Memos

Big social researchers

BSR01

Shares social media data, has more concerns about reproducibility than privacy
More motivated to get research results, make sure to follow terms of service
Spoke with online forum moderators
But didn’t consider ethical challenges related to the idea of BSR as human subjects research
From CS, publishes in management literature as well, aware of disciplinary differences in how data collection and BSR are addressed

BSR02

big social data is controlled by companies/organizations—database changes, data format, data sampling/data provided, terms of service, etc.
Idea of participants wanting credit for their contributions, depending on the community (e.g. on Wikipedia there is a value of openness and credit)

BSR03

Metadata quality, shortcomings of “authority records”
I was especially struck by the idea of specifically designing research questions in a way that takes into account ethical considerations. This can help guide big social researchers toward ethical research questions. However, it may limit the research questions that can responsibly be asked of research data. How might researchers ask questions that may be more sensitive, while still maintaining ethical standards?
only try to measure things that you think will require ethically sound methods

BSR05

weighing risk with reward
will the results of the research be important enough to justify risk to participants? esp with sensitive data, or identified data like the panel with voter records matched with Twitter
Openness vs privacy.

- How do researchers conduct these risk-benefit analyses? Have they been trained to do so?
- This is a classic strategy for understanding ethically-relevant harms, but most researchers appear to conduct these analyses informally.

- Analyses are done by talking to colleagues, reading relevant literature, and thinking about potential harms to participants (+ harms to reputation or other professional consequences).

Twitter ToS vs. better quality data
Reading, Conversations with collaborators a main way that he considered these issues

BSR06
Knowledge and thoughtfulness about responsible research is growing all the time - would have a better understanding now than in 2017

Collaboration with social scientists - as mentor, not coauthor
enough TweetIDs being published about natural disasters that he could reuse a lot of data, also collect their own data
knowledge about who uses twitter

BSR07
Different social media platforms have different context expectations for data. E.g. pinterest pins depend on the context of the board and related links, but are often taken out of context from the person who is pinning.

- different social media platforms have different expectations of privacy - Pinterest is by nature less private—pins are being repinned all the time
- However, most users wouldn't expect their pins to be used for research purposes—outside the context of their original intent/purpose of the pin
- "So there's some concern there [about privacy], but I don't know, I feel like it's outweighed by the fact that, you know, we're trying to document something that might be harmful and trying to help public health professionals. So yeah, I feel like on balance, it's an acceptable practice."

field of Journalism considers consent differently from qualitative researchers - not studying people, studying *content*

- this also contributed to them not feeling they needed IRB approval

Terms of service/fair use/IP confusion—didn't know how to think about it, just assumed it was okay, since they weren't publishing the full dataset

BSR08
protecting privacy
convenience of secondary datasets
pace of academic research is slow, but social media landscape changes quickly

BSR09

because they bought the data from Twitter, not allowed to share.
Some research questions work better with social media - e.g. weather disaster events
used filters to filter tweets for correct context, but hard because limited metadata—not many geotags, developed a method to extract location from content of tweets
tools to filter out bots
didn't think about consent bc of Twitter terms of service - the team bought twitter data from the company, so felt fine using it without specific consent from users.
also went through IRB, but felt if these larger entities okayed the research, they were okay without specific consent from users
user privacy more important - protect identity of individuals - username, user ID, pictured in photos or named in tweets, took excerpts from tweets rather than quoting the full tweet
had seen issues arise in previous social media papers (or in response to publication of previous papers), and responded to those issues

BSR10

strategies for inferring context - profile data, hashtags, etc
macroscopic level - by analyzing (topic modeling) more data, context will emerge.
self regulatory behaviors in a social network - low quality bots or fake news spike, but then peter out quickly as influential members of the network
Funder data sharing requirement - shared 1% of total tweet IDs (on project page, not repository) for convenience
social media terms of service change often

Qualitative researchers

QR01

had a data curation specialist on the research team who helped think through responsible sharing practices from the beginning of the research design
very thoughtful about data sharing, spent a lot of time working with repository to make sure as much could be published safely as possible.

QR02

people have contacted them directly - just people who had read what they'd published
graduate students at the top of the pile get access to your data
data is transmitted in a "pseudo kinship" relationship
passed down to grad students after you die
data might get "stale" - useful to historians and to anthropologists

QR03

Committed to data sharing - has experienced others wanting to use their data
interested in the benefits of data sharing - to communities and to researchers
idea that respondents want to be identified - want their stories to be shared

QR04

only do research on data when you think you can ethically do so
scope conclusions as appropriate
qual researchers look to computer science to see how they can implement strategies that they use in CS to scale up
ethnographic reports as historical records rather than data per se - do your life's work, then donate to libraries

QR05

Couldn't find standardized protocols for deidentification, so created own protocol
Included a readme and links to related articles
The study included institutional data from their university that they couldn't publish along with their research data

QR06

Had thought through curation with repository, but the terminology and ideas around data sharing were still pretty unfamiliar to them. Motivation to share data was funder mandate.
Context: team may think something is obvious, but maybe it's not obvious to somebody who hasn't worked on this project for years.
how much of the consent form do participants really understand?
decided to deidentify videos (blur faces) even though participants had consented to having their faces in the videos.

QR07

training in qualitative research is helpful - if you don't have it you have to cobble stuff together.
we need a good resource about deidentification of qualitative data. What are best practices?
How can we make the data useful (maintain context), while not compromising privacy of participants?

QR08

Talked about how extensive the explanation of context should be, and how to balance contextual information with privacy/deidentification
Used quotes mined from research papers, so this is secondhand information already and the context is once removed.
Data from a blog where people post about [health issues]—they considered it to be public, but they were also concerned with human participants.

QR09

reviewed the literature about how to do qual secondary analysis
research team knew and trusted each other
acted ethically according to “their own standards”

QR10

Secondary analysis was prompted by insights in the first analysis—analyzed own data.
Discussion within the research team about informed consent, but decided reconsent wasn’t necessary
Strategies for reducing harm for participants—removing quotes critical of their workplaces

Data curators

DC01

Considering how to reduce potential harms of responsible big social research—are there big social data that just shouldn’t be collected and curated because it’s too risky?
Considering how context can be communicated as part of the shared dataset, while still maintaining privacy. Finding balance.
Also discussed outreach and advocacy for library data services and data management.

DC02

Varied role and guidance provided by IRBs about data reuse.
Complexities of deidentification with qualitative data
Role of data curators to provide training and guidance on data sharing from the beginning of the research process.
Consent processes - tiered consent options
Restricted access/ access controls

DC03

qual curation is time consuming
consultation with colleagues
non-standardized metadata

DC04

Ensuring quality of transcripts - multiple transcribers and reviewers. Idea of data comparability in order to align with other studies. How to connect data to related studies. IP issues when datasets and data collection instruments are copyrighted.
Rarity of qualitative data sharing

DC05

Data creators provided a document to the data curators explaining the deidentification procedures they had implemented so far

- what rules they used, what identifiers were masked, etc. Helped the curator follow those rules during their review.

Difficulty of responsiveness of data creators

- can the curators reach them with questions? Can be difficult to get responses, and therefore, curators may not reach out with questions.

Standard questions asked for privacy/disclosure risk review - how many subjects, vulnerable pops?, etc.

DC06

Embedded data curator in a research project. Felt most comfortable talking about support for data sharing - providing access. Curating longitudinal studies, making sure metadata was standardized through the years, across different students, lab members.

DC07

Ideas about archives/historical documents versus research objects

- collecting proactively - like oral histories

Generative/iterative activities - Learning more about the collection as you collect - new hashtags, new revisions - a thing in motion. parameters change over time.

- similar to how in qualitative research, you can learn more about your research question over time.

Reproducibility - because a social media dataset may change over time, can’t really be used for reproducibility, but just transparency

- similar with qualitative data how you can’t draw universal conclusions from the research, just for that specific population, and your specific context

DC08

context-dependent - different research questions, communities, datasets require different treatment

risk and benefit - just “learning something new” isn’t enough if there is a big risk to the community

idea of community versus individual consent and risk - informed consent from an individual doesn’t really matter if the community might be at risk

could be an argument for community focus groups or advisory boards for social media research projects, rather than individual consent

who is the curator/researcher responsible to? Twitter? science? the community? weighing the responsibility to both and making decisions from there.

For social media data, repository can provide analytical output of data rather than full dataset

DC09

flowchart for consent decision-making

moving into the big data space by talking about data sharing and reuse at computational social science conference

IRB connections

starting early in the process to support good data management

challenges of high quality curation when it may not be the PI’s priority

interoperability of qual data analysis systems

DC10

Considered big social data less identifiable, suggested that more data could lead to broader conclusions

When curators aren’t trained as librarians, they may be less fluent in library and information science disciplinary ideas