Data Security in Practitioner-Academic Partnerships: An Agenda for Improvement *

Aidan Milliff *Massachusetts Institute of Technology*

A large and growing body of social science research uses personal information from research subjects—often collected in collaboration with non-academic partners like private enterprise, governments, or NGOs. Consideration of the complex data security obligations and obstacles that come along with academic/practitioner collaboration, however, has been piecemeal, and actual security practices vary widely from project to project. Identifying and implementing “good practices” depends on the scruples and best judgment of individual research teams because discipline-wide best practices are insufficiently consolidated. This essay makes three preliminary contributions toward the development of best practices. First, it proposes a general set of imperatives/obligations against which data security practices can be measured, and then identifies the primary threats to fulfilling these obligations. Second, it elaborates on the particular threats to data security that come along with practitioner partnerships. Finally, it suggests a non-exhaustive range of data-security practices to address different threats, and demonstrates in appendices how three off-the-shelf tools can be used to ameliorate threats to data security, especially in collaborative projects.

Introduction

Social scientists who work with personally identifiable information make two promises related to personal data: a promise to respondents that their sensitive, personally identifiable information will be kept confidential and not misused, and a promise to the academic community that their research will be conducted transparently and replicably. Developing workable practices for data security in social science research is a matter of finding a non-burdensome protocol for data handling that successfully optimize between these two promises.

When academics work with practitioner partners in industry, government, or the non-profit sector, either in the United States or internationally, they almost always make a third set of promises—to the practitioner partner—that make optimizing between data security

*This essay is an extension and revision of an informal set of data security tips co-authored with Minh Trinh and Lukas Wolters. Recommendations drawn directly from that document will be cited accordingly. Thanks to Lukas Wolters, Minh Trinh, Andy Halterman, and Lily Tsai for feedback. Thank you in advance to Molly Roberts, Jesse Driscoll, and participants in the “Multi-Method Tools for Data Security in Political Science Research” panel at APSA 2020. **Current version:** Prepared for the 2020 APSA Annual Meeting, 13 September 2020; **Author contact:** milliff@mit.edu.

and transparency even more difficult. Adding practitioner partners to the research workflow introduces three new challenges. First, practitioner’s high level of context-specific knowledge makes them better equipped to de-anonymize confidential data, which may violate promises made to respondents. Second, all but the highest-capacity partner organizations may constitute weak points in a data security architecture if they practice poor digital hygiene or are operating in a context where they are not well protected from government interference. Third, since most practitioner-academic partnerships involve some sort of data-sharing agreement or agreement about what data will be collected, partnerships decrease the level of flexibility academics have to make decisions that are conducive to data security. Academic researchers should pay special attention to these additional challenges when designing and implementing data security practices for work with non-academic collaborators.

This essay will proceed in four sections. Since data security practices in social science research remain largely *ad hoc* at the practical or technical level, even as the normative framework around participant privacy has progressed substantially in recent decades ([APSA Ad Hoc Committee on Human Subjects Research, 2020](#)), the first section will cover basics: recapitulating agreed-upon imperatives of data security, introducing three threat vectors against which academics should be prepared protect their data, and enumerating characteristics that “good solutions” should have. A second section expands on the obligations and threats, describing how partnering with non-academic collaborators to collect data changes data security challenges and adds to the threat environment. A third section proposes—in brief—a wide range of practices for data collection, storage, and dissemination that begin to address the threats introduced in sections one and two. Finally, three appendices demonstrate or provide step-by-step tutorials to use different low-overhead and relatively non-technical tools to improve data security in collaborative research.

Researcher Obligations and Threats to Data Security

Data used in social science research frequently includes personally identifiable information (PII), or “quasi-identifiers” that can facilitate identification of individuals in the data.¹ Personal information comes from a variety of different sources: information can be collected directly from the people it describes as in a survey or interview, it can be obtained from an aggregator like a government agency or private company, or it can be scraped from semi-public sources like social media sites. “Data security” is maintained when personal information is kept confidential, and when third parties are *not* able to use published information to link anonymous records in researchers’ data sets to real people. Data security failures are those instances where anonymous records are linked to real people, either because PII is improperly released, or because identities are reverse engineered from the available data.

Security failures compromise respondents’ privacy and can expose them to a variety of negative consequences. Depending on the attitudes, behaviors, and conditions that are described in data that becomes personally identifiable, respondents can face severe repercussions like legal sanction ([McMurtrie, 2014](#)) or retribution from non-state actors ([Venkatesh, 2008](#)), as well as more diffuse harms like the impact on personal life, employment, or reputation that could come from exposure of health or financial records ([Ohm, 2010](#)).

¹When referring to PII, this essay roughly follows the definition specified by the National Institute of Standards and Technology for use by the U.S. Government: “1) any information that can be used to distinguish or trace an individual’s identity such as name, social security number, date and place of birth, mother’s maiden name, or biometric records and 2) any other information that is linked or linkable to an individual such as medical, educational, financial or employment records,” ([McCallister, Grance and Scarfone, 2010](#)). “Quasi-identifiers,” per [Dalenius \(1986\)](#), are data that are not individually identifying nor linkable to an individual, but could be “unique and publicly known” for some individuals, such that it is possible to identify individuals from theoretically non-identifying information. As an example: There are nearly 7,000 graduate students at MIT. 64% identify as male, 61% identify as white students who are U.S. permanent residents. Dozens (at least) are from Colorado. 60 students are in the political science department. None of these attributes is personally identifying, but the combination of all attributes identifies (as of 2020) a single individual: the author. These pieces of information are quasi-identifiers because particular combinations are rare or perhaps unique.

Obligations

Academic social scientists have an obligation to conduct transparent, replicable research no matter what sort of data they use (King, 1995, 2007; Lupia and Elman, 2014; American Political Science Association, 2015; Moravcsik, 2015). When working with PII, researchers assume an additional obligation to protect the privacy of their respondents. Mediating between these two obligations, one which incentivizes transparency (or at least the appearance of transparency) and another which incentivizes absolute opacity and secrecy, is a difficult and necessary task for many scholars. Social scientists are already well socialized to understand the professional incentives for transparency. The obligations and imperatives related to protecting confidential information, though coalescing quickly in the past decade, are not as well understood.

Most researchers who collect personal data themselves (rather than using already collected data) explicitly promise privacy to their respondents via consent forms approved by institutional review boards (IRBs) (Fujii, 2012; Zechmeister, 2016); in these situations fulfilling obligations to respondent privacy and confidentiality is a clear imperative of institutional policy, professional ethics, and, depending on funding source, federal law (American Political Science Association, 2012). Even when researchers do not make direct promises to respondents—for example, through the growing practice of partnering with a company or government to use existing data containing PII (King and Persily, 2019)—there is a growing consensus within academia and government funding agencies that the researcher is ethically obligated to protect PII in “public” data as if they had acquired informed consent and collected the data directly from respondents (National Science Foundation, 2017; Vitak, Shilton and Ashktorab, 2016; Gibney, 2017; Markham and Buchanan, 2012; Shilton, 2016).

For political scientists, the most recent statement of general norms and principles comes from the American Political Science Association Ad Hoc Committee on Human Subjects Research (APSA Ad Hoc Committee on Human Subjects Research, 2020). Confidentiality (not sharing the identity of research participants) ought to be maintained in general, and

promises about confidentiality should be made explicitly. When research data are shared, political scientists should take care to redact to the degree that even well-informed parties cannot identify individual participants. The guidelines further note that privacy and confidentiality obligations accrue to the researcher, not journal editors or reviewers, and that when researchers decline to share data due to the risks associated with de-anonymization, they “should be prepared to justify their decision to journal editors, to reviewers, and in oral and written reports of their research findings.”

Researchers are therefore caught in a difficult situation: they have apparent professional incentives to a) share data as much as possible, and perhaps more critically b) to maintain copies of *all data* indefinitely. Following those incentives necessarily jeopardizes commitment to respondent privacy to at least some degree: researchers in the field of statistical disclosure control contend that risk of de-anonymization is a direct trade-off against data utility ([Templ, Meindl and Kowarik, 2020](#)). To protect sensitive data researchers need to address a number of different threats. Some threats are integral to the process of collecting and possessing PII, while others arise when nominally de-identified data are shared.

General Threats to Data Security

Social scientists using PII (or quasi-identifiers) in their research should consider three classes of threat to privacy and confidentiality, whether or not they are working with non-academic collaborators: theft, expropriation or seizure of data, and de-anonymization of data that has been intentionally released.

Theft

Theft can occur at any point between collection and destruction of PII, regardless of whether or not a public version of the data is released. The threat can be divided into two types—physical theft and theft from digital storage—but this section will mostly focus on digital storage theft, because physical theft follows logic that should be familiar to most.

Why should social scientists take the threat of data theft seriously given that there have been no high profile instances of PII theft from social science research projects? For one, theft of personal data from academic institutions is hardly rare, but so far has targeted student records, not research data. In 2016, 98 public school systems, private colleges, and public universities including large flagship systems like Michigan State University and UC Berkeley reported PII stolen from their networks ([Identity Theft Resource Center, 2016](#)). The 2016 Berkeley cyber attack, where hackers targeted (but likely failed to breach) financial information of 80,000 student, faculty, staff and alumni in the school’s online payment system, came less than a year after another breach of the financial aid office exposed information including social security numbers for ~250 students ([Gilmore, 2016](#)).

University networks are potentially vulnerable, and given recent trends, researchers would do well to assume that their data may eventually be an attractive target, or at least adjacent to one. The threat of data theft via computer networks is compounded by three other trends. First, in an important effort to make research useful and consequential outside the academy, many social scientists now share exciting results far beyond paywalled journals, especially when the research involves innovative data ([Williams June, 2017](#)). Publicity, or even research publications, make researchers data holdings a sort of public information, and could make a sensitive dataset into a target.

Second, social scientists are beginning to use larger and more granular administrative data sets that contain PII in their research (for example, [Chetty et al., 2014](#)).² Finally, exposure to the threat of theft increases when social scientists work with remote collaborators and are required to store PII on a network or frequently send it back and forth in order to facilitate collaboration ([Summers, 2016](#)). Collaboration and co-authorship are increasingly common in social sciences, though political science seems to lag behind other fields ([Henriksen, 2016](#)).

²The data agreement for this particular study seems robust. Chetty and Saez’s team never gained direct access to PII from the IRS, but rather sent code to run their analyses to the IRS, which returned results that had been stripped of PII ([Mervis, 2014](#)). It is cited here primarily as a example of the potential publicity researchers might attract by using novel administrative/personal data.

Expropriation

Another class of threat, the possibility of government expropriation of sensitive data, also impacts data security obligations. Like theft, expropriation threatens any data that researchers possess, whether or not they release a public version. Expropriation can be broken down into two threats: legal and coercive. The likelihood of each type varies widely by context. In terms of legal expropriation, researchers—at least U.S.-based researchers—do not necessarily have legal protection to refuse when American courts demand confidential data ([Knerr, 1982](#); [Traynor, 1996](#)). In one extreme situation in 1993, a sociology graduate student who refused to testify against former research participants suspected of vandalism was jailed for four months after being held in contempt of court ([Scarce, 2005](#); [American Sociological Association, 2017](#)).

Bringing data collected in other countries back to the United States does not necessarily protect researchers from the threat of legal expropriation. “Off-shoring” data is particularly weak protection if the country of origin has legal treaties with the United States. In 2011, tapes from an oral history of the Irish Republican Army held by researchers at Boston College were subpoenaed under the auspices of a U.S.-U.K. mutual legal assistance treaty and used as evidence in a murder investigation in Northern Ireland that implicated multiple research subjects ([McMurtrie, 2014](#); [Radden-Keefe, 2018](#)). Whether researchers do or do not successfully rebuff attempts to compel the release to confidential data, the threat is worth considering.

Governments—more often non-U.S. governments—can also attempt to coerce researchers to give up confidential data.³ Researchers, especially those who are obvious visitors/out-siders, are sometimes monitored by security services while working, and are thus forced to take precautions to protect the confidential data they are gathering ([Wood, 2007](#)). In a small number of instances, researchers are closely followed or questioned by police ([Menoret,](#)

³Isolated stories suggest that the greatest risk of coercive expropriation by the U.S. government in recent years is at border points—both for U.S. citizens returning from data collection in other countries and for non-U.S. citizens returning to the American institutions at which they work or study.

2014). U.S. citizens conducting research in a foreign country might have a legal advantage in these situations insofar as they can go home and not expect to be extradited, but in general, the option to leave does more to protect a researcher’s personal integrity than the security of the data they have collected.⁴ Researchers need to take other steps beyond a flexible-date airline ticket to ensure sensitive data remain secure.

De-Anonymization

Finally, the possibility that publicly released data can be de-anonymized threatens respondents’ privacy. Unlike theft and expropriation, the threat of de-anonymization *is* a function of academic norms of transparency and reproducibility, not simply a consequence of collecting PII. Also unlike theft and expropriation, which researchers should seek to prevent to the greatest possible degree, researchers will often knowingly tolerate some risk of de-anonymization in order to fulfill professional data-sharing obligations. Because the threat of de-anonymization is less straightforward than theft or expropriation, guarding against it requires a more nuanced approach.

De-anonymizing data that has been stripped of obvious PII like names and addresses can be surprisingly easy, so protecting respondent confidentiality requires more than simply deleting names, birth dates, and other unique identifiers from data before release. Respondent anonymity can be compromised from technically de-identified data in two deeply connected ways: matching unique combinations in sparse data to publicly available data, and identification using contextual knowledge. Identification attacks that exploit sparse data structures are simpler than many researchers expect. For example, experiments with U.S. Census data show that as of 2000, 87% of U.S. residents are uniquely identified by only three attributes: ZIP code, gender, and birth date (Sweeney, 2000). These (nearly) unique Zip-Gender-Birth date identifiers could be cross-referenced with a variety of publicly-available

⁴It also does little to protect the personal integrity of people who cannot leave, because they live in the place data is being collected. Protecting interlocutors and research assistants from consequences unrelated to data breaches is a separate and very important consideration.

records like voter rolls to match “anonymous” records to particular people.

Re-identification in sparse, high-dimensional data sets (data sets on transactions or consumer preferences are simple examples) can also be accomplished using data that seems much less connected to identity than ZIP code and birth date. In a study of Netflix user data, computer scientists found that even a small amount of “background knowledge about a respondent’s movie tastes” was sufficient to identify their record in the anonymized Netflix data. IMDB accounts (social media accounts that include names) with as few as 5-10 movie ratings could be reliably linked to Netflix records because after a handful of widely popular movies, a person’s corpus of movie ratings (combined with the time stamps of the ratings) is a surprisingly individual trait ([Narayanan and Shmatikov, 2008](#)).

Adversaries can also de-anonymize data using broad contextual knowledge that is not codified in public records. Academic publications that use anonymized data collected from a particular place—perhaps an administrative district, school system, or even a company/organization—commonly describe the setting from which data were collected, without identifying the setting or the individual respondents. These innocuous details are important for describing scope conditions of social science theories, but they can also be used to identify the setting from which data were collected. Once the setting is identified, the risk of de-anonymization of individual respondents is much higher. Take one example, a highly-cited study correlating behavior on social networks with offline networks of friendship, which used data collected from the Facebook profiles of students at an “anonymous, northeastern American university” ([Lewis et al., 2008](#)). Authors provided contextual information about the university’s unusual housing lottery system: “towards the end of their freshman year students were allowed to identify up to 7 alters who collectively constituted ego’s “housing group.” Entry into a housing group is necessarily a mutual choice; and while not guaranteed to share a room, all students in a housing group are guaranteed that they will be placed in the same upper-class residence for the duration of their studies.” A privacy researcher combined this quasi-identifying characteristic as well as geographic and enrollment-size characteristics listed

in the study with oddly-named college majors listed in the study’s dataset codebook—many colleges offer gender studies degrees, but most do not call it “Studies of Women, Gender and Sexuality”—to positively identify the “anonymous” university as Harvard college(Zimmer, 2008).⁵

Once the setting is identified, the risk of re-identification posed by sparsity and unique combinations of quasi-identifiers becomes more serious. In the Harvard College data (no longer publicly available owing to these exact concerns (*Tastes, Ties, and Time Dataverse*, 2010)) there are unique records along a number of dimensions including nationality. The dataset includes only one international student from Iran. If the university were truly anonymous, the Iranian student’s identity would be protected to a reasonable degree because of the moderate number of Iranian students in college in the United States—roughly 1,500 at the time of the study (Ditto and Baste, 2014). If the university is known to be Harvard College and the class is known to be the class of 2009, re-identifying the Iranian student is much less difficult—especially for adversaries with additional contextual knowledge or personal knowledge of the Harvard College class of 2009. Leaving aside the easy re-identification of people who are unique in a particular characteristic measured in the data, identifying the context from which data are collected can still seriously threaten privacy when the data are sufficiently high dimensional—in the case of social network data, a person’s friend network and list of interest is almost as good as a fingerprint (Narayanan and Shmatikov, 2009).

Of the three threats to data security, de-anonymization is the most complicated. Social scientists are obligated to release data after publication in most circumstances. Sadly, the straightforward techniques commonly used to protect respondent privacy in published data are inadequate protections against even moderate-effort attempts to re-identify individuals, especially in high-dimensional data.

⁵One might have guessed the “anonymous college” was Harvard based on even less evidence than Zimmer used: Four of the five study authors were on faculty at Harvard when the research was conducted, and the study describes the research assistants who collected the data as being “in the same network” as the study subjects (Lewis et al., 2008).

Data Security with Practitioner Partners

Research partnerships between academic social scientists and practitioners (governments, NGOs, or private companies) create special cases of the general data security threats— theft, expropriation, and de-anonymization—of the general data security threats described above. This section describes three special cases that ought to be common across practitioner-academic partnerships, but is not exhaustive. In this section, I assume that PII collected in the course of research needs to be protected against improper use by the practitioner partner, as well as by third parties. Surely, under some arrangements, this is not the case.

Working with practitioners changes the presentation of all three threats. Research partner computing and data storage systems—especially the systems of NGO partners—may be old and more vulnerable than university systems. Research partners may be more vulnerable to expropriation given the need to maintain good relationships with local and national governments where they work. Perhaps most importantly, research partners are likely to be experts in the exact type of contextual knowledge that facilitates re-identification of nominally anonymous respondents, and, especially in the case of government and NGO partners, may have pre-existing relationships with anonymous respondents in which sensitive data would give them unfair or improper leverage.

Theft

Practitioner partners—NGOs in particular—might amplify the threat of data theft. Even many highly-capable NGO partners use outdated digital infrastructure and may not follow best practices for digital hygiene and information security. Practitioner-partners other than large companies are fairly likely to be a weak point in network security, and any PII-containing data that is stored on or transferred through their network is more likely to be vulnerable to theft. Negotiating changes to these practices, or avoiding poorly-secured networks all together, is a difficult if not impossible demand for organizations that are, in

general, under-resourced and stretched thin to begin with. Potential problems stemming from research partner network security and digital hygiene are a function of the project’s data plan—as described below, establishing strong protocols early is important.

Expropriation

Practitioner-partners (especially those outside the United States) might introduce an additional expropriation risk that could compromise data security. While researchers may enjoy the freedom to “go home” with their data in hand, local practitioner-partners might be subject to coercive pressure and threats from government, as well as pressure from funders. The exposure of practitioner-partners to government pressure puts any data held by the practitioner partner at risk, and may leave researchers with little leverage to ensure data security, even though they remain the ones ultimately responsible for doing so.

De-Anonymization

Practitioner partners in academic research can complicate efforts to prevent de-anonymization of data. Further, the threat of de-anonymization applies earlier in the data-gathering process academics are collaborating with practitioner partners. If sensitive data is passed back and forth between research partner and academic, the imperative to prevent de-anonymization might expand beyond published data to include non-public data shared with the research partner. Practitioners like NGOs, governments, and companies are valuable research partners *because* of their fine-grained contextual knowledge. The information asymmetry between researcher and partner is bound to be especially severe when researchers look to a practitioner partner (like a local NGO) to help recruit respondents from the population that practitioner serves (Neal et al., 2015). The more a research partner knows about the context and the population of respondents, the more points of external leverage they have to re-identify anonymous records from seemingly non-identifiable information.

Why does this matter? First, when respondents share sensitive information with re-

searchers it is not always the case that they consent to share that information with a third-party practitioner, even (or perhaps especially) if they have a pre-existing relationship with that practitioner. One of the common arrangements that brings academic researchers together with NGO partners, for instance, is program evaluation. If NGO partners were to de-anonymize data in which respondents shared attitudes or experiences related to NGO services, the negative consequences for respondent wellbeing could be meaningful.

Consequences might be particularly severe because, second, local practitioners (or, more fairly, individuals with bad intentions in otherwise well-intentioned organizations) are likely to have leverage to *use* sensitive information to disadvantage respondents. If, for example, a respondent admits to criminal activity in a survey and their survey response is de-anonymized by a partner social service organization, that organization could use the information to deny the respondent benefits. In a real (and now infamous) example from sociology research, disclosing personally-identifiable data on informal economic activity to local power-brokers in Chicago public housing allowed those brokers to extract unpaid “taxes” from the respondents (Venkatesh, 2008). It is clear in this case that the “practitioner partner” did not have the best interests of the respondent at heart, and that the researcher exposed his subjects to material harm by compromising their privacy (Waterston, 2009). Researchers must be especially careful about de-anonymization when working with practitioner partners because partners are likely to be uniquely capable of identifying anonymous respondents, and well-positioned to use sensitive information to respondents detriment.

Ideas for Improvement

Social scientists who use sensitive data ought to carefully assess data security threats and take appropriate measures to counter them. This task is equally necessary and somewhat more difficult when researchers are collecting and storing PII in collaboration with a practitioner partner. The following section will briefly introduce some of the measures that social scientists can take to address theft, expropriation, and de-anonymization threats to PII.

The following recommendations are not nearly an exhaustive list, nor are they necessarily representative of the cutting edge in security research. Rather, they are chosen to be *feasible* for social science researchers who are not accomplished programmers and do not have large budgets they can allocate to data security. Good data security practices will balance between obligations to privacy and transparency, ideal data security practices will achieve such a balance using techniques that have a low barrier to entry in terms of technical ability, time commitment, and cost.

Characteristics of Practical Solutions

Data security practices are only valuable if they can be implemented *correctly and consistently*. Practical solutions should be cheap enough and simple enough to implement that they can be *expected* of all researchers, regardless of their access to resources. At the same time, they need to be robust enough that researchers can credibly promise respondents that their confidential information is safe.⁶ Finally, given that some new and very robust data security techniques rely on complicated statistical and computer science principles ([The OpenDP Team, 2020](#)), this essay stipulates that in order to be practical for researchers who face professional incentives to be transparent, data security practices need to be intuitive enough (easy enough to explain how they work) that implementing them doesn't adversely impact the professional credibility of researchers. The limit of "practicality" in this regard is a moving target. As the median political scientist (including researchers who primarily use qualitative methods in their own work) becomes more familiar with computational methods, the viability of sophisticated differential privacy tools ought to become relatively lower.

The remainder of this section presents one general principle and non-exhaustive list of specific solutions that address the threats of expropriation, theft, and de-anonymization in practical ways. The list of solutions includes specific tools for use in practitioner-academic collaborations.

⁶There are no fool-proof solutions for data security, so any "robust solution" or promise made is more precisely a promise that the probability of a data breach is quite low.

Data Minimization as a General Best Practice

The surest way to avoid compromising the privacy of respondents is to *not collect PII and quasi-identifiers*. Certain quasi-identifiers like age, race, and location of residence are variables that affect a very broad array of social science outcomes, so collecting them is a necessary part of good science. Right now, though, many social scientists feel professional incentives to maximize data collection beyond these basic quasi-identifiers in order to ensure that a particular collection effort bears fruit—a researcher never knows what robustness checks will be necessary to satisfy a reviewer or produce interesting work from costly-to-collect data in the event that a main hypothesis is not supported. But data that are never recorded cannot be stolen or expropriated (unfortunately they might still be reverse-engineered from the data that are recorded).

One logistically simple technique that addresses multiple data security threats—both with and without practitioner partners—is a spartan impulse during the research design phase. Researchers can ask: “what is the minimum amount of PII necessary to test my hypotheses and evaluate the most likely alternative explanations?”, or “where can PII be substituted for randomly-generated ID numbers or quasi-identifiers recorded at a less-granular level?”, and “are the potential consequences of not knowing a respondent’s address (or name, or birth date) tolerable from a social science perspective?” This practice of “data minimization” or “privacy by design” is beginning to be formalized for web-based and big-data applications by security researchers in computer science ([Hansen et al., 2004](#); [Schaar, 2010](#); [Kung, 2014](#), etc.), largely in response to legal standards for data collection and storage in the EU ([Kinch, 2016](#); [Information Commissioner’s Office, 2017](#)). While the formal architecture might not be applicable to the relatively small scale (in both rows and columns) of most social science research projects, the general principle of minimizing collection can address all three threats (theft, expropriation, de-anonymization) to at least some degree. The “practicality” and professional acceptability of data minimization likely depends on early adoption by relatively senior researchers, and integration of data minimization-based justifications into already

accepted research design tools like pre-registration. Junior scholars and graduate students will only be able to refuse to do certain robustness checks on data-minimization grounds if the norm is first established by researchers with more job security.

Preventing Theft

Feasible theft-prevention measures address two possible avenues of threat: in-person theft of/from physical storage media and remote theft over networks. Protocols to protect against in-person theft are likely intuitive to computer users and should require little explanation: all devices should be protected by a password that is not re-used elsewhere, device hard drives and storage media like USB drives should be encrypted (on Mac OSX this option is built into system preferences, for Windows the built-in option is called BitLocker), two-factor authentication should be used everywhere it is available, and devices should be stored behind locked doors when not being used. When data are deleted, the files should be overwritten multiple times, or better yet, the storage medium should be destroyed to prevent recovery. Many open-source tools that overwrite deleted files are available free of charge ([Stanford University Information Security Office, 2005](#)). Preventing network theft requires a more complicated architecture, and high-quality protection against network threat vectors makes for a very inconvenient workflow. The risks and benefits of three network security workflows—cloud storage, local storage, and air-gapped storage—are discussed below:

Cloud Storage: Storing sensitive data on cloud-based servers like Dropbox or Amazon Web Services is extremely convenient—especially for research with collaborators—and provides excellent protection against hardware failure. Cloud storage, though, offers only moderate security protections against data theft. The data stored on Dropbox are well-encrypted using 256-bit AES keys. Encryption doesn't matter, though, if an adversary has a researcher's Dropbox password, and Dropbox admitted in 2016 that close to 70 million user names and passwords had been stolen in 2012 ([Turner, 2016](#)). Cloud storage services will remain

attractive hacking targets, so they will continue to pose theft risks. Further, the design of cloud storage technology requires an internet connection to access files. This means it will be impossible to isolate PII from an internet connection if using cloud storage. Isolating PII from network connections is one of the most robust security techniques to prevent theft via networks, and it is unavailable to researchers committed to storing PII on cloud services.

Researchers who are committed to using cloud storage by institutional policy, agreement with a research partner, or simple force of habit should consider additional steps when using cloud storage. One cautious step would be anonymizing or cleaning data (as if it were going to be released) before uploading to cloud storage, such that a data breach at a cloud storage provider would not jeopardize respondent anonymity any more than a planned data release would. Another possibility is adding additional layers of encryption to sensitive data destined for cloud storage (perhaps following the PGP protocol outlined below). Encrypting data before uploading to Dropbox, which already uses its own encryption, creates a system that is “single-point safe”: multiple independent systems including the cloud storage password or service-provider encryption *plus* the user-added PGP encryption must be defeated at the same time to access the data. Finally, all major cloud storage providers including Google Drive, Dropbox, Amazon, Box.com, and Microsoft offer two-factor authentication. Researchers should get in the habit of using it wherever available.

Local Storage: Storing sensitive data on a work computer (or an external hard-drive that is used by plugging it into a work computer) is also convenient, and presents a far lower target profile than password-protected cloud storage services like Dropbox. At the same time, local storage is somewhat more vulnerable to hardware failures than cloud storage. The common solution to this hardware failure vulnerability—backing up data to other storage media—increases the number of locations from which it may be stolen, and thus increases risk very slightly. Local storage on a computer that is used for non-PII work makes it impossible to maintain strict segregation (an “air gap”) between confidential information and other

files. The only way to truly mitigate the risk of theft via networks is to make sure PII is stored on a machine that is not accessible over any networks. Normal computers certainly do not meet this standard, and plugging external media like hard drives into computers that are internet-connected makes the media as vulnerable as if they were network-connected themselves.

Researchers who choose local storage to prevent data theft should take a number of additional steps. First, they should ensure that data storage is as local *in fact* as it is in theory. Data stored on a machine that is backed up by a service like Code42/Crashplan or Carbonite, for example, is not “locally stored” in the sense that it is inaccessible from other machines. Many universities install backup software on machines they own, giving them the same vulnerabilities as cloud storage. Second, researchers should limit remote access to their machines. Again, many universities and other organizations set up university-owned computers to allow for remote IT support. Machines used to store sensitive data should disable remote access, restrict most incoming connections, and perhaps disable the function that acknowledges/responds to “pings” or other network connection attempts (in Mac OS X this can be done in the Security and Privacy settings, in Microsoft systems, use the “Defender” firewall in settings). Of course, individual computers are not particularly likely targets for theft when compared to cloud storage platforms, especially to adversaries who do not know what is on the computer. The low probability of becoming a target in the first place might be a point in favor of using local storage workflows for data security.

Air-Gapped Storage: Storing sensitive data on a server that is not connected to a network (no wi-fi, no ethernet cable, no connection to other computers) is by far the most cautious workflow to reduce the risk of theft via networks. This design establishes an “air gap” (physical separation) between the sensitive data and any devices that are connected to an external network. This workflow is extremely inconvenient: accessing data requires physical access to the server, and transferring non-sensitive data from the server to another machine

requires physical storage media and a protocol for re-formatting the media after each use.⁷

An isolated server design is also somewhat expensive and extremely vulnerable to hardware failure. The security benefits of the system hinge on the fact that sensitive data are stored *only* on the secure server, so, if the server fails the data are likely lost. Running a reliable server on a local network is not a trivial task, so this workflow is probably more feasible for labs, departments, or research centers than individual researchers.

An inexpensive, modular air-gapped system is theoretically possible to build using inexpensive single-board computers like the Raspberry Pi. A terminal could be built where individual users could retrieve their Raspberry Pi (Versions 1 and 2 are not internet-capable without an after-market wifi adapter) from a locker, plug it into a power supply, keyboard, and monitor, and manipulate sensitive data on the network-isolated terminal. Since Raspberry Pis are woefully under-powered for doing even simple statistical analyses, this air-gapped terminal would function somewhat like a “glove box” in a laboratory. Sensitive data could be stored in its original form on the non-network connected computer, but researchers could also use the computer to separate PII from non-identifiable data—in effect de-contaminating the data so that it became safe to transport to a network-connected computer—and then extract non-identifiable data for analysis while leaving PII behind in the secure system. This “glove box” option might be a good compromise between the high cost of a true secure network and the low computing power of the Raspberry Pi. Even air-gapping has recently been shown to be vulnerable to particularly sophisticated attacks, but such attacks seem unrealistically difficult to execute, and thus present relatively little risk ([Callan, Zajic and Prvulovic, 2015](#); [Wolff, 2013](#)).⁸

Preventing Theft when Working with Partners: The viability of theft-prevention tools

⁷Cleaning the storage media used to transfer data on and off the isolated system is necessary to keep the system isolated. A famous attack that used “dirty” storage media to access an air-gapped system was the Stuxnet worm, a U.S. Government-created computer virus that was used to destroy centrifuges at an Iranian uranium enrichment facility. The Stuxnet worm reached the facility’s air-gapped internal system via USB drive ([Langner, 2011](#)).

⁸Thanks to Andy Halterman for this note of caution, and for discussing the proposed system.

described above for research with partners depends both on the capabilities of the partner and on the structure of the partnership agreement. For research projects where the social scientist's home institution and the partner's office are in different cities (or countries) for instance, a network-isolated or air-gapped system is likely impractical unless the partnership agreement does not require that raw data be accessible by the practitioner partner. In some cases, negotiating an agreement that allows for the use of more secure, less convenient storage is not difficult because practitioner partners are primarily interested in data-derived products like reports and analyses, and are interested in having academic partners create those products. If social scientists work proactively with their partners to identify the analyses and deliverables that the partner wants, and then budget time and resources to deliver those products to the partner, they may be able to negotiate a data management agreement that does not require raw data sharing, thereby making more secure and restrictive data security protocols practical. If partner organizations need access to raw data, cloud storage is likely to be the solution that gives the researcher the most control and ability to monitor data security. Data that are shared via email attachment or physical storage media cannot be monitored after the fact, but cloud storage platforms like DropBox allow file owners to monitor access.

Preventing Legal Expropriation

Useful ways to address the threat of legal expropriation can be grouped into three activities: seeking immunity, attempting to tie one's hands such that compliance is impossible, and misdirection. The risk proposition for preventing legal expropriation is very different from the risks associated with theft prevention. Attempts to prevent legal expropriation all basically shift risk of harm or adverse consequences *from* respondents whose personal information is recorded *to* the research team. Social scientists trying to prevent legal expropriation of data have to think about the risks they are personally assuming when they try to avoid legal expropriation of data, but more importantly they need to think of the risks with which they

are encumbering practitioner partners and research staff. Decisions about legal and coercive risks *must* be discussed and made collaboratively for research projects that involve partners.

Further complicating the picture: the various tools to prevent legal expropriation are not interchangeable. If U.S. citizens attempt to protect research data by misdirecting investigators *in the United States*, for instance, they may be committing a crime. Conversely, if researchers were to seek immunity in order to protect research data gathered in any of a number of countries, doing so might have two negative effects: tipping off the government to an opportunity for expropriation of sensitive data, and drawing attention to the research project, thereby exposing respondents, staff, and partners to a greater threat of harassment or coercion. Hand-tying strategies have ambiguous consequences for researchers and partners. For U.S. researchers in non-U.S. contexts, the ability to say “I do not have access to the data you want” might be powerful, especially in countries that do not have mutual legal aid treaties with the United States. At the same time, non-compliance can be risky in terms of physical safety for the researcher in the short term, but more durably (and probably more seriously) for local staff and partner organizations that depend on good relationships with their government in order to operate. These possible consequences must be carefully weighed.

Legal Waivers/Seeking Immunity: For IRB-approved research taking place in the United States, researchers may be eligible to apply for a confidentiality waiver from the National Institutes of Health ([National Institutes of Health, 2017](#)).⁹ This certificate is designed to protect identifiable, sensitive information from subpoena by U.S. courts; in other words, it gives researchers grounds on which to defy a legal demand to release identifiable information covered by the certificate. The NIH certificate application process is no more onerous than the IRB approval that researchers using PII already undergo, and the legal protections it promises are certainly better than the status quo (which provides no specific protection for

⁹Per the policy guide on their website, NIH certificates of confidentiality can only be issued for health-related research. Many issues that social scientists study, however, could be considered to be applicable to public health research.

confidential research data). It is important to know, though, that NIH certificates have, in some instances, been insufficient legal tools to prevent the disclosure of PII in criminal and civil proceedings ([Beskow, Dame and Costello, 2008](#); [Bryon Adinoff and Chezem, 2013](#)). Outside the United States, researchers should consult with partners about the wisdom of voluntarily disclosing information about a research project to a government agency.

Hand-Tying Strategies: A researcher can do a variety of things to “tie her hands” and make it impossible to comply with a demand for data. One low-cost and technologically simple tool for researchers collecting data in a *different legal jurisdiction from their home institution* uses PGP (“Pretty Good Privacy” - an open-source encryption software commonly used by journalists) to set up a “vault” into which researchers can deposit information. PGP encryption uses a pair of keys (strings of alphanumeric characters) called the “public key” and the “private key” to encrypt and decrypt information. The intuition behind the encryption is that any files encrypted using a particular public key can *only* be decrypted using the corresponding private key (See [Foundation, 2014](#), for a good introduction). While PGP is most commonly used to encrypt email traffic (B sends a message to A that is encrypted with A’s public key; A uses her private key to decrypt B’s message), social scientists can use it to create a “vault” that they can deposit into easily, but cannot access in the field.

To create a vault, a researcher must first generate a key pair,¹⁰ and then stores the private key on a local drive, at their home institution or somewhere else that is not accessible during data collection. It is crucial that the private key *not* be accessible to the researcher once she is in the field; this means it cannot be stored on the cloud, available in an email, or carried with the researcher on local media like a USB drive. Once in the field, the researcher can use software like GNU Privacy Guard (see footnote 11) to encrypt data using the public key, and then either send that encrypted data in an email, upload the encrypted data to the cloud, or simply keep it on her hard drive. No matter where the encrypted data are stored,

¹⁰A number of different software packages can be used to generate key pairs and manage encryption. Two popular, and well-regarded implementations of the PGP framework are [GNU Privacy Guard](#) and [OpenPGP](#).

they cannot be decrypted without the private key. After encryption, the researcher destroys the unencrypted data. At this point, the researcher has effectively *tied her hands*: she no longer has unencrypted PII, and she does not have access to the key that could decrypt the files until she returns to her home institution. The PGP lock-box can even be implemented without internet access, if the encrypted files are stored on the hard-drive where they are encrypted (This “lock-box” protocol comes from [Trinh, Milliff and Wolters, 2017](#)). Other methods could theoretically serve the same purpose without using computer encryption. A researcher could mail local media back to her home institution, for example. A PGP-based system, though, can be easily backed up to guard against data loss from hardware failures without increasing the threat of decryption. In this way it serves as a nice balance between preventing data loss and preventing unwanted data access.

Failing to comply with demands for data through hand-tying strategies described here is a short-term solution until the researcher can either leave the jurisdiction of whoever wants the data or mount a legal challenge of their own. This stop-gap measure has downsides for social scientists—using it to protect sensitive data outside one’s home country likely jeopardizes future visa applications—but the potential downsides are more serious for partner organizations and staff, who are less likely able to avoid legal threats or harassment in the long term. Again, a bare minimum standard for researchers interested in using hand-tying strategies to avoid legal expropriation of sensitive data should be that everyone involved in the research is *aware* of the strategy and consents to its use. Social science researchers might be able to modestly decrease the chance that partners and staff are punished for hand-tying if their data sharing agreement leaves local partners (or even research staff) unaware of the particulars of what are in the data, but research partners should be included in a discussion of risks both as a matter of equity and because they are likely to have better knowledge of the context-specific risks.

Misdirection: Especially when working outside the United States, researchers might consider creating artifacts of politically innocuous research projects, disclosure of which would

not compromise confidential data (Trinh, Milliff and Wolters, 2017). Having decoy data, or even better, real data from a separate project that contains no sensitive information, could plausibly be enough to satisfy inquiring police officers, immigration officials, or other government representatives. This remedy is unlikely to weather robust scrutiny, and, being caught in a lie has obvious risks. At the same time, it requires little technological sophistication to implement, and, in combination with a file-naming protocol that obscures real data (or, better yet, in combination with the PGP lock-box described above), it provides an additional layer of security in exchange for increased risk. A safer version of the same strategy is to pre-specify innocuous language that *accurately* describes the substance of a research project while making it appear uninteresting and politically irrelevant. Many academics may find that such descriptions come to them quite naturally. Research partners, or academics working at institutions in the country where data is being collected, are likely to be good sources of guidance on how to describe research in ways that government officials find uninteresting.

Preventing De-anonymization

Measures to prevent de-anonymization of release data must, as described above, do far more than simply remove PII like names and addresses. Given the complication and seriousness of the de-anonymization threat to data security, a number of security researchers and social scientists have developed methods to guard against de-anonymization. Three such methods are briefly reviewed here. These three methods are aimed at preventing de-anonymization of publicly available, anonymized versions of sensitive data. As noted above, though, researchers working with partners must contend with additional risks to respondent privacy posed by research partners with strong contextual knowledge that could be used to identify respondents in nominally anonymous data. Researchers working on projects where a) data will be shared with practitioner partners and b) respondents have not explicitly consented to personally-identifiable data being shared with the practitioner partner should plan to take these or other de-anonymization prevention steps before sharing data internally.

Statistical Disclosure Control and k -anonymity: One of the earliest anonymization algorithms in the digital age, by [Samarati and Sweeney \(1998\)](#), propose a strategy for preventing de-anonymization that relies on modifying identifiable data such that no value of any attribute in the data is shared by fewer than k records (see also [Sweeney, 2002, 2007](#)). Take, for example, the attribute “age.” If, in a dataset, no individual value for “age” appears for fewer than three records, then the data has 3-anonymity with respect to age. This principal is more commonly implemented with respect to “quasi-identifier tuples”, or combinations of attributes that could collectively lead to identification like age-gender-birthplace, or the above mentioned ZIP Code-birth date-gender. Data are k -anonymous with respect to the quasi-identifier tuple ZIP Code-birth date-gender if for any combination of the three attributes found in the dataset, there are at least k records with the same exact combination. K-anonymity is *manufactured* in identifiable data by suppressing values of identifiable attributes (replacing either unique names or all names with NA), or by generalizing values of identifiable attributes (replacing all birth years with birth decades).

K-anonymization has some drawbacks. First, since anonymization is accomplished by redacting real records that correspond directly to actual people, adversaries can still make inferences about individuals they know to exist *somewhere* in a k-anonymized dataset. If adversaries are looking at a sociology survey to discover the sexual orientation of “Albert,” male, age 35, ZIP Code 60637, and known to have responded to the survey, the adversary can look at sexual orientation for all records that match Albert’s quasi-identifier tuple and infer from those data the probability that Albert is gay. Advancements to the k-anonymization framework (including tools implemented in the software package `sdcMicro` demonstrated in an appendix) provide ways to quantify the risk of re-identification for particular respondents in k-anonymized data, allowing researchers to more consistently gauge the efficacy of the tools. Besides this problem, K-anonymization is hard to implement in high-dimensional data sets, where the unicity of quasi-identifier tuples is remarkably high ([de Montjoye et al., 2013](#)).

Finally, since the suppression tactic used to create K-anonymization pursues unique values (data can be k-anonymized by suppressing unique values of an attribute rather than the whole attribute), the process disproportionately censors outlier data, which can change the distributional characteristics of K-anonymized data relative to its identifiable form ([Angiuli, Blitzstein and Waldo, 2015](#)). With these caveats in mind, K-anonymity is an attractive solution because it is intuitive, relatively easy to implement on a given dataset, and widely used.

A related tool, part of the broader research area around Statistical Disclosure Control (SDC: tools to prevent re-identification of individuals in microdata), focuses on aggregation: limiting both the geographic and quantitative resolution at which data are reported. Techniques for decreasing data resolution while maintaining analytical validity often come from government statistical services like the Bureau of the Census or the General Accounting Office ([Kingsbury, 2001](#)). Like K-anonymity, these techniques are oriented toward eliminating unique records in publicly released data. The Census, for example, limits the geographic resolution at which results are available. The smallest U.S. Census geographic unit, the “block”, has only a few hundred residents and covers as little as 30,000 square feet. In order to prevent sensitive data (like income) from being linked to individuals, the Census does not report most demographics at the block level. Another technique which focuses on limiting the quantitative resolution of data is “top-coding,” or binning all responses values that are in the tails of the distribution of responses such that no bin has only one response in it. This is most common with income data, where brackets are evenly sized up to a certain ceiling, above which there is only one bracket for “\$d or more”. Limiting resolution is an effective technique for preventing adversaries from identifying individual records in a dataset, but like many SDC tools, it buys security at the cost of analytical value or “informativeness” of the data. For academic researchers, aggregated data may not satisfy an obligation to produce replication data for a paper that relies on analyses conducted using respondent-level data. Aggregation can be done carefully such that the summary statistics of a dataset are well pre-

served, but it necessarily obliterates high-leverage observations which may be major drivers of the results of statistical analysis.

Simulation, Noise, and Differential Privacy Algorithms: A different set of techniques used to prevent de-anonymization more radically change the raw data. Instead of releasing a redacted and de-identified version of data, they argue, the safest course of action is to either simulate a new version of the data that produces the same statistical inferences as the original data but contains no records that correspond to real individuals, or to alter the data by introducing randomness or switching values such that de-anonymization is more difficult. These methods rely on one of two statistical principals, neither of which will be discussed in great detail here. The first class, multiple imputation methods, proposes releasing data sets that are either fully simulated ([Rubin, 1993](#); [Raghuathan, Reiter and Rubin, 2003](#)) or partially simulated ([Reiter, 2003](#)) from the actual population of the original dataset. These methods should, for certain types of statistical analysis, yield results that are similar to those of the original dataset, satisfying *most* users, but certainly not all. Raghuathan et al. propose that users with specific needs or users who wish to conduct unusual analyses would need to apply for access to the original, confidential data rather than use the simulated public version. Imputation methods completely eliminate the possibility of de-anonymization since no real individuals exist in the data to be de-anonymized. They do, however, raise significant barriers to entry when it comes to using the data. The available frameworks require sophisticated statistical analysis to replicate results from the original data. Variance estimates for the framework in Raghuathan et al., for example, require a Bayesian bootstrap, which is computationally intensive and non-trivial to implement.

The other class, differential privacy methods, mask values of numerical data using additive noise or value switching. A number of available differential privacy algorithms show that noise is added in such a way as to preserve the original data distribution, while obscuring individual observations ([Kargupta et al., 2003](#)). In addition, a team of social scientists is cur-

rently developing a privacy-preserving data-sharing protocol that uses a differential privacy architecture that will eventually be built into the Harvard Dataverse infrastructure such that social scientists uninterested in learning the math behind differential privacy can still reap its benefits (Gaboardi et al., 2016; Evans et al., 2020). This framework shows substantial promise, and, in a relative advance compared to multiple imputation, can be analyzed as if it were the raw data. Both of these techniques though, may temporarily suffer from the fact that they are based on advanced statistics and are somewhat difficult to implement pending specialized packages for common statistical software. The professional incentives for their use depend largely on whether or not the social science academic community as a whole perceives them as credible substitutes for un-altered replication files. The trade off proposed by these methods is excellent privacy protection at the potential cost of skepticism .

Maintaining Anonymity in Text and Qualitative Data: The de-anonymization prevention tools described above are aimed at preventing re-identification in tabular data, but not all social science research creates or uses tabular data to test hypotheses. A substantial body of research analyzes quasi-identifying or PII-containing data that are either naturally represented in text (historical or legal documents, social media data), or can be coerced into text (interviews). Text data that encode sensitive or quasi-identifying information are often very easy to de-anonymize given basic contextual knowledge. Picking out unique quasi-identifier tuples in text may not be as easy as in tabular data, but text data and natural language can be uniquely identifying in its pragmatics (context, implication, etc) even if identifying data have been removed from the semantics (words) and syntax (organization of words). Think how easy it is to have a conversation in which all speakers know the conversation is about a third party, without ever saying words or phrases that positively identify that third party.

Many uses of text data in social sciences in both qualitative and quantitative empirical research do not pose difficult data security challenges, either because they analyze corpora

of text that are already public (Katagiri and Min, 2019), or because they analyze very narrowly-scoped text like text responses to specific survey questions that makes identification difficult by design (Roberts et al., 2014). Other projects use data that are semi-public (like tweets), and are easy to link to a unique identity, but which are treated more like private or confidential data as norms around social media data evolve. Finally others use clearly private, longer narrative texts like transcripts of interviews, for work that is traditionally analyzed qualitatively. For these later two categories of text or natural language data, researchers need to pay attention to de-anonymization concerns when sharing materials to replicate analyses, especially with practitioner partners who may be well equipped to identify respondents using contextual knowledge.

I present a new workflow that applies a widely used, unsupervised text analysis model (the Structural Topic Model by Roberts et al. (2013)) and its corresponding user-friendly R package (Roberts, Stewart and Tingley, 2018), to conduct reproducible and privacy-protecting analysis and presentation of patterns in corpora of sensitive text that cannot be shared in full without risking de-anonymization of respondents. Topic modeling for privacy protection uses a standard and increasingly popular toolkit for a fundamentally different purpose: most unsupervised models for text analysis are used for dimension reduction and efficient comparison of documents in corpora of text that are so large that a researcher or team could not feasibly read and analyze them. I use topic modeling for dimension reduction and efficient comparison of documents in a corpus that is quite small (in fact, so small that a single researcher could transcribe it all from audio), but for which presentation of raw, high-dimensional data without some “flattening” poses a threat to the privacy of the speakers represented in the text.

Topic modeling helps identify patterns in the contents of documents under a set of assumptions about the relationship between semantic choice and meaning: topic models (starting with Latent Dirichlet Allocation in Blei, Ng and Jordan (2003)) model the appearance of a given word in a document as a function of some latent or unobserved category, a “topic”

that the word is used to describe. A fitted topic model produces summaries for each document: a vector (summing to 1) of topic proportions which describes the prevalence of each latent category in a document. Per [Grimmer and Stewart \(2013\)](#), identification of the substantive meaning of a topic/cluster returned by the model is the responsibility of the researcher, not the model. The topic prevalence can be compared across documents to identify patterns in the ways that topics relate to each other—when a document discusses topic 1, it is also likely to discuss topic 8—and with structural topic models, but not all other varieties of topic model, topic prevalence can be related to document metadata to identify further patterns—respondents over the age of 35 have higher topic prevalence for topic 1 than respondents under the age of 35.

Topic modeling is useful for privacy protection because it focuses exclusively on *morphologic* patterns (words and their meanings), discarding both semantic and pragmatic information in natural language. The data format that topic models ingest (and thus, the primary data format that needs to be shared in order to make topic modeling transparent and reproducible) is a document-term matrix: a matrix that represents how many times each term in the vocabulary of the entire corpus is used in each document. A single-sentence document like “this is a test” would be represented as a vector with 1s for the terms “this”, “is”, “a”, “test” and zeroes for all other terms that appeared in the totality of documents.¹¹ Crucially, the vector for the documents “this is a test” and “test is a this” are identical: word order is not encoded in a document term matrix.

Because syntax is stripped from the documents used in topic modeling, it is extremely difficult to re-assemble the original natural language given only a document term matrix. For longer documents (i.e. documents that are multiple sentences and thus contain multiple verbs, multiple subjects, etc.), re-assembling the original document from a DTM is practically

¹¹Three of the four words would probably be dropped in pre-processing of documents. Most topic models discard extremely common “stop words” because their appearance says very little about content. Pre-processing also usually drops terms that appear only in one document because those terms are also uninformative for model fitting. This has the nice side effect of making it hard to re-identify respondents based on known vocabulary quirks.

impossible.¹² This means that a document term matrix, so long as none of its constituent terms are themselves identifiers, is hard to de-anonymize.

Topic modeling, in sum, is a powerful dimension reduction tool to show patterns in sensitive text, allowing researchers to show, without disclosing sensitive information that risks de-anonymizing research subjects, that the conclusions they are drawing from qualitative analysis of a sensitive corpus are actually reflected in reproducible patterns that obtain across all documents. Below, I show an example of this new use of topic modeling, adapted from [Milliff \(2020\)](#).

Conclusion

Ensuring the security of respondent’s sensitive information while fulfilling professional obligations to transparency is an evolving challenge. As the popularity of collecting and analyzing large, identifiable data sets grows in disciplines like political science, the ethical and professional consequences of a potential data breach grow as well. Examples from the academy in the last two decades ([Venkatesh, 2008](#); [McMurtrie, 2014](#), among others) already hint at the grave consequences that the release of sensitive data can have for research subjects; with these examples in mind, the academy should not be content to wait for an even larger crisis to prompt re-examination data security expectations in social science research.

This essay has proposed a preliminary road-map for that re-examination. It has proposed a set of obligations that researchers have to their colleagues and their respondents, identified factors that threaten researchers’ ability to fulfill their obligations, explored how the growing phenomenon of practitioner-academic collaboration amplifies threats to data security, and finally discussed some potential ways to address these threats. It is eminently possible to reduce the risk of data security failures. Whether or not better practices are ultimately adopted, though, depends largely on whether social science disciplines incentivize good data

¹²Our example sentence would actually be easy to re-assemble under the assumption that the original document followed standard grammatical conventions. There is only one likely order in which the four words in the example would appear in standard English.

security practices and tolerate the compromises that such practices require.

References

- American Political Science Association. 2012. “A Guide to Professional Ethics in Political Science.”
URL: <http://www.apsanet.org/portals/54/Files/Publications/APSAEthicsGuide2012.pdf>
- American Political Science Association. 2015. *Replication Data and Research Transparency*. Washington, D.C.: American Political Science Association.
- American Sociological Association. 2017. “A Real Case Involving the Protection of Confidential Data.” Report.
URL: <http://www.asanet.org/teaching-learning/faculty/teaching-ethics-throughout-curriculum/case-99-real-case-involving-protection-confidential-data>
- Angiuli, Olivia, Joe Blitzstein and Jim Waldo. 2015. “How to De-Identify Your Data.” *ACM Queue* 13(8).
- APSA Ad Hoc Committee on Human Subjects Research. 2020. Principles and Guidance for Human Subjects Research. Technical report American Political Science Association Washington, D.C.: .
URL: https://www.apsanet.org/Portals/54/diversity%20and%20inclusion%20prgms/Ethics/Final_-_Principles%20with%20Guidance%20with%20intro.pdf?ver=2020-04-20-211740-153
- Beskow, Laura M, Lauren Dame and E Jane Costello. 2008. “Certificates of Confidentiality and the Compelled Disclosure of Research Data.” *Science* 322(5904):1054–1055.
- Blei, David, Andrew Ng and Michael Jordan. 2003. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research* 3(993-1022).
- Brandom, Russell. 2014. “New documents reveal which encryption tools the NSA couldn’t crack.” *TheVerge.com* .
URL: <https://www.theverge.com/2014/12/28/7458159/encryption-standards-the-nsa-cant-crack-pgp-tor-otr-snowden>
- Bryon Adinoff, Robert R. Conley, Stephan F. Taylor and Linda L. Chezem. 2013. “Protecting Confidentiality in Human Research.” *American Journal of Psychiatry* 170(5):466–470.
- Callan, Robert, Alenka Zajic and Milos Prvulovic. 2015. FASE: Finding Amplitude-modulated Side-channel Emanations. In *Proceedings of the 42nd International Symposium on Computer Architecture*. Georgia Institute of Technology Atlanta, Georgia: pp. 592–603.
- Chetty, Raj, Nathaniel Hendren, Patrick Kline and Emmanuel Saez. 2014. “Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States.”
- Dalenius, Tore. 1986. “Finding a Needle in a Haystack, Or: Identifying Anonymous Census Records.” *Journal of Official Statistics* 2(3):329–336.
- de Montjoye, Yves-Alexandre, César A. Hidalgo, Michel Verleysen and Vincent D. Blondel. 2013. “Unique in the Crowd: The privacy bounds of human mobility.” *Scientific Reports* 3(1).

- Desai, Sonalde and Reeve Vanneman. 2015. “India Human Development Survey II (IHDS-II).”
URL: doi.org/10.3886/ICPSR36151.v2
- Ditto, Steven and Larisa Baste. 2014. *Infographic: Iranian Students in the United States*. Washington, D.C.: Washington Institute for Near East Policy.
- Evans, Georgina, Gary King, Margaret Schwenzfeier and Abhradeep Thakurta. 2020. Statistically Valid Inferences from Privacy Protected Data. Working paper, Harvard University Cambridge, MA: .
- Foundation, Electronic Frontier. 2014. “An Introduction to Public Key Cryptography and PGP.”
- Fujii, Lee Ann. 2012. “Research Ethics 101: Dilemmas and Responsibilities.” *PS: Political Science & Politics* 45(04):717–723.
- Gaboardi, Marco, James Honaker, Gary King, Kobbi Nissim, Jonathan Ullman and Salil Vadhan. 2016. “PSI: A Private Data Sharing Interface.”
- Gibney, Elizabeth. 2017. “Internet Research Triggers Scrutiny.” *Nature* 550:16–17.
- Gilmore, Janet. 2016. “Campus alerting 80,000 individuals to cyberattack.”
URL: <http://news.berkeley.edu/2016/02/26/campus-alerting-80000-individuals-to-cyberattack/>
- Grimmer, Justin and Brandon M. Stewart. 2013. “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.” *Political Analysis* 21(3):267–297.
- Hansen, Marit, Peter Berlich, Jan Camenisch, Sebastian Clauß, Andreas Pfitzmann and Michael Waidner. 2004. “Privacy-enhancing identity management.” *Information Security Technical Report* 9(1):35–44.
- Henriksen, Dorte. 2016. “The rise in co-authorship in the social sciences (1980–2013).” *Scientometrics* 107(2):455–476.
- Identity Theft Resource Center. 2016. Data Breach Report 2016. Technical report Identity Theft Resource Center.
- Information Commissioner’s Office. 2017. Preparing for the General Data Protection Regulation (GDPR): 12 Steps to Take Now. Technical report United Kingdom Information Commissioner’s Office London: .
- Kargupta, Hillol, Souptik Datta, Qi Wang and Sivakumar Krishnamoorthy. 2003. Random Data Perturbation Techniques and Privacy Preserving Data Mining. In *IEEE International Conference on Data Mining*. IEEE.
- Katagiri, Azusa and Eric Min. 2019. “The Credibility of Public and Private Signals: A Document-Based Approach.” *American Political Science Review* 113(1):156–172.

- Kinch, Nathan. 2016. *Data Minimization - The New Design Rules for Startups*. TechCrunch.com.
- King, Gary. 1995. "Replication, Replication." *PS: Political Science and Politics* 28:443–499.
- King, Gary. 2007. "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing." *Sociological Methods and Research* 36:173–199.
- King, Gary and Nathaniel Persily. 2019. "A New Model for Industry–Academic Partnerships." *PS: Political Science amp; Politics* p. 1–7.
- Kingsbury, Nancy R. 2001. Record Linkage and Privacy: Issues in Creating New Federal Research and Statistical Information. Technical Report GAO-01-126SP United States General Accounting Office Washington, D.C.: .
- Knerr, Charles R. 1982. *What To Do Before and After a Subpoena of Data Arrives*. New York, NY: Springer New York pp. 191–206.
- Kung, Antonio. 2014. PEARs: Privacy Enhancing ARchitectures. In *Privacy Technologies and Policy*. Number LNCS 8450 Second Annual Privacy Forum, APF 2014 Athens, Greece: Springer.
- Langner, R. 2011. "Stuxnet: Dissecting a Cyberwarfare Weapon." *IEEE Security Privacy* 9(3):49–51.
- Lewis, Kevin, Jason Kaufman, Marco Gonzalez, Andreas Wimmer and Nicholas Christakis. 2008. "Tastes, ties, and time: A new social network dataset using Facebook.com." *Social Networks* 30(4):330 – 342.
- Lupia, Arthur and Colin Elman. 2014. "Openness in Political Science: Data Access and Research Transparency: Introduction." *PS: Political Science and Politics* 47(1):19–42.
- Markham, Annette and Elizabeth Buchanan. 2012. "Ethical Decision Making and Internet Research: Recommendations from the AoIR Ethics Working Committee."
- McCallister, Erika, Tim Grance and Karen Scarfone. 2010. Guide to Protecting the Confidentiality of Personally Identifiable Information (PII). NIST Special Publication 800-122 National Institute of Standards and Technology Gaithersburg, MD: .
- McMurtrie, Beth. 2014. "Secrets from Belfast." *Chronicle of Higher Education* (January):1–41.
- Menoret, Pascal. 2014. Repression and Fieldwork. In *Joyriding in Riyadh*. New York: Cambridge University Press.
- Mervis, Jeffrey. 2014. "How Two Economists Got Direct Access to IRS Tax Records." **URL:** <https://www.sciencemag.org/news/2014/05/how-two-economists-got-direct-access-irs-tax-records>

- Milliff, Aidan. 2020. Facts Shape Feelings: An Information-Based Framework for Emotional Responses to Trauma. Working Paper, 2019-06 Massachusetts Institute of Technology Cambridge, MA: .
- Moravcsik, Andrew. 2015. *One Norm, Two Standards: Realizing Transparency in Qualitative Political Science*. Houston, TX: Society for Political Methodology.
- Narayanan, Arvind and Vitaly Shmatikov. 2008. Robust De-anonymization of Large Sparse Datasets. In *IEEE Symposium on Security and Privacy*. IEEE pp. 111–25.
- Narayanan, Arvind and Vitaly Shmatikov. 2009. De-anonymizing Social Networks. In *30th IEEE Symposium on Security and Privacy*. IEEE.
- National Institutes of Health. 2017. “Certificates of Confidentiality: Background Information.”
- National Science Foundation. 2017. “Interpreting the Common Rule for the Protection of Human Subjects for Behavioral and Social Science Research.”
URL: <https://www.nsf.gov/bfa/dias/policy/hsfaqs.jspthird>
- Neal, Jennifer Watling, Zachary P. Neal, Mariah Kornbluh, Kristen J. Mills and Jennifer A. Lawlor. 2015. “Brokering the Research-Practice Gap: A typology.” *American Journal of Community Psychology* 56(3-4):422–435.
- Ohm, Paul. 2010. “Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization.” *UCLA Law Review* 58(3):1701–1778.
- Radden-Keefe, Patrick. 2018. *Say Nothing: A True Story of Murder and Memory in Northern Ireland*. New York: Penguin Random House.
- Raghunathan, Trivellore E, Jerome P Reiter and Donald B Rubin. 2003. “Multiple imputation for statistical disclosure limitation.” *Journal of official statistics* 19(1):1.
- Reiter, Jerome P. 2003. “Inference for Partially Synthetic, Public Use Microdata Sets.” *Survey Methodology* 29:181–88.
- Roberts, Margaret E., Brandon M. Stewart and Dustin Tingley. 2018. “stm: R Package for Structural Topic Models.” *Journal of Statistical Software* .
- Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G. Rand. 2014. “Structural Topic Models for Open-Ended Survey Responses.” *American Journal of Political Science* 58(4):1064–1082.
- Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley and Edoardo M. Airoldi. 2013. The Structural Topic Model and Applied Social Science. In *NIPS 2013 Workshop on Topic Models: Computation, Application, and Evaluation*. NIPS.
- Rubin, Donald B. 1993. “Discussion: Statistical Disclosure Limitation.” *Journal of official statistics* 9:462–8.

- Samarati, Pierangela and Latanya Sweeney. 1998. “Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression.”
- Scarce, Rik. 2005. *Contempt of Court: A Scholar’s Battle for Free Speech from Behind Bars*. Lanham, MD: Rowman and Littlefield.
- Schaar, Peter. 2010. “Privacy by Design.” *Identity in the Information Society* 3(2):267–274.
- Shilton, Katie. 2016. “Emerging Ethics Norms in Social Media Research.”
- Stanford University Information Security Office. 2005. Disk and Data Sanitization Policy and Guidelines. Technical report Stanford University Palo Alto: .
- Summers, Scott. 2016. *Organising, Storing and Securely Handling Research Data*. Essex, England: UK Data Service.
- Sweeney, Latanya. 2000. “Simple Demographics Often Identify People Uniquely.”
- Sweeney, Latanya. 2002. “k-anonymity: A model for protecting privacy.” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10(05):557–570.
- Sweeney, Latanya. 2007. “Systems and methods for deidentifying entries in a data source.” US Patent 7,269,578.
URL: <https://www.google.com/patents/US7269578>
- Tastes, Ties, and Time Dataverse*. 2010.
URL: <https://dataverse.harvard.edu/dataverse/t3>
- Templ, Matthias, Bernhard Meindl and Alexander Kowarik. 2020. Introduction to Statistical Disclosure Control SDC. Technical report Zurich University of Applied Sciences Zurich, Switzerland: .
URL: https://cran.r-project.org/web/packages/sdcMicro/vignettes/sdc_guidelines.pdf
- The OpenDP Team. 2020. The OpenDP White Paper. Technical report Harvard University Cambridge, MA: .
URL: https://projects.iq.harvard.edu/files/opendp/files/opendp_white_paper_-_11may2020.pdf
- Traynor, Michael. 1996. “Countering the Excessive Subpoena for Scholarly Research.” *Law and Contemporary Problems* 59(3):119–48.
- Trinh, Minh Duc, Aidan Milliff and Lukas Wolters. 2017. “Data Security.” Working Paper.
- Turner, Karen. 2016. “Hacked Dropbox login data of 68 million users is now for sale on the dark Web.”
- Venkatesh, Sudhir. 2008. *Gang Leader for a Day*. New York: Penguin.

- Vitak, Jessica, Katie Shilton and Zhara Ashktorab. 2016. Beyond the Belmont Principles: Ethical Challenges, Practices, and Beliefs in the Online Data Research Community. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing*. Association for Computing Machinery San Francisco, CA: pp. 941–53.
- Waterston, Alisse. 2009. “Exoticizing the Other and the Author.” *North American Dialogue* .
- Williams June, Audrey. 2017. “Publicize your Research.” *The Chronicle of Higher Education* .
- Wolff, Josephine. 2013. *Great, Now Malware Can Jump the “Air Gap” Between Computers*. Slate.com.
- Wood, Elisabeth Jean. 2007. Field Research. In *The Oxford Handbook of Comparative Politics*. Oxford: Oxford University Press.
- Zechmeister, Elizabeth J. 2016. Ethics and Research in Political Science: The Responsibilities of the Researcher and the Profession. In *Ethics and Experiments*, ed. Scott Desposato. London: Routledge.
- Zimmer, Michael. 2008. “More on the ‘Anonymity’ of the Facebook Dataset - It’s Harvard College.”
URL: <http://michaelzimmer.org/2008/10/03/moreon-the-anonymity-of-the-facebook-dataset-its-harvard-college/>

Supplemental Information: Three Privacy Protection Demos

PGP “LockBox”

Using PGP (pretty good privacy) encryption to make research data temporarily inaccessible is a good way combat threats of theft and expropriation of sensitive data. PGP works across all major operating systems, and can successfully encrypt many types of files, including .csv tabular data, many types of text files, and .mp3 audio files. PGP is useful for generating temporary inaccessability because it uses one key (the “public key”) to encrypt data, and a separate key (the “private key”) to decrypt. Data encrypted with a particular public key can *only* be decrypted using the particular private key that matches it. The two keys together are called a “key pair.” As far as the maintainers of PGP’s open-source implementation know, the encryption standard has not been broken, though some commercial tools that use PGP have had flaws ([Brandom, 2014](#)).

When a user has access to a public key, but not the corresponding private key, they can encrypt data and then transport or copy the encrypted data as they please, but they cannot reverse the encryption process. For this reason, PGP is often used by journalists who want sources to be able to share private information via otherwise unsecure channels like email.

A PGP lockbox for social science research serves a slightly different purpose with the same basic tools. Whereas PGP encryption is normally used to transfer data such that it is inaccessible to anyone but the target recipient, social scientists can use the same standards to transport and store data such that it is temporarily inaccessible to *everyone*, by storing the private key somewhere that it cannot be applied to the data while data collection is ongoing (i.e. local storage on a computer at a researcher’s home institution). The rest of this section shows step-by-step instructions for setting up and using a PGP lockbox to encrypt sensitive data and make it temporarily inaccessible to all parties, including the researcher.¹³

¹³Instructions and screenshots are specifically for GPG Tools on Mac OSX, but the process is similarly simple using Windows and Linux. Good resources for both exist online.

Ingredients:

- Two computers (any OS, any variety)
 - One computer remains at home institution
 - Second computer used during data collection
- Open PGP Software like [Gnu Privacy Guard/GPG Tools](#)

Setup:

1. Download and install GPG Tools or other software that implements the OpenPGP framework onto **both computers**
2. Using the **computer that will not be carried during data collection**, open the GPG Keychain application, click “new” in the top left corner, and follow instructions to generate a new key pair with the default options.

Figure 1: Creating a new keypair in GPG Keychain for Mac OS X

3. Once the key is successfully created, a prompt will ask you to upload the key pair to a key server, where other users can find your *public* key, and encrypt files with it so that only you (using your new private key) can decrypt. *Unless you are planning to have other users encrypt files with your new key pair, you can select “No, thanks!”*

Figure 2: Confirmation of new key pair, GPG Keychain for Mac OS X.

4. Now, it's time to export your public key in order to transfer it. Right click (CTRL + click) on your new key in GPG Keychain, and select "Export" (you can also email the public key to yourself). Make sure the box labeled "include secret key in exported file" is *not checked*, and save the key.
5. Transfer the file with the public key to [the data collection computer](#) however you like.
6. Double click on the key file transferred to [the data collection computer](#). Clicking should automatically open GPG Keychain and import the new public key.
7. Verify that your [stay-home computer](#) has both public and private keys, and that your [data collection computer](#) has only the public key. The leftmost column in GPG Keychain (see figure) shows the "type."

Figure 3: Verifying key pair type on the [stay home computer](#). The top line shows that both secret/private and public keys for "Data Security Test" fingerprint 4534... are stored on this machine.¹⁵

Figure 4: Verifying key pair type on the [data collection computer](#). The top line shows that only the public key for “Data Security Test” fingerprint 4534... is stored on this machine.

8. Nice work! You could theoretically use this key pair for all your PGP needs, but it is probably more cautious to create a separate keypair if you plan to use PGP in emails, etc.

Now that the lockbox is set up, how do you use it? Once again, the objective is that it functions like a timed safe at a convenience store: once you deposit something into it, getting it back is not possible at a moment’s notice, no matter how much you may want it. Data encrypted with your new public key will become accessible when you have access to the private key, stored *only* on the [stay-home computer](#).

Using the Lockbox:

0. Before you leave your home institution to collect data, make sure your [stay-home computer](#) is password protected. Put it in your desk, lock it if you can. Be sure the private key is only in GPG Keychain, and not in some directory that you can access remotely. Turn off remote access/ssh.
1. On the [data collection computer](#), collect your data. At regular intervals during data collection, encrypt your data and destroy the unencrypted copies:
 - (a) In Finder, navigate to the file you wish to encrypt, for example a photo of Chance the Dog.
 - (b) Right click (Ctrl + click) on the file, navigate to “services” and then select “OpenPGP: Encrypt File.”

Figure 5: Menus that appear after right-clicking a file in Finder.

- (c) A GPG Tools window appears, allowing you to select a key with which to encrypt the file. Use the key created above, and add a file-specific passphrase that is different from the key-specific passphrase created above.

Figure 6: Selecting a key for encryption with a passphrase.

- (d) Follow the prompts from GPG Tools.
- (e) Now you have an encrypted file! Verify the file name now ends in .gpg. *Delete the unencrypted file and empty the trash.*¹⁶

¹⁶As many people know, deleted files are often still recoverable. Unfortunately, the solid state hard drives (SSDs) in many new computers make it harder to “overwrite” deleted files than old HDDs did. On a Mac, you can and should still overwrite deleted files when feasible. Open terminal and enter the following prompt to “overwrite” free space on your internal SSD, but the process is slow! `diskutil secureErase freespace 4 /Volumes/Macintosh\ HD`. The numeral 4 is the option for 3-pass secure erasing with the U.S. Department of Energy algorithm. Other options include: US Department of Defense algo. 7-pass erasing (2), Gutmann algorithm 35-pass secure erase (3), overwriting with zeros (0), or a single-pass random overwrite (1).

2. Once files are encrypted, treat them as you would normally. Backing them up is not a bad idea, so long as none of them are stored where the private key is.
3. Upon return to your home institution (or when you need to analyze the data), transfer the encrypted files to your **stay-home computer** and reverse the encryption process.
 - (a) Navigate to the encrypted file on your **stay-home computer**, right click, and select “Services » OpenPGP: Decrypt File”
 - (b) A window will prompt you for a password—enter the passphrase you set earlier.

Figure 7: Prompts for decryption.

- (c) Success! A decrypted copy of your file should have appeared in the same directory! Open it up and go to work

The PGP Lockbox keeps everyone’s hands off your data, including yours. This means the system only works if you can wait to analyze your data until you have returned to your home institution. Keeping the private key on your **data collection computer** to decrypt and encrypt the data at your convenience offers only as much protection as password-protecting a file.

*Statistical Disclosure Control Tools: **sdcMicro***

Various tools exist for measuring and implementing k -anonymity and other statistical disclosure control principles in microdata. One highly-developed suite of tools in R is the **sdcMicro**

package (Templ, Meindl and Kowarik, 2020). This application shows how to use `sdcMicro` to measure de-anonymization risk in nominally anonymous data, and how to implement *non-perturbative* changes in microdata to decrease the risk of de-anonymization.¹⁷

I use data from a massive household survey of Indian citizens, the India Human Development Survey II (IHDS-II) (Desai and Vanneman, 2015), to demonstrate how `sdcMicro` can be used to decrease the risk of de-anonymization in sensitive data. IHDS-II surveys over 200,000 individuals in more than 40,000 households across all 35 states and union territories (prior to the creation of Telangana and the dissolution of Jammu and Kashmir), covering standard demographic information, household finance, education, health, and a wide range of other topics. I use a small subset of the data to construct a statistical disclosure “problem”: a range of quasi-identifying variables for which risk must be gauged and disclosure-mitigation steps taken, and a range of sensitive variables for which the values should not be matchable to specific individuals.

Though the SDC literature—which grows out of technical research at statistical agencies and the International Household Survey Network—primarily focuses on quasi-identifying variables that are part of public record, I use a broader set of quasi-identifiers that are relevant to protecting research participants from de-anonymization by knowledgeable local partners. When local knowledge is at play, it is worth including variables like religion, caste, marital status, etc. for which a research partner with substantial local knowledge would know the values *without consulting public record*.¹⁸

For this example, I use the quasi identifiers of age, state, district, village name, marital status, and caste or religion, along with the potentially sensitive information of how much income a respondent receives annually from government schemes and benefits. IHDS-II wisely

¹⁷Perturbative methods like value swapping, post-Randomization, and simple additive noise are also implemented in `sdcMicro`, but it seems they have been overtaken in popularity by more sophisticated differential privacy algorithms with more elegant statistical properties.

¹⁸In many parts of India, it is actually conceivable that religion and caste are part of public record (with some uncertainty) given naming conventions. Many observant Sikh women, for example, take the name Kaur as either a middle name, or in place of their family name. The male equivalent Singh is a weaker signal of religious identity.

replaces village names with a unique code in their data to prevent re-identification. For the purpose of this example, I treat the codes as identifiable, even though they are not.

Ingredients:

1. Dataset with quasi-identifiers
2. A computer running R 2.10 or newer
3. An installation of the `sdcMicro` package from CRAN—this demo uses version 5.5.1¹⁹

Using `sdcMicro`:

1. Set up your “SDC Problem” by creating an SDC Object:
 - (a) Load necessary packages and import data as a `data.frame` object

```
library(readr); library(sdcMicro)
ihds <- read_tsv('/your/file/path.tsv')
```

- (b) Create an SDC object using your data. SDC objects take a number of arguments. See comments in the code chunk below for a brief description of each

```
sdc <- createSdcObj(# Your microdata, as a data.frame object
  dat = ihds,
  # Column names: categorical quasi-identifiers
  keyVars = c("district", "male", "mar_stat",
              "rel_caste", "state", "vill_code",
              "age"),
  # Column names: numeric quasi-identifiers
  numVars = NULL,
  # Cluster ID
  hhId = IDPSU,
  # Vector of sample weights
  weightVar = WT,
  seed= 02139)
```

- (c) Print the SDC object for an initial read-out of the unicity of records in the dataset. Pay special attention to two features: the proportion of records that violate k -anonymity for $k \in \{2, 3, 4\}$, and the size of the smallest categories for your key variables.

¹⁹Use this code in a `.R` script to install and load. `if (!require('sdcMicro')) install.packages('sdcMicro'); library('sdcMicro').`


```

print(sdc)
The input dataset consists of 204376 rows and 13 variables.
--> Categorical key variables: district, male, mar_stat, rel_caste,
state, vill_code, age
--> Weight variable: WT
--> Cluster/Household-Id variable: IDPSU
-----

Information on categorical key variables:

Reported is the number, mean size and size of the smallest category
>0 for recoded variables.
In parenthesis, the same statistics are shown for the
unmodified data.
Note: NA (missings) are counted as seperate categories!

Key Variable          Number of categories          Mean size
district              372(372)                    549.398(549.398)
male                  2(2)                        102188.000(102188.000)
mar_stat              6(6)                        34062.667(34062.667)
rel_caste             7(7)                        29196.571(29196.571)
state                 33(33)                       6193.212(6193.212)
vill_code             39(39)                       5240.410(5240.410)
age                   100(100)                     2043.760(2043.760)
Size of smallest (>0)
      29      (29)
    101964 (101964)
      341      (341)
     5388      (5388)
      272      (272)
       58      (58)
        5      (5)
-----

Infos on 2/3-Anonymity:

Number of observations violating
- 2-anonymity: 147648 (72.243%)
- 3-anonymity: 187432 (91.709%)
- 5-anonymity: 202378 (99.022%)
-----

```

- Now, begin modifying the data to reduce identifiability. Start by recoding variables that have a large number of small “bins,” like age, to be less granular. The function `globalRecode`, applied to your SDC object, will recode specified variables to be less

granular.²⁰ Simply specify the SDC object, the column you want to recode, and what you want the new categories to be. Then print the SDC object to evaluate the effect of recoding on k -anonymity. When we recode “age” from specific ages to decade bins, the number of observations that are unique across our large number of quasi-identifiers drops from 72% of the dataset to 21% of the dataset. More gains are possible from this single operation by creating even wider bins for age, but wider bins are less useful for analysis. Consider also using the functions `topBottomCoding()` and `groupAndRename()` to provide similar functions for numerical and categorical variables, respectively.

```
sdc <- globalRecode(sdc, column = "age",
                    breaks = seq(from=min(sdc@manipKeyVars$age),
                                to=max(sdc@manipKeyVars$age), length.out = 10))
print(sdc)
```

Infos on 2/3-Anonymity:

Number of observations violating

- 2-anonymity: 44088 (21.572%) | in original data: 147648 (72.243%)
- 3-anonymity: 77362 (37.853%) | in original data: 187432 (91.709%)
- 5-anonymity: 123333 (60.346%) | in original data: 202378 (99.022%)

3. Once satisfied with recoding, try value suppression. The function `localSuppression()` implements an algorithm to prune the dataset into k -anonymity (where k is an argument supplied by the user) by suppressing *individual values* of quasi-identifier variables. The algorithm used by the package suppresses quasi-identifier values for particular observations that have the highest risk of de-anonymization in the existing format of the data. Users can (and should) use the “importance” argument in the function, in order to constrain the algorithm’s choice about which variables to suppress in a given observation. Variables ranked as most important are used as last-resort suppression. Users should also note that `localSuppression()` runs slowly, especially for large datasets and datasets that have a high number of key variables. It continues pruning until k anonymity is achieved for 100% of observations. Note that in order to achieve 3-anonymity across 7 key variables (an unusually high number), 89,865 values are suppressed—roughly 44 cells for every 100 observations in the dataset. When suppression functions this aggressively, users should consider deleting certain quasi-identifier variables entirely, or using perturbative techniques like post-randomization or one of the variety of available differential privacy algorithms. Note, also, that the variables specified as high-importance in the function are suppressed very sparingly. Specifying theoretically important variables as “high importance” during local suppression minimizes the rate at which observations in SDC-treated data will drop out of key regressions due to missingness.

²⁰Counterintuitively, the function `microaggregation()` does something else


```
sdc <- localSuppression(sdc, k=3, importance = c(6,1, 2, 3, 7, 5, 4))
# which vars (rank in order of sdc@keyvars) should be maintained?
# Variables with higher "rank" (1-n) will be last for suppression
```

```
sdc          # To confirm k-anon and see what was suppressed
```

The input dataset consists of 204376 rows and 13 variables.

```
--> Categorical key variables: district, male,
mar_stat, rel_caste, state, vill_code, age
--> Weight variable: WT
--> Cluster/Household-Id variable: IDPSU
```

Information on categorical key variables:

Reported is the number, mean size and size of the smallest category >0 for recoded variables.

In parenthesis, the same statistics are shown for the unmodified data.

Note: **NA** (missings) are counted as separate categories!

Key Variable	Number of categories	Mean size
district	373 (372)	496.634 (549.398)
male	3 (2)	102177.000 (102188.000)
mar_stat	7 (6)	33936.833 (34062.667)
rel_caste	8 (7)	29154.143 (29196.571)
state	34 (33)	4254.879 (6193.212)
vill_code	40 (39)	5137.103 (5240.410)
age	10 (100)	22251.556 (2043.760)
Size of smallest (>0)		
	20 (29)	
	101951 (101964)	
	308 (341)	
	5345 (5388)	
	152 (272)	
	12 (58)	
	351 (5)	

Infos on 2/3-Anonymity:

Number of observations violating

```
- 2-anonymity: 0 (0.000%) | in original data: 147648 (72.243%)
- 3-anonymity: 0 (0.000%) | in original data: 187432 (91.709%)
- 5-anonymity: 42928 (21.004%) | in original data: 202378 (99.022%)
```

Local suppression:

KeyVar	Suppressions (#)	Suppressions (%)
district	19628	9.604
male	22	0.011
mar_stat	755	0.369
rel_caste	297	0.145
state	63965	31.298
vill_code	4029	1.971
age	1169	0.572

4. After recoding and suppressing, users should re-measure disclosure risk before exporting modified datasets. `sdcmicro` provides various metrics for disclosure risk, nicely summarized in a print function. There does not seem to be a universally accepted threshold for how much risk is tolerable, but researchers should decide on thresholds they feel they can defend. Risk measures, plus a full summary of changes can also be output as a report.

```
measure_risk(sdc) # This runs slowly
  report(sdc, internal = T, verbose = T) # generates HTML report
print(sdc, "risk")
```

Risk measures:

Number of observations with higher risk than the main part of the data:

in modified data: 0

in original data: 0

Expected number of re-identifications:

in modified data: 9.99 (0.00 %)

in original data: 357.17 (0.17 %)

Information on hierarchical risk:

Expected number of re-identifications:

in modified data: 1023.10 (0.50 %)

in original data: 31574.67 (15.45 %)

- Users should also consider measuring l -diversity, a measure of disclosure risk related to k -anonymity. l -diversity measures, for a group of k observations that have identical values across a set of quasi-identifiers, the number l of well-represented values for some sensitive attribute. A dataset is l diverse if every group of k observations is represented by l different values for a sensitive attribute. In practical terms, if a 3-anonymous dataset is only 1-diverse for some sensitive attribute, an adversary looking for a person known to be represented in the dataset and having known quasi-identifiers might be able to learn sensitive information about the person simply because all people who share a set of quasi-identifiers also share a value for sensitive information. Within reason, higher l -diversity is better for privacy. Given the unusually high number of quasi-identifiers in this example, achieving high l -diversity would require very drastic modifications to the data.

```
print(ldiversity(d_sub_new, # New dataset
  keyVars = c("male", "age", "rel_caste",
    "district", "state", "mar_stat"),
  ldiv_index = "ben_income")) # Sensitive Variable
```

L-Diversity Measures

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.000	1.000	2.000	3.882	5.000	37.000

- Once satisfied (perhaps after multiple iterations through the above steps) users can export their data, though the process is made slightly cumbersome by `sdcmicro`. All variables other than the quasi-identifier/key variables must be re-assembled separately from the original dataset as they are unchanged during the process.

```
ihds_new <- cbind.data.frame(ben_income = ihds$ben_income,
  disab_pension = ihds$disab_pension,
  hhid = ihds$hhid,
  resp = ihds$resp,
  WT = ihds$WT,
  sdc@manipKeyVars)
```

- The “information” costs of the SDC modifications are measured in terms of distance between the old and new values of continuous values (and differences in eigenvalues) the report described in step 4, but users looking to measure the “information” costs in more practical terms or for mostly categorical variables should consider comparing the performance of pre- and post-modification data in substantively meaningful regressions. Unlike other privacy tools that rely on simulation or noise to obscure

sensitive information, SDC tools *will* change the central tendencies and dispersion of key variables. The relevant question, then, is whether the change is tolerable for the purposes of the research. Table 1 shows the difference in regression coefficients for the same OLS model fit to pre- and post-modification data. The differences, depending on perspective, are substantial, and tens of thousands of observations are dropped because NAs have been induced in the course of local suppression. Whether these differences are acceptable, either for primary analysis or for sharing data, is up to the researcher.

Table 1: Comparison between regressions on pre-modification and post-modification data

	<i>Dependent variable:</i>	
	Pre-modification	Post-modification
	ben_income	
disab_pension	5,073.788*** (237.984)	4,872.312*** (327.661)
rel_caste - Forward caste	232.416*** (67.173)	408.418*** (105.006)
rel_caste- OBC	258.449*** (63.640)	381.109*** (100.315)
rel_caste - Dalit	467.991*** (65.320)	617.885*** (101.783)
rel_caste - Adivasi	347.042*** (79.689)	518.968*** (119.222)
rel_caste - Muslim	109.470 (71.851)	233.291** (108.347)
rel_caste - Christian, Sikh, Jain	252.876** (106.496)	594.696*** (163.238)
mar_stat - Married	−354.006*** (96.193)	−315.579* (187.374)
mar_stat - Unmarried	−21.081 (99.345)	−275.114 (195.852)
mar_stat - Widowed	127.066 (111.180)	−25.319 (238.022)
mar_stat - Separated/Divorced	−207.232 (208.632)	−95.489 (607.482)
mar_stat -Married no gauna	476.453 (325.016)	−942.231 (576.728)
age (numeric)	11.078*** (1.083)	
age(11,22]		196.695*** (42.749)
age(22,33]		52.223 (77.683)
age(33,44]		80.673 (91.009)
age(44,55]		−45.490 (95.144)
age(55,66]		299.156*** (113.114)
age(66,77]		1,127.492*** (171.932)
age(77,88]		1,095.003*** (352.118)
age(88,99]		461.485 (960.741)
male	−95.988*** (26.038)	−38.081 (31.644)
Observations	204,376	127,955
R ²	0.050	0.049
Adjusted R ²	0.048	0.046
Residual Std. Error	5,702.491 (df = 203990)	5,508.105 (df = 127562)
F Statistic	27.772*** (df = 385; 203990)	16.787*** (df = 392; 127562)

Note:

*p<0.1; **p<0.05; ***p<0.01

Unlike `sdcMicro` and the PGP lockbox, this final demonstration focuses on a tool for privacy-preserving presentation of text data—especially text data in small corpora that are primarily collected for qualitative analysis. The benefit of STM for privacy preservation is that the main data format that must be shared in order to reproduce analyses, the Document-Term Matrix, naturally makes de-anonymization difficult in its standard pre-processing steps. Table 2 shows the DTM realization of a document analyzed using STM for privacy preservation in Milliff (2020).²¹ Though a motivated reader could learn *something* about the themes discussed in the document by reading the DTM alone, it would be extremely difficult (likely not possible with any degree of certainty) to reconstruct the document to the extent that contextual knowledge could be used to re-identify the respondent. Turning a DTM back into a document would require an adversary to: 1) reverse the process of stemming—turning stems back into words with proper conjugation and declension; 2) re-arrange the words into the order they appeared in the document and re-insert meaning-critical punctuation (especially full stops); and 3) re-conjure the missing stop words like articles, personal pronouns, direct object and indirect object pronouns, etc.

Another of the major benefits to STM as a deanonymization-prevention tool is user-friendliness. The optimization algorithm that fits structural topic models is complex, but using the `stm` package in R is straightforward, especially with thorough instructions in the package vignette by Roberts, Stewart and Tingley (2018).

Using STM to prevent de-anonymization follows largely the same steps as normal use for digesting large, public corpora. The important modifications come in pre-processing and presentation of the model findings.

First, researchers fitting topic models to sensitive data should do an additional set of pre-processing in order to eliminate personal identifiers before using STM’s built-in tools

²¹This section should probably be re-done with a corpus that *can* be shared, so that the tool can be validated.

to stem the text, remove stopwords, and create a DTM. The process of creating a DTM is likely to do a fairly good job of removing identifiers in its standard function. Identifiers, by definition, occur in one or very few records, so STM pre-processing may automatically drop them as sparse terms. Because identifiers are particularly risky, though, additional steps should be taken to ensure they are cut out of the data. Two possibilities exist: larger corpora could be stripped of identifiers using a Named Entity Recognition (NER) model like the pre-trained models in the python library `spaCy`. The NER model uses statistical (as opposed to rules-based) entity recognition to identify spans of text indicating people's names, particular locations, etc. A researcher could use the pre-trained tool to find and delete information like names and locations that is unique enough to aid de-anonymization and too unique to provide much value in the topic model fitting. NER models are likely to remove identifying information, but not certain. Instead of NER models, researchers could also use brute force: for corpora that are small enough to read, researchers could go through and manually delete identifiers like addresses, cross streets, names of people and locations, in order to ensure they do not end up in the model fit. This process is more labor intensive, but provides better assurances.

Second, for STM specifically because it allows users to estimate topic contents and prevalence as a function of document-level covariates, researchers must take steps—perhaps including the statistical disclosure control tools shown above—to ensure that the prevalence and content covariates they include (and which would be necessary to reproduce the model) are not easy to de-anonymize. The same cautions about disclosure control apply to document-level covariates which are used after model fitting to estimate the association between topic prevalence and respondent characteristics.

Third, researchers should be aware of the importance of un-processed documents in interpreting STM and other topic models. The topics that are generated by a topic model are not guaranteed to be substantively meaningful, and they require substantial interpretation by the user to figure out what, if anything, they mean. One accepted way to label the topics

is reading the documents that have the highest proportions for each topic, and then deciding what thematically links those documents ([Grimmer and Stewart, 2013](#)). Verifying the interpretation of the model, therefore, is easiest if some documents are shared. Researchers have two choices for dealing with this. First, they might take their chances with refusing to share full documents given privacy concerns. It is uncommon to share interview notes for qualitative interviews as part of “replication files,” so researchers might be able to avoid sharing STM documents as well. Second, researchers can split the longer interviews into shorter documents (even paragraph length works) and preserve the order and respondent information by specifying them as prevalence/content covariates in the STM. Under this system, the documents that might be shared to verify model interpretation would be sufficiently short to lessen the risk of de-anonymization. Of course, the most transparent path still poses some de-anonymization risk, and is potentially a weak point in the attempt to use STM for privacy preservation.

The remainder of this demo shows the topic model fit from [Milliff \(2020\)](#), which uses STM to present trends in the contents of interviews about emotional and political responses to violent trauma. The sensitive data used in the topic models are the transcripts of 31 in-depth interviews (semi-structured) conducted in January 2018 with the surviving relatives of homicide victims who were killed between 2015 and 2017 in Chicago, IL. In the interviews, which lasted between 90 and 180 minutes, respondents share their experiences of trauma, their interactions with the state, and their thoughts on the causes of violence with surprising candor. Respondents were recruited with help from a non-academic partner: a social service organization that provides free case management and services to families of homicide victims.

A tool like STM is useful for sharing the results gleaned from these interviews because the views and experiences shared in the interviews are potentially sensitive—perhaps the most sensitive are assignments of blame for the death of a family member—and because the narrative format of the interviews would make re-identification possible even if identifiers like name, place, dates, etc. were deleted. Staff from the partner organization would be able

to easily re-identify respondents given full interview transcripts. Some respondents would be identifiable by a broader audience as well: a number of the homicides discussed in the interviews were covered in local press or memorialized in music.

The goal of this topic model is to show, in a transparent and reproducible way, how the author reached conclusions about the correlates of anger at the perpetrator of homicide vs. anger at other targets based on primarily qualitative analysis of the interviews. An STM fit at the paragraph level with ten topics shows that discussion of anger (topic 5) is positively correlated with conversations about the motive behind the homicide (topic 3) and that when respondents are talking about confusion with respect to what happened (topic 6) they are not using words from the anger topic.

The same model can also be used to estimate associations between respondent-level meta-data and topic prevalence. Since respondent transcripts are broken into many paragraphs, these estimates group documents by respondent. This presentation supports qualitative analysis about *who* and *what circumstances* were most likely to be associated with high levels of anger directed at the perpetrator.

STM results in this application are not a stand-alone presentation of the rich interview evidence in this application. In [Milliff \(2020\)](#), STM results support traditional qualitative interpretation of evidence and single case vignettes—themselves carefully written to avoid including information that could be cross-referenced against public sources—by showing that key patterns obtain across the whole sample, and are not cherry picked from particularly evocative interviews or dramatic stories. The paper further negotiates between privacy protection and transparency by including the “top document” paragraphs for each topic. The author read the 25 top documents in order to label each topic—three of the top 25 are included in an appendix of the paper.

Topic Correlation

Figure 8: Inter-topic correlation for topics where $|r| > 0.1$ with Topic 5 (anger, blame).

Figure 9: Bivariate associations between respondent characteristics and topic proportions

Stem	Count
anywher	1
carri	4
church	4
day	2
doesnt	1
even	1
everi	1
everywher	1
garbag	1
gun	1
happen	1
kill	1
laundromat	1
littl	1
mean	1
much	1
news	1
nothing	1
one	1
period	1
realli	1
see	1
shot	1
sometim	1
start	3
street	1
take	1
time	1
took	1
wife	1
without	1
work	1
wouldnt	1

Table 2: DTM vector corresponding to a single document in the corpus.