Worst Case Scenarios in Sharing Large-scale Sensitive Data

Amy O'Hara
Massive Data Institute, Georgetown University
amy.ohara@georgetown.edu

2019

INTRODUCTION

While working for the Census Bureau, I created a unit whose mission was to increase the awareness and acceptance of administrative data for research purposes. I sought data including as much identity, temporal and spatial detail as possible to enable linkages across files and over time. I successfully negotiated data sharing arrangements because of the agency's statutory authority. The Census Act prohibits data use for non-statistical purposes (e.g., surveillance, enforcement, or marketing) and from releasing information that could reveal the identity of any business or individual in the data. I pursued arrangements with data owners across all levels of government and industry, vastly increasing the number and variety of data sources[1] to improve Census' economic and population statistics.

While running this unit, I constantly considered the risks of acquiring and using universe-level datasets. I often asked myself, "What is the worst thing that could happen?" On any given day, there were many, many answers to that question. My high frequency worst things worries were:

- Data breaches
- Information degradation
- Insider threats
- Volatility in sources
- Data misuse
- IT constraints (computing and security)
- Errors in published numbers
- Negotiations falling through
- Suboptimal linkages
- Parties revoking agreement terms

[1] https://www2.census.gov/about/linkage/data-file-inventory.pdf

I was constantly seeking best practices to prevent the worst things from happening and developing plans to deal with actual disasters. As the signatory on most of the data sharing agreements, I was accountable for adhering to the terms and conditions within (and I especially wanted to avoid fines and imprisonment from violating IRS agreements). I encourage applying my "what could go wrong" thinking to a large-scale data sharing initiative. Social media, search engine, rideshare and job search platform, patient and customer encounter, and cell phone data contain large volumes of time- and context-specific information, include legitimate and illegitimate entries, corrections and deletions, and are available with hazy degrees of permissible uses and consent. The data generation process and owner incentives for sharing are very different than with government data.

I explore four of the worst things that I think could happen involving large scale social data sharing. I define and describe potential consequences for each, suggest possible methods for avoidance, and then consider the potential functions of an institution to facilitate and moderate responsible data sharing.

THE WORST THINGS THAT COULD HAPPEN

No data sharing

One of the worst things that could happen would be the absence of sharing. If data owners refuse to allow access, knowledge production suffers and researchers will use inadequate sources that result in misleading or conflicting measures. Without data sharing, inferences are drawn from small, expensive samples that fail to capture movements observable in population-level data.

How can we avoid this worst case? What are incentives for data sharing? Data owners want to avoid the lawsuits, embarrassment, and staff burden that sharing entails. Can we propose pathways of sharing with varying degrees of control?

Owners could share internal, restricted data with outsiders by requiring researchers to join their organization as consultants or employees, generating analyses that are beneficial or benign to the company. Embedding within the organization allows control over access and output. This works with corporate data, state agency data, and student data. If encouraging this path, help owners standardize a transparent process. How do people apply, how are they selected, will the owner filter the results? Create a disclaimer for outputs and standardize language for COI/affiliation listing. Develop a process to securely retain analytic extracts.

Owners could use an intermediary to manage access. This works in multiple academic and nonprofit settings within sectors and topics, effectively outsourcing governance and monitoring. It only works when owners have complete trust in the intermediary, and incentives are clear and stable. GovLab is exploring data collaboratives, the University-Industry Demonstration Partnership explores the structures in industry and academia that affect partnerships.

Bad data sharing

Bad data sharing can result in broken trust, penalties and lawsuits, compromised identities, and erroneous inferences. What are the causes bad data sharing? In most cases, it stems from a poor setup or poor controls.

A poor setup involves weak agreements and contracts that cause bad data sharing. Some agreements lack formality and specificity. They fail to adequately define terms of use, user requirements, or data management and security protocols. Data are also shared badly when metadata and provenance

measures are missing. This happens in government and industry, where institutional knowledge is relied on in place of documentation. Numerous groups offer guidance and examples to improve agreements. Guides and toolkits are online, and community support is available in some domains. How applicable are resources that focus on community-academic partnerships to industry-academic partnerships? Does more general guidance, such as Grabus and Greenberg (2017) resonate with decision makers for data sharing in the private sector?

Poor controls also result in bad data sharing. Poor data management protocols can result in unauthorized access or information retrieval. Poor controls may result in data uses outside the agreed upon terms and conditions, jeopardizing the business interests of the data generator and the trust of their clients. Poor controls can also result in output problems, risking over disclosure that compromises data subject privacy. How could you avoid this worst case? Better controls are needed. Multiple research teams are pursuing ways to automate agreement formation and data usage controls, including policy development at Research Data Alliance and formal models such as Karafili et al (2017). But smart contracts are not imminent. The Census Bureau has layers of clearances and monitoring on the actors, worksites, and analyses, but this control is costly and burdensome. More connected systems and automated monitoring must be developed and periodically inspected by trusted auditors.

Data are discarded

If owners throw out the data, sharing cannot take place. Data may be discarded by owners who see no value in retention, or only see expense and liability. Owners may also be deleting data by overwriting it, often a result of legacy systems designed when storage was expensive. Data may be discarded due to legal constraints. Examples include federal laws that limit months of retention, such as 24 months for the National Directory of New Hire (NDNH) data or 48 months of National Change of Address (NCOA) data. Can a trusted institution retain data for secondary use, to enable historical or longitudinal study when firms have no incentive to keep it? Would such an arrangement be durable over time? How can the additional privacy risks (Altman et al. 2018) be addressed? Should laws be changed to authorize longer retention windows for government data like NDNH and NCOA? When a record is deleted, can any information about its original existence be retained to enable record count checks in later versions?

How do retraction and revision factor into data sharing arrangements? When juvenile records are expunged, tweets are deleted tweets, or account holders request erasure, how are data stores handled? What version control and metadata can reflect evolving databases? What is the effect on replication? Are there liability issues that make sharing some data risky for owners?

When an owner reviews output and suppresses or demands changes, data are effectively discarded. This censorship affects the validity of results. This occurs in arrangements private companies but also with state and federal governments but is seldom discussed. How can this be addressed, through preregistration, by making the terms of review and release public?

Public outcry

When data subjects, advocacy groups, lawmakers, or the media are surprised by secondary data uses or data sharing, they can be alarmed and angry. The public outrage may be legitimate, due to duplicitous or clueless data owner actions, or it may be due to a lack of transparency on the justifications for use and security protocols. In any case, public outcry often halts or ends data sharing. How can you avoid this worst case? Data owners and analysts need to be more transparent with the public about how and

why the data will be used. This is hard. In government, I did federal register notices, privacy impact assessments, websites, and seminars. They seldom reached the audiences that matter. Practical tools such as Finch's engagement matrix (ADRF Network Working Group Participants 2018) can be tested and deployed. Clear but sensitive language needs to accompany all output explaining that important research was made possible by clients or users like you with minimal intrusion and risk. It needs to be as commonplace and recognizable as PBS' "Made possible by viewers like you" and the American Humane Society's "No animals were harmed" certification.

WHAT CAN BE DONE?

There are ever growing mounds of consumer, user and usage data, reflecting encounters, transactions, and statuses. Fortunately, there are many responsible owners and providers, as well as experts to help curate and preserve data. Can another institution accelerate more responsible data sharing? Others have called for cross-sector intermediaries, including Groves and Neufeld (2017) and Bernholz (2016). Some international initiatives[2] may also be relevant. If a new institution took action, how much of the following would it do?

Identification. An institution could seek useful sources, monitor emerging sources, and sponsor data collection. It could actively build the catalog. Or connect those with data to those who want it. How technically involved in the data should an institution be? Should it help users understand universes, data treatment, and limitations?

Negotiation. An institution could set norms for working with firms, exploring the value proposition between parties. Would it negotiate on behalf of individual projects or a set of uses?

Agreements. An institution could manage agreements after negotiation, handling their time limits, maintenance, and modification. Would an institution help enforce negotiated terms of use? Would an institution consider information ownership and licensing or subscription terms? How involved with payments would an institution want to be on behalf of a data owner?

Data Transfer and Access. An institution could manage a data environment with current period and/or historical data and provide controlled access. Does it enable a federated system? Or is it an aggregator building a repository? Would an institution act as a trusted third party and obtain identified data to conduct joins? Should it anonymize data for researchers? Would it provision through remote access or host researchers? Would they handle screening and monitoring of researchers (could they outsource through institutional Google sign-in or use Experian to validate identity)? Would an institution assist with output review?

Gather tools and models. An institution could engage with those who explore privacy, ethics, and security controls. Would an institution act as an IRB? Would an institution work on messaging and monitor perception? Would it advance transparency, helping subjects see how their data are being used? Would it explore data trust or commons models, would multiple approaches be needed depending on data type and source or expected users?

[2] Such as the UK Consumer Data Research Centre (https://www.cdrc.ac.uk/) or Canadian Social Media Data Stewardship (http://socialmediadata.org).

CONCLUSION

These observations are drawn from my federal experience where I hoped for the best and planned for the worst. Will companies engage? Ongoing conversations with companies are needed, building on existing work by the GovLab and Future of Privacy Forum (Harris 2017), to understand their motives and incentives. Cross-disciplinary and cross-sector efforts are needed to set norms on data sharing and archiving and to invest in technical and governance models. An institution supporting large-scale social data sharing must be durable enough to withstand public scrutiny and corporate leadership changes. With a clear vision, well-defined scope, and persistence, it will make significant progress.

References

ADRF Network Working Group Participants. 2018. *Communicating about Data Privacy and Security.* Administrative Data Research Facilities Network Working Paper 1. https://repository.upenn.edu/admindata_reports/1?utm_source=repository.upenn.edu%2Fadmindata_reports%2F1&utm_medium=PDF&utm_campaign=PDFCoverPages.

Altman, Micah, Alexandra Wood, David R O'Brien, and Urs Gasser. 2018. "Practical Approaches to Big Data Privacy Over Time." *International Data Privacy Law* 8 (1): 29–51. doi:10.1093/idpl/ipx027.

Bernholz, Lucy. 2016. *Trusted Data Intermediaries.* Accessed May 2, 2019. https://pacscenter.stanford.edu/publication/trusted-data-intermediaries/.

Grabus, Sam, and Jane Greenberg. 2017. "Toward a Metadata Framework for Sharing Sensitive and Closed Data: An Analysis of Data Sharing Agreement Attributes." In *Proceedings of the 11th Annual Conference on Metadata and Semantic Research,* edited by Emmanouel Garoufallou, Sirje Virkus, Rania Siatri, and Damiana Koutsomiha, 755:300–311. Cham: Springer International Publishing. doi:10.1007/978-3-319-70863-8_29.

Groves, Robert M., and Adam Neufeld. 2017. *Accelerating the Sharing of Data Across Sectors to Advance the Common Good.* https://mccourt.georgetown.edu/massive-data-institute/dataforsocialgood.

Harris, Leslie. 2017. *Understanding Corporate Data Sharing Decisions: Practices, Challenges, and Opportunities for Sharing Corporate Data with Researchers.* https://fpf.org/wp-content/uploads/2017/11/FPF_Data_Sharing_Report_FINAL.pdf.