Who consents to sharing their tweets with researchers? A comparative analysis of selection bias in linked survey and social media data

Presentation to the DIGISURVOR Workshop

Dr. Conor Gaughan
Postdoctoral Research Associate
Cathie Marsh Institute for Social Research (CMI)
University of Manchester

✉ @conor.gaughan@manchester.ac.uk

Prof. Alexandru Cernat
Professor in Social Statistics
Cathie Marsh Institute for Social Research (CMI)
University of Manchester

✉ alexandru.cernat@manchester.ac.uk

With: Rachel Gibson, Marta Cantijoch, Riza Batista-Navarro

Linking Survey Data with Digital Data

- Traditionally, survey data is considered the “gold standard” in quantitative social and political research

- This has often meant relying almost exclusively on self-reported measures when seeking to understand important issues relating to human attitudes, behaviours and social networks.

- However, driven by the digital revolution and explosion of new data sources, survey research is entering a new era of development that centres on its linkage with other forms of observed or digitally generated data

- One of the fastest growing areas of this new research agenda within social science is the augmentation of survey data with digital trace data (DTD) (Guess et al., 2019; Eady et al., 2019; Stier et al. 2020a, 2020b; Silber et al., 2022).

What is Digital Trace Data (DTD)?

- Digital trace data (DTD) refers to the data we leave behind as a result of our interactions with various digital technologies

- The digital revolution has led an explosion in new data sources we can use to study human attitudes, behaviours and social networks

- This can include:

- **Social Media Activity** (Likes, shares, comments, follows, posts on platforms like Facebook, Twitter/X, and Instagram)
- **Web Browsing Data** (Search queries on search engines like Google or Bing)
- **Mobile and App Usage** (GPS location tracking, app usage logs, or push notification interactions)
- **Online Transactions** (Online purchase history, cart abandonment data, reviews and ratings)
- **Communication Data** (Emails, chat logs, call records, SMS)
- **Sensor Data** (Smart home devices or wearable trackers like smart watches and Fitbits)
- **Streaming and Media Consumption** (Viewing history on Netflix, YouTube or Prime, listening history on Spotify etc.)

All images free for use under the Pixabay Content License

SURVEY DATA

- ⊕ Representative Sampling
- ⊕ Tailored Questions
- ⊕ Anonymity
- ⊕ Self-Reported Information

- ⊖ Sampling Biases
- ⊖ Measurement Biases
- ⊖ One-Dimensional

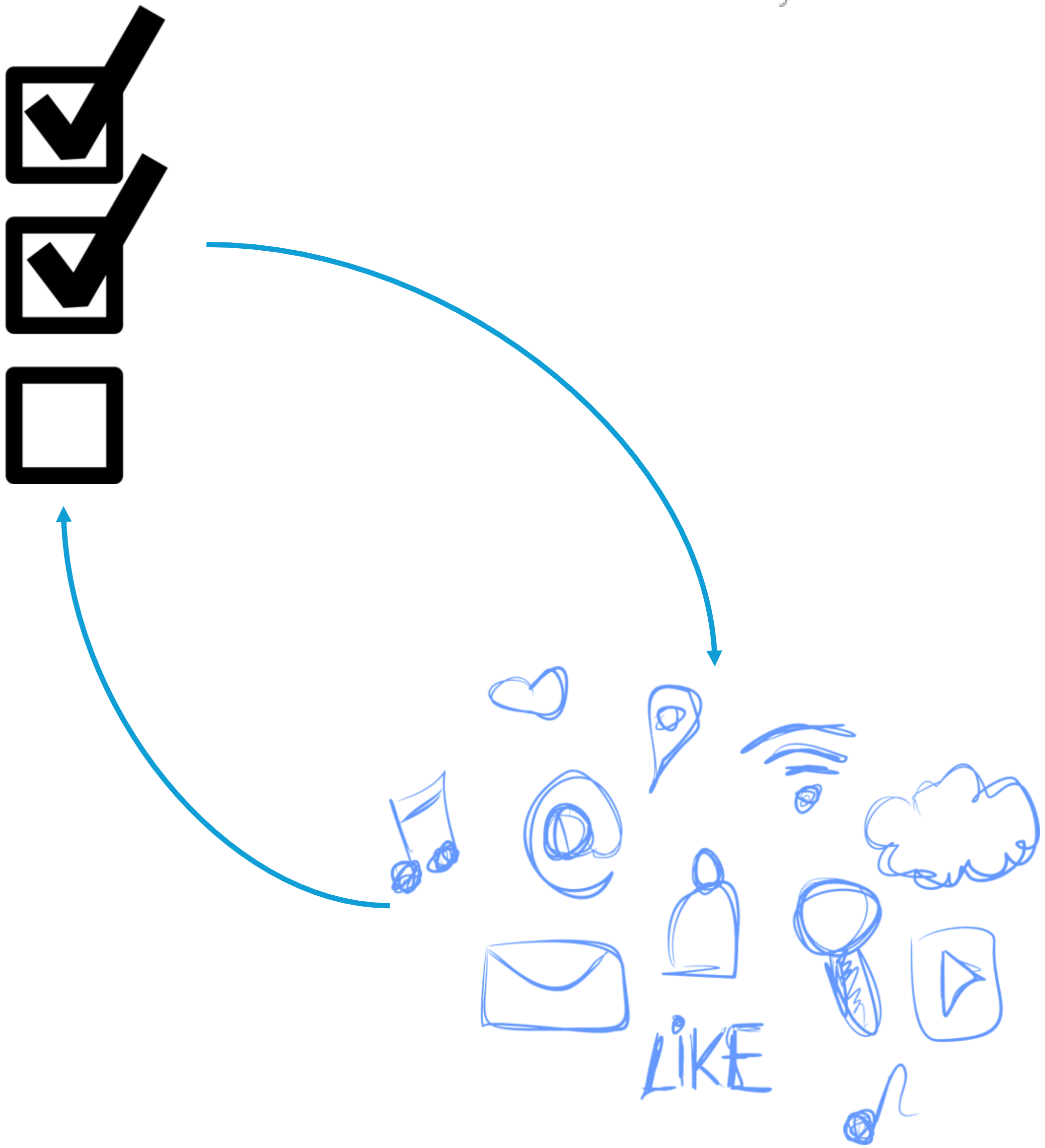
DIGITAL TRACE DATA

- ⊖ Non-Representative
- ⊖ Unstructured
- ⊖ Highly Sensitive

- ⊕ Non-Reactive
- ⊕ Real-Time Data
- ⊕ Unlock New Insights
- ⊕ Publicly Accessible (Sometimes....)

How Can Combining Survey and DTD Improve Social Science Research?

- Survey research is entering a new era of development which centres on its linkage with external data sources such as participant DTD (Groves, 2011; Stier et al., 2020)
- As discussed, both forms of data can be informative sources of information, but each come with their own drawbacks
- Early research has shown that linking DTD to surveys can help to address measurement bias in the survey data, using participant DTD to calibrate and verify their responses (Al Baghal et al., 2020; Cernat et al., 2024; Guess et al., 2019a; Haenschen, 2020; Henderson et al., 2021; Murphy et al., 2019)
- Other research has shown how linking survey data to DTD can help to add a "human face" to the DTD, allowing us to better understand how both the usage and impact of digital technologies vary within the population (Guess et al., 2019b; Guess et al., 2023)
- **However, linking the two data sources can introduce entirely new biases into our data which we need to be aware of**



Selection Bias in Social Media Data

- One of the major issues with using social media data in research is in the **selectivity** of social media users

- Social media users are usually **not representative** of the general population

- There are also selection biases that exist in usage **across** different social media platforms

- This is further compounded when linking social media data with surveys, where **certain people are more likely to consent** to linkage than other

Linked Survey-to-Twitter/X Data

- Six linked survey-to-Twitter/X datasets were collected as part of a previous ERC funded grant project across a four-year period (2020-2024) in five countries during six national election campaigns:

1. **United States 2020:** Presidential Election
2. **Germany 2021**: Parliamentary Election
3. **France 2022:** Presidential Election
4. **Poland 2023:** Parliamentary Election
5. **United Kingdom 2024:** Parliamentary Election
6. **United States 2024:** Presidential Election

- Survey samples ranged from between 5,000-6,000 respondents who were asked a series of questions including standard sociodemographic characteristics as well as political attitudes, behaviours and online activity

- They were also asked which social media platforms they used and were invited to share their Twitter/X handles with the research team if they had an account so that their data could be collected

Research Questions:

RQ1: What are the **selection biases** associated with the **usage of X** and **consent to data linkage**?

RQ2: What are the **differences between countries** in the selection into X usage and consent to data linkage

RQ3: How does selection into data linkage **change over time**?

Two Binary Dependent Variables:

DV1: Uses X? [Yes:1, No: 0]

*Conditional on **DV1**:*

DV2: Consented to Linkage? [Yes: 1, No: 0]

Independent Variables:

Sociodemographic:

- Age [18-25, 30-44, 45-64, 65+]
- Gender [F, M]
- Education Level [Low, Medium, High]
- Employment Status [Employed, Not Employed]
- Household Income [Low, Medium, High]

Attitudinal/Behavioural:

- Political Attention [Low, Medium, High]
- Ideology [Liberal, Moderate, Conservative]
- Privacy Concern [Low, Medium, High]
- General Trust [Low, Medium, High]
- Trust in Social Media Companies [Low, Medium, High]
- Frequency of Political Posts (X) [Daily, Monthly/Weekly, Yearly, Never]
- Frequency of Political Replies (X) [Daily, Monthly/Weekly, Yearly, Never]

Country: Time	**N**	**Uses X**	**Consented**
		(% of N)	**(% of Users)**
FR: 2022	4,736	1,092 (23%)	821 (75%)
GR: 2021	4,035	657 (16%)	490 (75%)
PL: 2023	4,118	679 (16%)	524 (77%)
UK: 2024	4,033	1,551 (38%)	1,103 (71%)
US: 2020	5,377	2,127 (40%)	1,547 (73%)
US: 2024	4,457	1,787 (40%)	1,138 (64%)
Overall	**26,756**	**7,891 (29%)**	**5,632 (71%)**

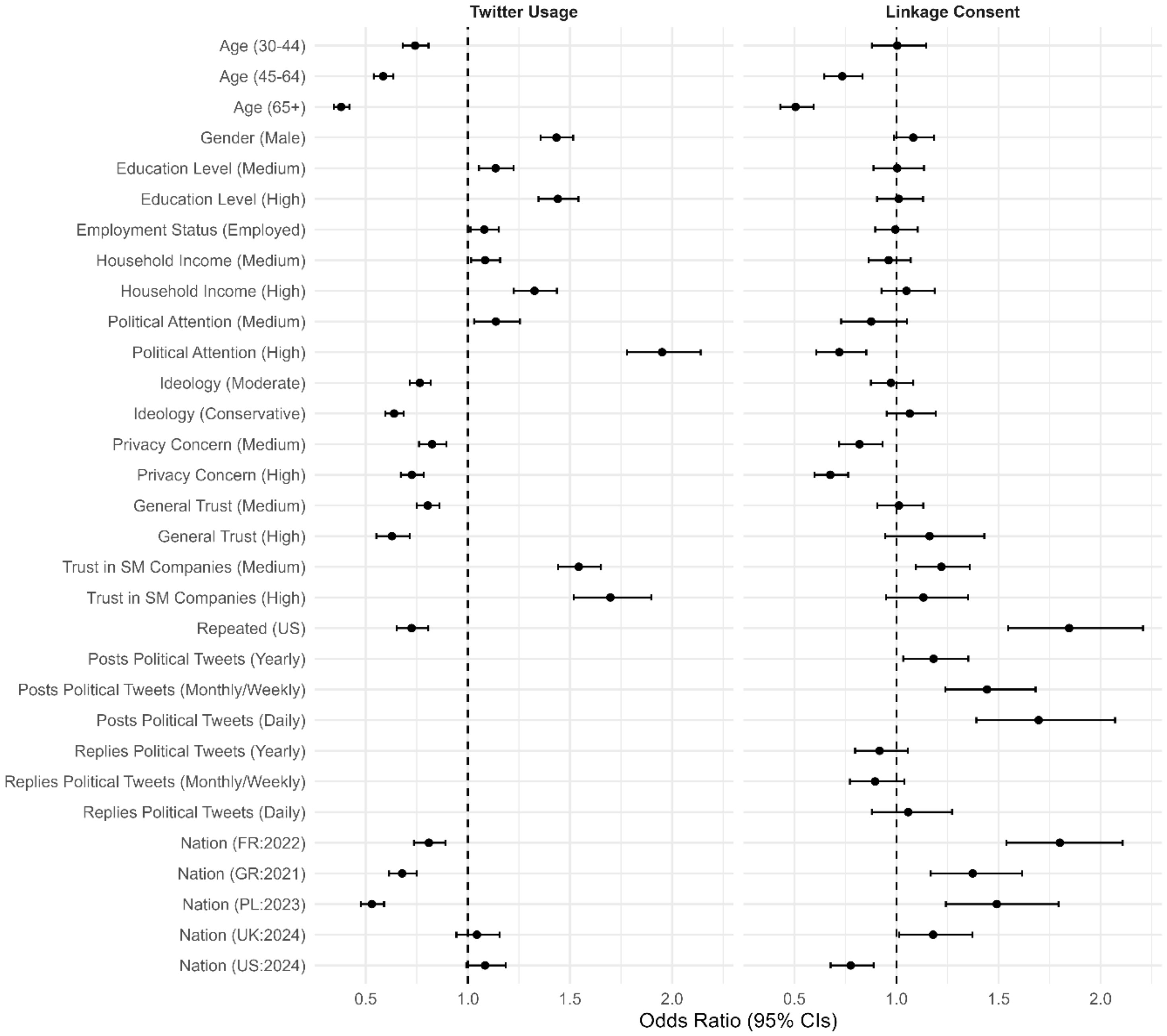
Three Logistic Regression Models:

RQ1: Pooled data with all countries and years

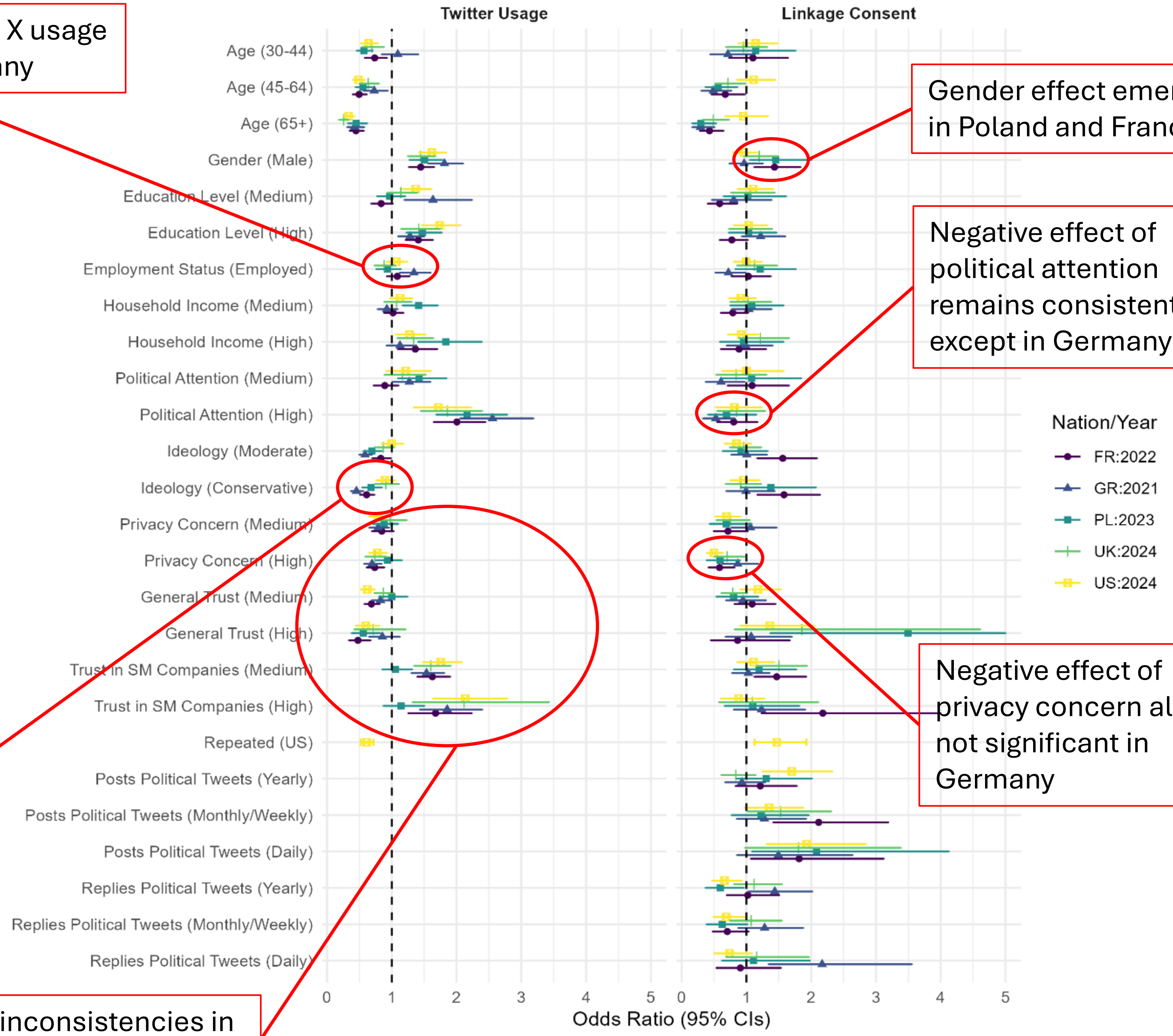
RQ2: Separate models **for each country** (US 2024)

RQ3: Separate models for **US 2020 vs 2024**

RQ1: Overall Bias



RQ2: Comparative Differences



Effect of employment on X usage only significant in Germany

Gender effect emerges in Poland and France

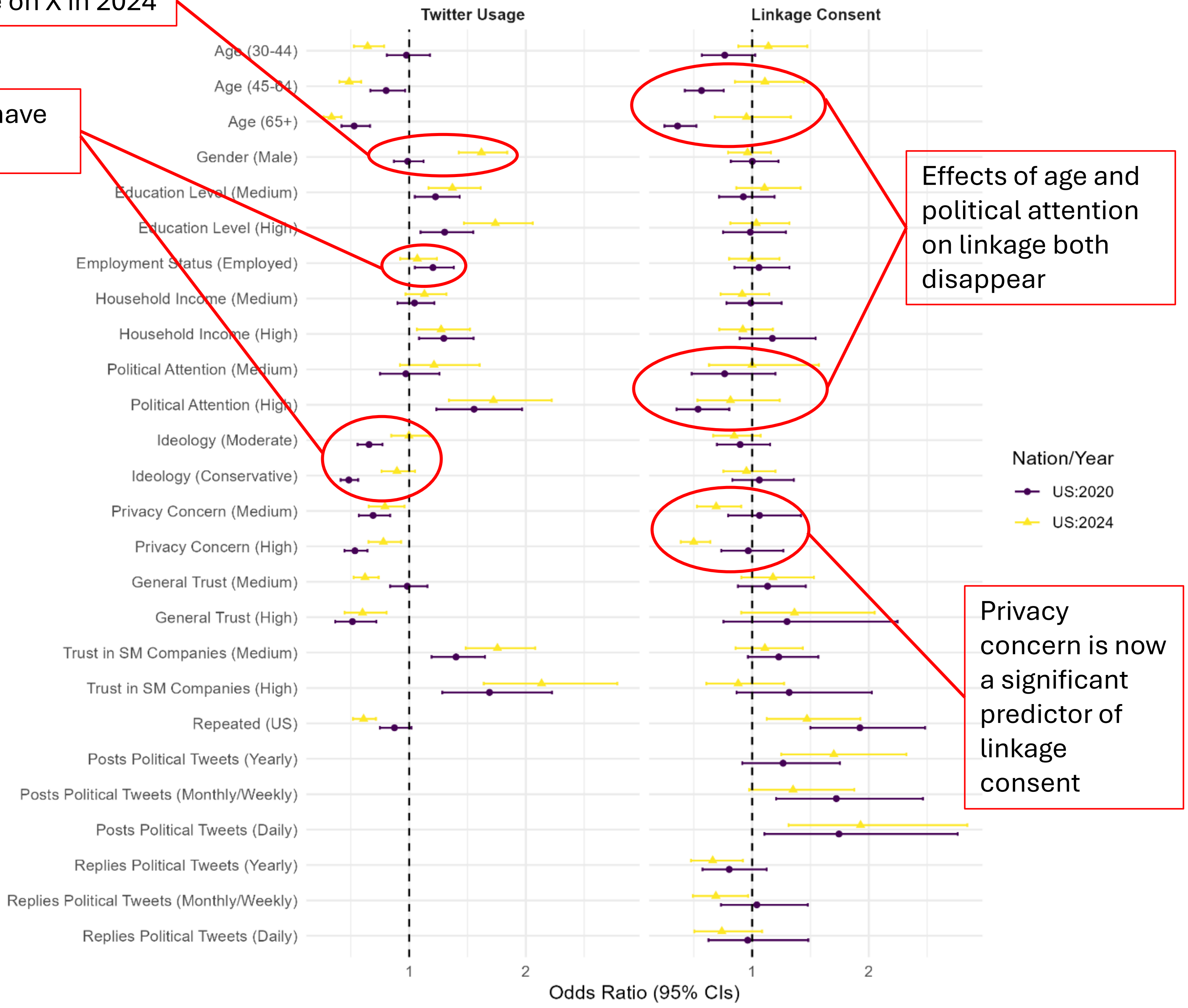
Negative effect of political attention remains consistent except in Germany

Negative effect of privacy concern also not significant in Germany

Liberal bias in X usage only significant in Poland, Germany and France

Further inconsistencies in privacy concern and trust

RQ3: Time Differences



Males more likely to be on X in 2024

Employment and Liberal bias have both disappeared

Effects of age and political attention on linkage both disappear

Privacy concern is now a significant predictor of linkage consent

Selection Biases When Using X Data

X Users

1. On average, X users are **younger, wealthier**, more **highly educated**, more likely to be **employed**, and more likely to be **male**

2. They also pay significantly more **attention to politics** and are more likely to lean **liberal** ideologically

3. They have **lower levels of general trust** but **higher levels of trust in social media companies** and **less concern for data privacy**

4. **Ideological**, **employment**, **privacy** and **trust** effects all vary between countries

5. **Ideological**, **employment** and **gender** effects all vary between countries

Linkage Consent

1. On average, X users who consent to linkage are **younger**, **pay less attention to politics** and demonstrate **lower concerns for data privacy**

2. **Higher levels of political posting** on the platform is also positively associated with likelihood to consent

3. Gender effects emerge in Poland and France where **males were more likely** to consent

4. The negative effects of **political attention** and **privacy concern** on linkage consent not significant in Germany

5. Effects of **age** and **political attention disappear** between 2020 and 2024 in the US

6. At the same time, negative effect of **privacy concern** has emerged

Conclusions

- Those that use X and consent to linking their data with surveys **differ significantly** from a general population sample

- There is also **significant variation** in effects **across countries** and **time,** so caution is needed when comparing groups

- This also has implications for models or algorithms developed in one **geographic** or **temporal** context may not be transferable

- However, one of the strengths of linking survey and social media data makes it possible to **estimate and correct for these biases**

UK Research and Innovation

Smart Data Research UK

MANCHESTER 1824
The University of Manchester

Thank You!

Dr. Conor Gaughan
Postdoctoral Research Associate
Cathie Marsh Institute for Social Research (CMI)
University of Manchester

✉ @conor.gaughan@manchester.ac.uk

Prof. Alexandru Cernat
Professor in Social Statistics
Cathie Marsh Institute for Social Research (CMI)
University of Manchester

✉ alexandru.cernat@manchester.ac.uk

With: Rachel Gibson, Marta Cantijoch, Riza Batista-Navarro