Linking Survey and Social Media Data: Bridging the Gap Between Data Protection and Open Research

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

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

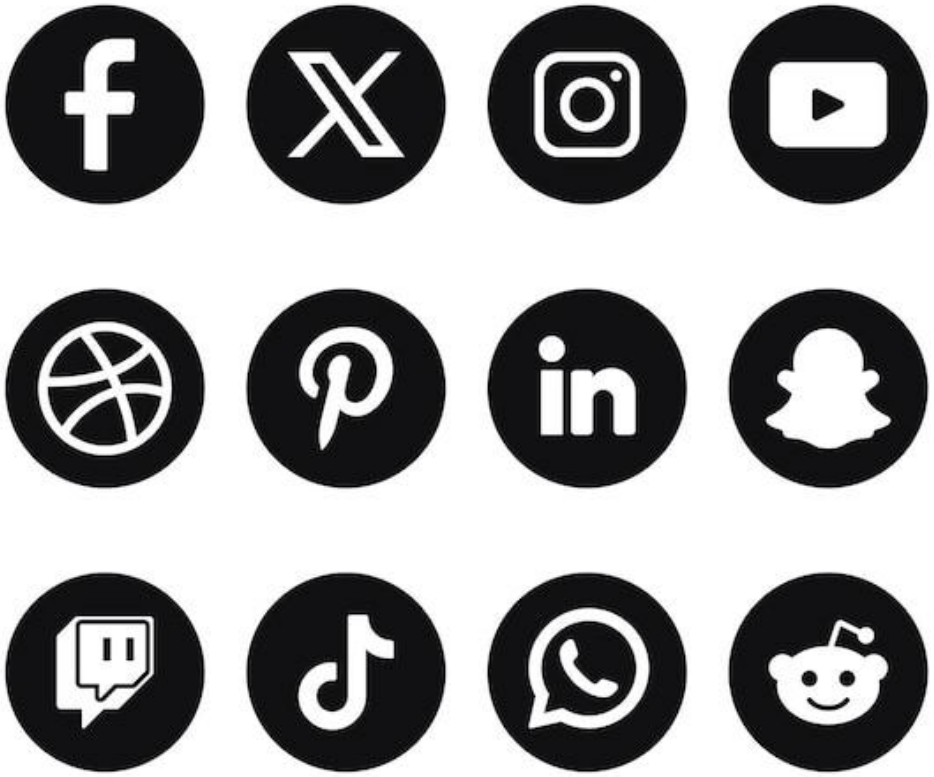
Standalone Social Media Data in Research

- The use of social media data in research is now pervasive across many disciplines (As of 2020, there were over 110,000 research publications with the term “social media” in their title, Aichner et al. 2021)

- Using SMD can offer several strengths over conventional data sources:

1) Non-reactive
2) Real-time analytics
3) Publicly accessible (sometimes…)
4) Vast and multimodal
5) Unlock new insights

- A methodological testbed (comp sci), observational space (social sci, marketing), recruitment tool (survey research), or study object in and of itself



Working with Social Media Data (ELPs)

Ethical

- How do we ensure voluntary participation and informed consent?
- How do we protect anonymity and confidentiality?
- How do we protect participants from the potential for harm?

Legal

- What are the relevant laws relating to the collection, use, archiving and sharing of social media data?
- What are the relevant national and international data protection legislation?
- EU/UK GDPR, California's CCPA, Canada's PIPEDA, Kenya's DPA, Sweden's DA etc.

Platform-**s**pecific

- What are the specific Terms of Service for developers on each social media platform?
- What are their rules on collection, use, archiving and sharing of their data?
- How do these terms interact with the relevant ethical and legal considerations?

User Expectations

- Recent high-profile controversies around online data sharing and use of SMD for research have increased public concern for how digital data is used (Fiesler and Hallinan, 2018; Hallinan et al., 2020).

- SMD seen as moderately sensitive relative to other personal data (health, demographics etc.) but still prefer consent before using it (Hemphill et al., 2022).

- **4 in 5** expect to be asked for consent before their SMD is used, **9 in 10** expect it to be anonymised before publication (Williams et al. 2017)

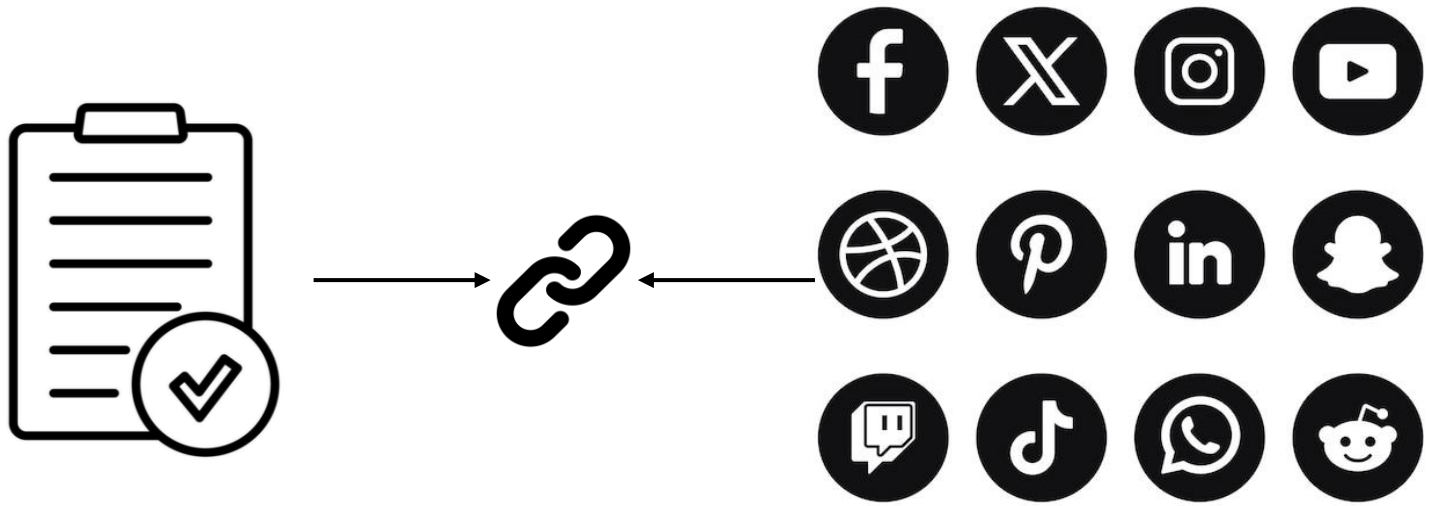
The “Third Era” of Survey Research

- Survey research is entering a new era of development which centres on its linkage with other forms of observed or digitally generated data (Groves, 2011)

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

- Of these DTD, social media data (SMD) has been an especially popular information source for social scientists given its public accessibility, allowing researchers to study human networks and behaviours in real-time and on a massive scale

- The eagerness of researchers to collect data from social media platforms has vastly outpaced the adaptation of our traditional ethical frameworks in social science and this needs to be addressed (Sloan et al., 2020; Webb et al. 2017; Williams et al., 2017)



Strengths	Weaknesses
✓ Value imputation	✗ Can introduce new biases
✓ Assess measurement error	✗ Asymmetries in data quality and quantity
✓ Reach “harder-to-reach” populations	✗ Increases risk of disclosure (think Netflix prize privacy breach in 2006!)

Sharing Linked Survey-to-SMD

- Researchers have begun to consider the ethical and practical implications of using SMD as an *isolated* data source but have largely neglected it in the context of linkage with other forms of data (Sloan et al. 2020)

- The innately *disclosive* nature of SMD (direct PII like names, biographies, profile pictures, locations) makes it impossible to share in its raw format without exposing sensitive survey information

- However, researchers are also expected to abide by the principles of **open research,** making their datasets more Findable, Accessible, Interopable and Reusable (FAIR)

- The importance of open research is compounded if the data has been publicly funded (maximise taxpayer value) or collected from vulnerable or harder to reach populations

Guiding Question of Our Research

RQ:

How can we as researchers bridge the gap between data protection and open research when working with linked social media data?

This has been explored within the context of the ELPs framework for working with social media data, the issues that are compounded when linking these data with survey responses, and the methodological challenges we face

Making Linked SMD Shareable

"Data protection law does not require you to adopt an approach that takes account of every absolute or purely hypothetical or theoretical chance of identifiability. It is not always possible to reduce identifiability risk to a level of zero, and data protection law does not require you to do so…

Effective anonymisation is about finding the right balance between managing this risk while keeping the utility of the data."

(p.12)

Information Commissioner's Office (ICO) Anonymisation, pseudonymisation and privacy enhancing technologies guidance, Chapter 2, October 2021

Three Areas of Guidance on Effective Anonymisation:

1) **Identifiability:** Can an individual be directly identified in a dataset, or indirectly "singled out" from other participants?
2) **Linkability:** Can an individual be identified when the data is linked with other external data sources?
3) **Inference:** Is there any potential to reasonable infer, guess or predict personal data based on the data provided?

Assessing the Risk of Identifiability

- The “**Motivated Intruder Test**”:

How likely is it that a motivated intruder would be able to successfully identify an individual in your anonymised data?

- The test accounts for identifiability risk in the context of the realistic cost of identification by other motivated actors in human, economic, temporal and technological terms

- A motivated intruder would be an individual who is:

✔ Reasonably competent

✔ Access to appropriate resources

✔ Uses investigative techniques

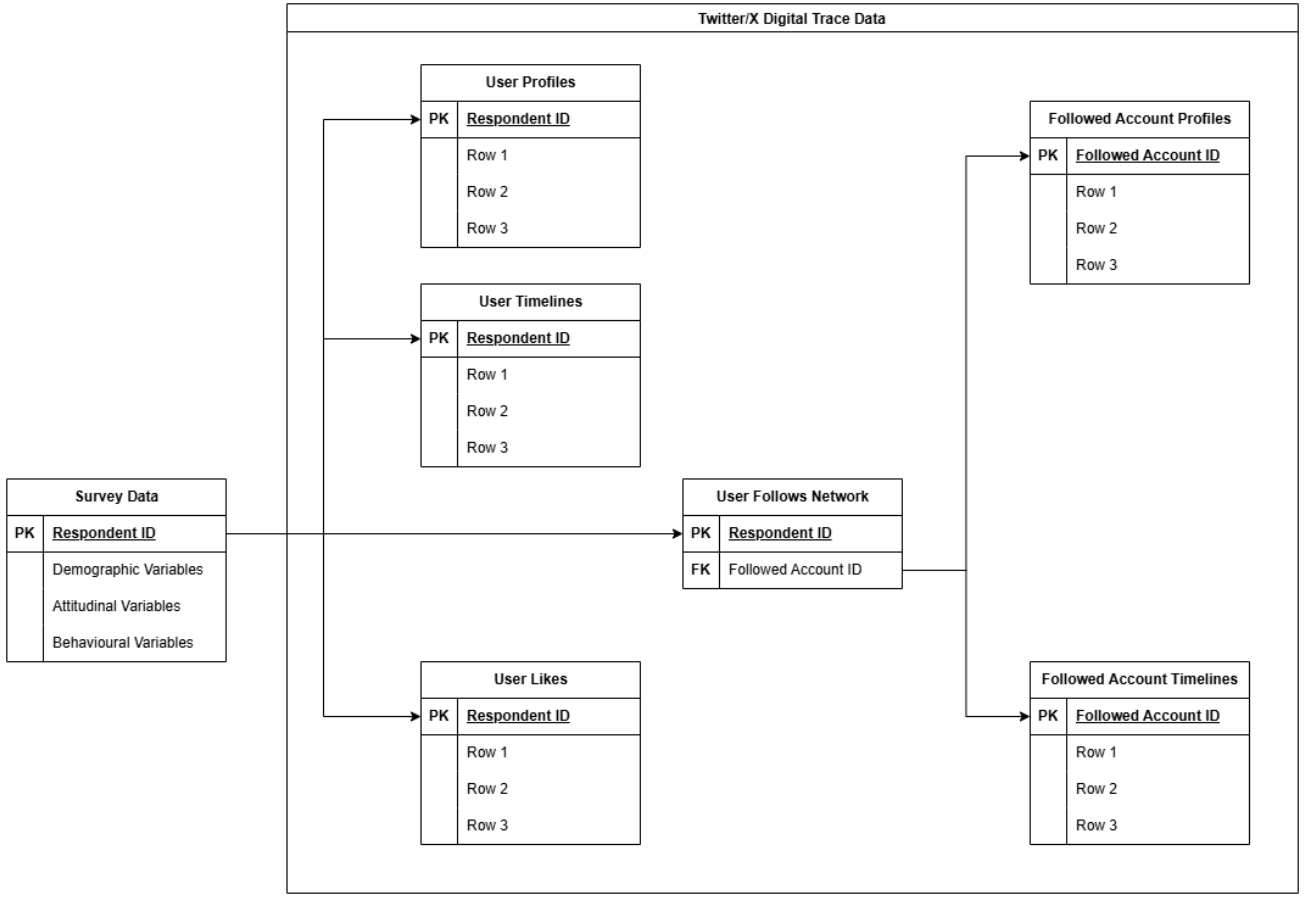
✘ Specialised knowledge

✘ Access to specialist equipment

✘ Resort to criminal activity

Statistical Disclosure Control (SDC)

- SDC is a set of methods and techniques used to reduce and test disclosure risks
- Even if one anonymises *direct* identifiers, the public nature of SMD makes all variables potential *quasi-identifiers*
- Three key SDC assessments:
  - k-anonymity (prevents singling out)
  - l-diversity (prevents linkage attacks)
  - t-closeness (prevents inference)
- SDC techniques include binning, aggregation, suppression, randomisation, deletion, and perturbation
- However, the more we **protect** the data, the more **utility** we destroy…



Data Protection-Data Utility Trade-off

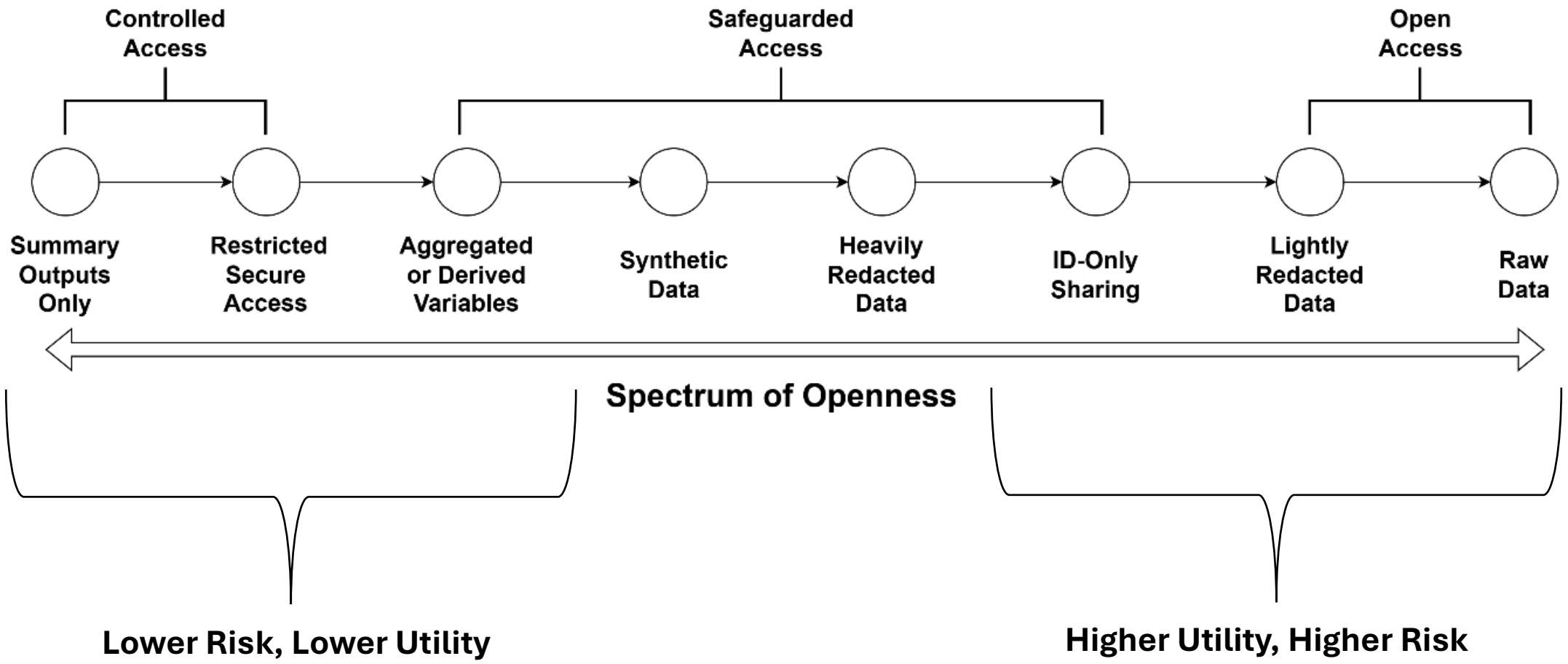
- We conceive of the conflict between data access and data protection as a **“spectrum of openness”**

- On the one side, researchers can opt for **complete data protection** where no linked data is shared (unless via controlled access) and only high-level summaries are published

- On the other side, researchers could opt for **complete openness**, releasing the data in its raw state, entirely unedited, unfiltered and unprocessed

- Then there are several positions in between, depending on how far a researcher wishes to lean one way or the other

Spectrum of Openness



Activity Questions:

1. Where has some of your own research sat along this spectrum in regard to data sharing? This can be specific to SMD or linked SMD, or any data at all!

2. What types of methods and strategies have you followed for ensuring your data has been safe to share with others?

3. Do you have any reflections on times when you have followed one or more of the options on the spectrum? What worked well and what were the limitations?

4. Are there any steps along the spectrum that we've missed? Perhaps something that worked well in your own research.