SoK: A Privacy Framework for Security Research Using Social Media Data

Kyle Beadle[1], Kieron Ivy Turk[2], Aliai Eusebi[1], Mindy Tran[3], Marilyne Ordekian[1], Enrico Mariconti[1], Yixin Zou[3*], Marie Vasek[1*]

UCL

Abstract—The use of social media data in research is common, spanning fields from computer science to social science, from human-computer interaction to law and criminology. However, social media data often contains personal and sensitive information. While prior work discusses the ethics of research using social media data, focusing on ethics broadly can be insufficient to unravel granular privacy risks and possible mitigations. Focusing on research papers that use social media data to study security-related topics, we systematically analyze 601 papers across 16 years, covering a wide array of academic disciplines. Our findings highlight a lack of transparency in reporting— only 35% of papers mention any considerations of data anonymization, availability, and storage. Applying Solove's taxonomy to classify the identified privacy risks in the social media setting, we observe that Solove's taxonomy was prescient in capturing aggregation risk, but the volume, timeliness, and micro details of data, combined with modern data science, yield risks beyond what was considered 20 years ago. We present the implications of our findings for various stakeholders: researchers, ethics boards, and publishing venues. While there are already signs of improvement, we posit that some small behavioral changes from the academic community may make big difference in user privacy.

Methodology

10,363 papers identified

↓

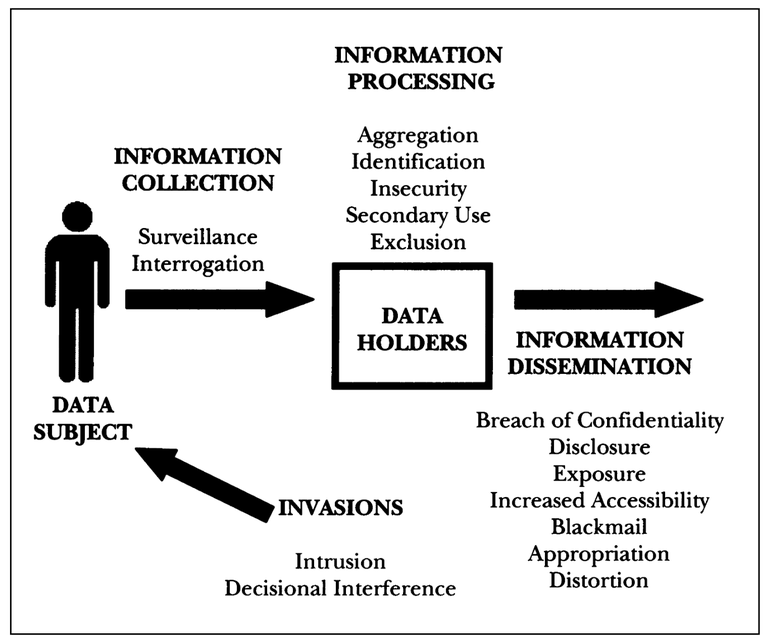
6,657 papers after duplicates removed

↓

776 papers after duplicates removed

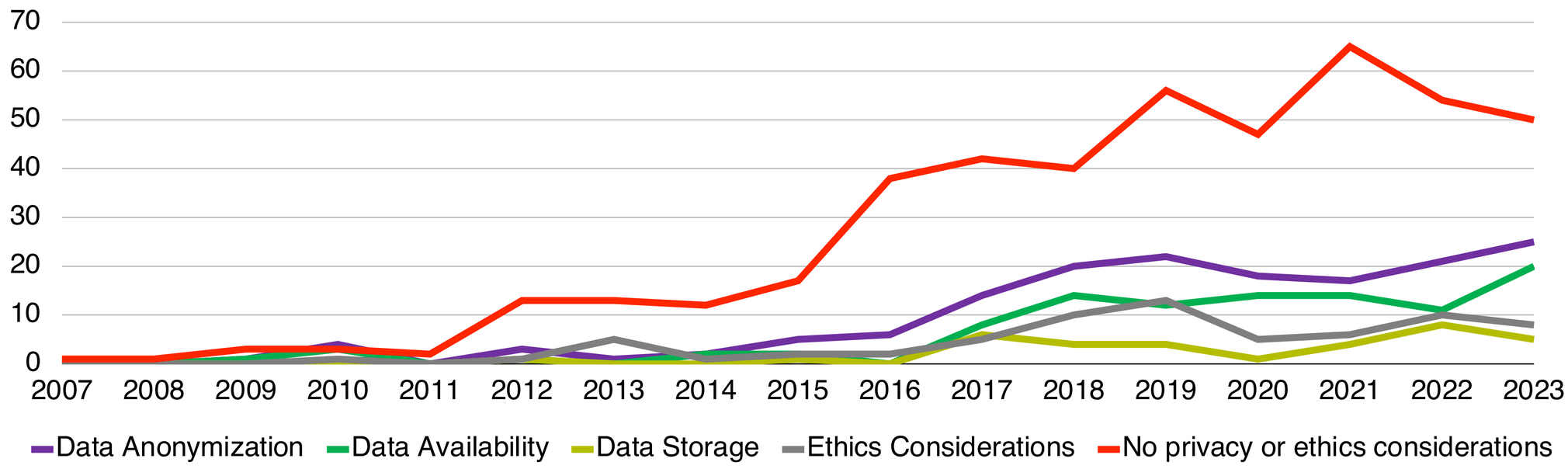
↓

601 papers included



Daniel J Solove. A taxonomy of privacy. University of Pennsylvania Law Review, 154:477, 2005.

Only 35% of security and privacy papers using social media data mention any considerations of data anonymization, availability, and storage.



Privacy Risks - Aggregation

Risk Manifestation	• Reuse of existing datasets • Combining datasets to detect sensitive attributes
Risk Prevention	• Consider implications before research • Expected use of data
Trade-offs	• Oversight gap • Challenges for open science

Top 5 Platforms Studied in Our Dataset

Platform	Count (%)
X / Twitter	296 (49.2%)
Facebook	108 (18.0%)
FORUM	73 (12.1%)
Reddit	47 (7.8%)
YouTube	37 (6.2%)

Privacy Risks - Insecurity

Risk Manifestation	• Storing data on an exposed server • Improperly storing data locally
Risk Prevention	• Encrypting data • Secure remote access
Trade-offs	• Resource requirements • Data loss or theft

More Risk Prevention

- Certificate of Confidentiality
- Privacy Risk Analysis
- Data Donation

Implications

- Researchers disclosing risk
- Ethics boards/IRBs understanding risks
- Venues setting and enforcing expectations of social media data privacy

Key Takeaways

- Tools exist to respect user privacy—we must hold ourselves and each other accountable to implement them.
- Initiate privacy-conscious research design, not just compliance.
- Encourage documenting and reporting privacy decisions.

kyle.beadle.22@ucl.ac.uk

https://kylebeadle.com

UKRI Engineering and Physical Sciences Research Council

UNIVERSITY OF CAMBRIDGE