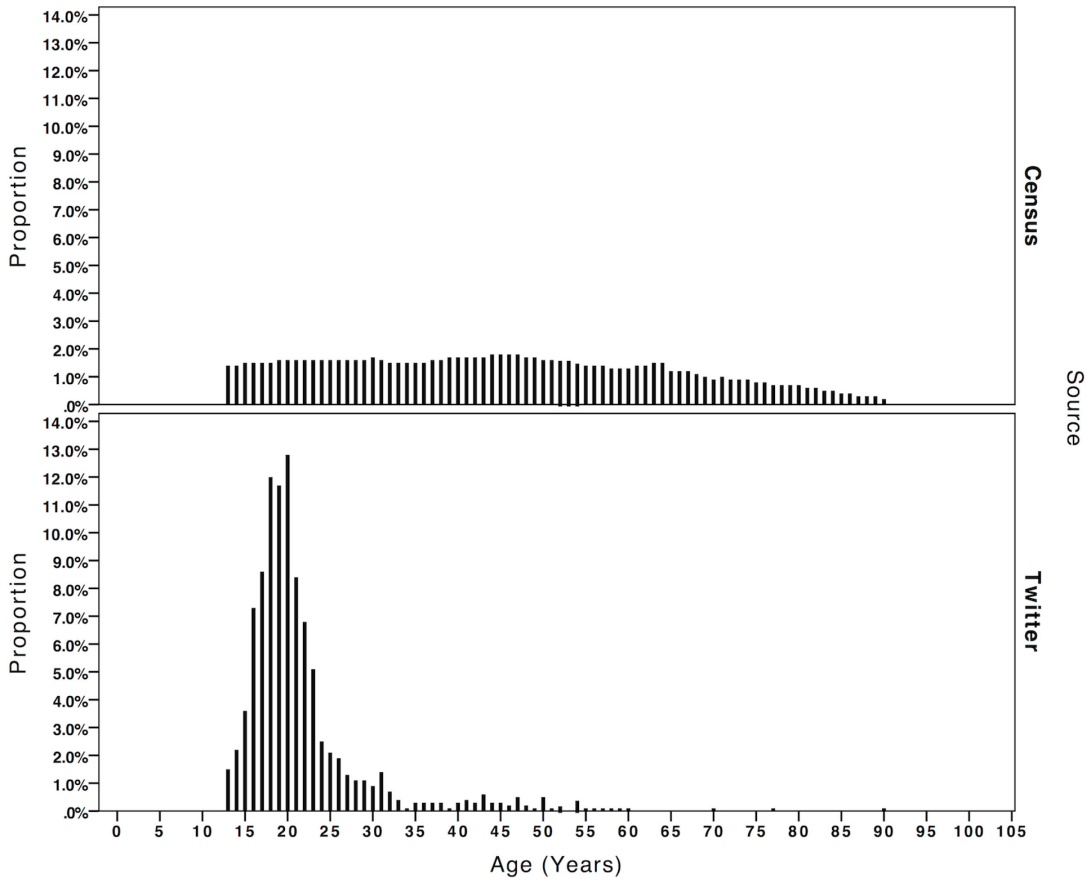
- However, recent research suggests that users who geo-tag tweets are not representative of the Twitter population:
  - Male users more likely to geo-tag
  - Geo-taggers tend to be older
  - Occupational group has an impact (NS-SEC)
  - Geo-taggers have different user interface languages
  - Geo-taggers tweet in different languages
- The differences are sometimes small but always significant

SOCIAL DATA SCIENCE LAB.

Sloan, L. & Morgan, J. (2015) Who Tweets with Their Location?: Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. PLOS ONE 10(11): e0142209. doi:10.1371/journal.pone.0142209

Demographics: Age

- Identifying age from signature data
- Preliminary analysis suggests usable age data for 0.35% of Twitter users
- Note that 0.35% of 645m is 2.25m (approx 40% of which is English language)

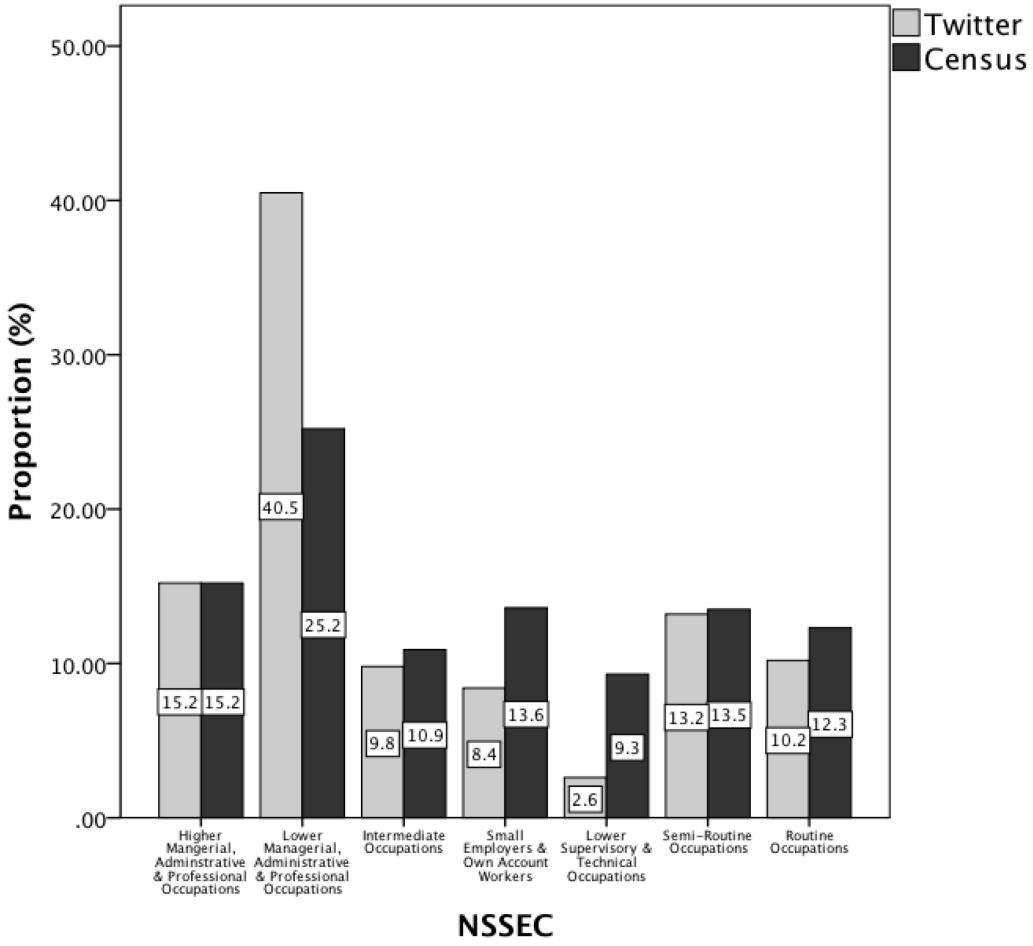


SOCIAL DATA SCIENCE LAB.

Sloan, L. et al. (2015) Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data. PLOS ONE 10(3): e0115545. doi:10.1371/journal.pone.0115545

Demographics: Occupation

- Identify occupation from signature data
- Linked to SOC2010 codes
- Enables allocation into NS-SEC groups



SOCIAL DATA SCIENCE LAB.

Sloan, L. et al. (2015) Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data. PLOS ONE 10(3): e0115545. doi:10.1371/journal.pone.0115545

References

Edwards et al. (2013) Computational social science and methodological innovation: surrogacy, augmentation or reorientation?, *International Journal of Social Research Methods*, 16:3

Gayo-Avello (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

Mislove et al. (2011) Understanding the demographics of Twitter users, *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*

Schwartz et al. (201) Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach, *PLOS ONE*, 8:9 (DOI: 10.1371/journal.pone.0073791)

Sloan et al. (2013) Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter, *Sociological Research Online*, 18:3 (http://www.socresonline.org.uk/18/3/7.html)

Sloan et al. (2014) Going Viral in Social Media – Networks and Intercepted Misinformation, *Software Sustainability Institute*, Cardiff University

Sloan et al. (2015) Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data. PLOS ONE 10(3): e0115545. doi:10.1371/journal.pone.0115545

Sloan, L. & Morgan, J. (2015) Who Tweets with Their Location?: Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. PLOS ONE 10(11): e0142209. doi:10.1371/journal.pone.0142209

Williams, M., Burnap, P. and 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* , pp. 0-19.

SOCIAL DATA SCIENCE **LAB.**

Out later this year: Sloan & Quan-Haase (*Dec 2016*)

SAGE Handbook of Social Media Research Methods