QMMR Qualitative & Multi-Method Research

Original Article

Qualitative and Multi-Method Research Fall 2021 - Spring 2022, Vol. 19.2 / 20.1 https://doi.org/10.5281/zenodo.6448133

Data Security in Human Subjects Research: New Tools for Qualitative and Mixed-Methods Scholars [1*]

Aidan Milliff [2†]
Massachusetts Institute of Technology

Political science research in both qualitative and quantitative traditions frequently uses data that contain personal information about research participants. Personal information can enter the research process in different ways; sometimes researchers collect it directly via a survey or an interview, other times they gather it from an aggregator like a government agency or private company or semi-public sources like social media. In many cases, the personal data that political scientists collect is both *personally-identifiable*[3] and *sensitive*, meaning that disclosure could expose respondents to severe repercussions like legal sanction (McMurtrie 2014) or retribution from non-state actors (Venkatesh 2008), as well as more diffuse harms like the negative impacts on personal life, employment opportunities, or reputation (Ohm 2010).

Scholars who use sensitive and personally-identifiable information (PII) in their research may struggle to balance two objectives which are in tension with one another: to keep sensitive data confidential to protect the privacy of human subjects,[4] but also conduct research that meets the method-specific standards of transparency as expected by the political science profession. Researchers often promise interviewees, study participants, or ethnography subjects that the information they share will be confidential unless they explicitly consent to being identified.[5] At the same time, professional bodies like the Qualitative Transparency Deliberations (Jacobs et al. 2021) and the APSA Ad Hoc Committee on Human Subjects Research (2020) call for researchers to provide at least parts of the underlying evidentiary record while still respecting privacy and maintaining confidentiality of sensitive, identifiable information. Some researchers may therefore perceive professional incentives to a) share data as much as possible, and b) maintain copies of *all data* indefinitely.[6]

While there is increasing clarity about the normative *standards* for privacy protection and qualitative transparency that political scientists should seek to uphold, the process of meeting those standards in practice remains largely *ad hoc*, and up to the discretion of individual researchers. To maintain data security in practice (i.e., to protect sensitive,

1 * Thank you to Andy Halterman, Minh Trinh, Lily Tsai, Lukas Wolters, and the referees and editors at QMMR for feedback. Thank you also to Molly Roberts, Jesse Driscoll, and participants in the "Multi-Method Tools for Data Security in Political Science Research" panel at APSA 2020.

2 † Ph.D. Candidate, Massachusetts Institute of Technology.

3 Personally identifiable here means that the data contain sufficient information to reasonably infer the identity of the individual who the data represents, directly or indirectly (McCallister, Grance, and Scarfone 2010).

4 This essay follows the common rule definitions of privacy and confidentiality, in which privacy refers to a research participant's desire (and right) to control what other people know about him or her, and confidentiality refers to the way that researchers (promise to) handle participants' data, typically focused on protecting their privacy.

5 This promise is frequently part of the consent forms required by Institutional Review Board (IRB) processes (Fujii 2012; Zechmeister 2015), and is probably only omitted in specific circumstances like elite interviews. Even when using pre-existing data that contains PII (King and Persily 2019), there is a growing consensus that researchers are obligated to guard "public" data as if they had secured informed consent and collected it themselves (Gibney 2017; Shilton and Sayles 2016).

6 The new APSA guidelines suggest that political scientists facing pressure to prioritize transparency in a way that harms research participants should contact the APSA Committee on Professional Ethics, Rights, and Freedoms.

identifiable data from misuse, disclosure, or reverse-engineering) researchers need to address a range of threats that accrue when sensitive, personally- identifiable data are collected and stored, and when de-identified data are shared. Although threats to data security (and viable solutions) vary widely depending on the research context and methods used, this article attempts to provide practical advice for designing data security protocols that meet reasonable standards for privacy protection and qualitative transparency.

I focus primarily on one common threat to data security and respondent privacy—the re-identification of participants—that can occur in both qualitative and quantitative human subjects research and is a threat across the lifespan of a research project. Re-identification can occur when adversaries are able to reverse-engineer the identity of research participants from sources that have nominally been de-identified or stripped of personal information. In the second section, I describe how the threat of re-identification arises in political science research and I describe general characteristics of good practical solutions to manage re-identification threats while respecting the importance of qualitative transparency. In the third section, I introduce a complication that is also widespread in political science research: re-identification threats increase and become harder to manage for research projects that involve partners like civil society organizations, community groups, research assistants, or translators. Finally, in the fourth section, I turn to solutions.

I propose some practical tools for managing the threat of re-identification in qualitative and multi-method data, including two novel practices that rely on open-source, easy to use tools. I conclude by situating these tools in the broader, evolving landscape of threats to data security in political science research.

Re-Identification and other Threats to Data Security

Social scientists who collect and analyze sensitive data face a wide range of threats to the confidentiality of participant data. These threats are important to consider at all stages of a research project; according to recently revised ethics guidelines from APSA, ensuring participant privacy and safety is the obligation of each individual researcher (APSA Ad Hoc Committee on Human Subjects Research, 2020). In this section, I briefly describe three of the many possible threats to data security: theft, expropriation, and re-identification. I then focus more specifically on re-identification for two reasons. First, re-identification is a threat that can be especially sensitive to the way researchers try to balance data security and transparency goals. Second, strategies to guard against re-identification are likely more generalizable than strategies to guard against theft and expropriation, which depend heavily on research context and legal jurisdiction.

One of the threats to data security is the possibility that data might be stolen. Theft can occur at any point between when data are collected and destroyed. Why should political scientists worry about theft? Theft of personal data from academic institutions is already common, but so far has targeted student records, not research data (see e.g., Identity Theft Resource Center 2017). Research data may become a target in the future, as social scientists use (and store) larger and more sensitive administrative data sets. The threat of theft might also increase in collaborative projects, where co-authors store PII on a network or frequently send it back and forth (Summers 2016).

Another threat to data security arises if researchers are forced, by law or otherwise, to give up data they have collected. This possibility, expropriation, threatens any data that researchers possess. Actors with bad intentions might also try to get data through coercion. Researchers are sometimes monitored by security services while collecting sensitive data (Wood 2009) or in rare instances, closely followed or questioned (Menoret 2014). United States citizens conducting research abroad might be able to leave without risk of extradition, but leaving generally protects a researcher's physical integrity, not the data they have collected.[7] Legal threats to data security are often overlooked, but researchers in the United States, for example, can be obliged to comply when American courts demand sensitive, identifiable 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 held in contempt of court and jailed (Scarce 2005). Bringing data across international borders is hardly an ironclad solution. In 2011, tapes from an oral history of the Irish Republican Army held by researchers at Boston College were subpoenaed under a provision in a mutual legal assistance treaty between the US and the United Kingdom; these tapes were then used to implicate the research participants in a murder investigation (McMurtrie 2014; Radden Keefe 2018).

A third threat to data security—the re-identification or reverse-engineering of personal information from nominally anonymous data—is more amorphous than the

7 Leaving also does too little to protect local colleagues.

first two.[8] Re-identification is a risk that varies depending on data sharing practices. Linking data to respondents can be surprisingly easy in both qualitative and quantitative data, even if PII are removed before sharing. Though the examples below describe re-identification in quantitative data, the same logic applies to descriptions of interview subjects or ethnographic interlocutors: providing context can sometimes positively identify an individual.

Re-identification can occur when unique combinations of attributes are matched to publicly available references, or when contextual knowledge allows an adversary to recognize an individual in the data. Sparse data structures are less anonymous than researchers expect. As of 2000, 87% of US residents are uniquely identifiable by three attributes which would be easy to match with public records: ZIP code, gender, and birth date (Sweeney 2000).

Re-identification doesn't just rely on demographic variables. In a study of Netflix user data, computer scientists found that small amounts of "background knowledge" about a respondent's movie tastes was sufficient to identify their anonymized account (Narayanan and Shmatikov 2008, 2). IMDB accounts (social media accounts) with as few as 5-10 movie ratings could be reliably linked to Netflix accounts because aside from a few popular movies, a watch-list is a surprisingly individual trait (Narayanan and Shmatikov 2008). Adversaries can also use broad contextual knowledge to identify anonymous respondents. Academic publications often try to describe the research setting without identifying it.[9] While important for assessing generalizability of results, these details can also be used to identify the data collection setting, increasing the risk of de-anonymization. Knowing the data collection setting aids de-anonymization. Unique records with respect to age or occupation become more identifiable if the data are known to come from a particular city, school, or company.

Re-identification is the most nuanced threat to data security because it often depends on the extent to which researchers share their data, either in publications, as replication material, or even with their research partners. Some of the techniques commonly used to protect respondent privacywhen sharing these data are not always adequate protection against motivated adversaries.

Data Security with Research Partners

Researchers often work with partners and collaborators—people who are not themselves academic researchers but aid in collection of data either for employment or for mutual interest/benefit. Though some researchers work "solo" or collaborate only with other academics, a substantial number of scholars work with partners, especially to do field research (Kapizewski, MacLean, and Read 2015). Working with partners including NGOs, governments, companies, research assistants, translators, and enumerators or guides change the presentation of all three data security threats.

Theft may be easier if partners' computing and data storage systems are more vulnerable than university systems. Even many highly capable partner organizations (never mind individuals) may have poor digital hygiene/ information security practices, making data that passes through their net- work more vulnerable to theft. Negotiating changes to information security practices or avoiding poorly secured networks all together, may be a difficult addendum to research agreements.

Partners may increase a project's vulnerability to expropriation if they need to maintain good relationships with governments where they work. Unlike researchers who may enjoy the freedom to "go home" from a research site, research partners could be subject to coercive pressure from government or, for organizations, their own funders. This exposure puts any data held by the partner at risk and may leave researchers with little leverage to fulfill their data security obligations.

Perhaps most importantly, partners are likely to be experts in the research context and thus particularly well-suited to identify individuals represented in the data that researchers collect.[10] This can complicate efforts to keep data anonymous. NGOs, governments, companies, and individuals are often valuable research partners *because* of their contextual knowledge, but the more they know about the context and the population being studied, the more points of external leverage they must re-identify individuals in de-identified records, quotations, or notes. When respondents share sensitive information with researchers, they may not want that information shared with a partner organization or members of the project team who reside locally. One common academic partnership arrangement, for example, is program evaluation (qualitative or quantitative) for a partner that serves the population that a researcher aims to study. If partners re-

8 Re-identification technically refers to discovering respondent identity in data from which PII has been stripped. De-anonymization refers to inferring respondent identity even though the data never contained PII. I treat them together because, as I describe below, various examples have shown that people can be identified from data that are thought to be *anonymous*, not just de-identified.

9 See, for example, the Facebook data from Lewis et al. (2008), which is no longer available because it was partially de-anonymized (Zimmer, 2008).

10 I assume here that sensitive information needs to be protected against improper use by the partner, as well as by third parties.

identify data including negative attitudes or experiences related to the services, the consequences could be bad for respondents if local partners have leverage to retaliate against them. If, for example, a respondent admits to criminal activity and their response is re-identified by the research partner, the information could be used to deny the respondent benefits. In a real example from qualitative sociology research, disclosing data on informal economic activity to a gang "research partner" active in Chicago public housing allowed the gang to extract unpaid "taxes" from the respondents (Venkatesh 2008).

Preventing Re-Identification: Ideas for Improvement

This section introduces tools that might help scholars address the risk of re-identification, and the special risks that come from working with research partners.[11] The tools recommended here are not exhaustive, not necessarily appropriate for all research contexts, not "silver bullet" solutions, nor representative of the cutting edge in security research. Instead, they are meant to be *feasible* for most researchers. Data security practices only work when implemented, so I focus on measures that are inexpensive, non-time-consuming, and technically simple.

Data Minimization as a General Best Practice

The best way to protect respondent privacy is to *not collect sensitive information or the PII necessary to link it to individuals*. Variables like age, race, and location of residence affect many social science outcomes and must be measured. But many researchers, both in quantitative and qualitative research, feel pressure to measure everything possible, whether to respond to hypothetical reviewers or to "make something" from costly-to-collect data even when main hypotheses are unsupported.

A spartan impulse during research design addresses many key data security threats: data that are never recorded cannot be stolen, expropriated, or accidentally released.[12] "Data minimization," or "privacy by design" entails collecting the minimum amount (and minimum granularity) of both sensitive information and potentially identifying information necessary to test hypotheses plus the most likely alternative explanations. Though the specifics of data minimization would vary across projects, the general intuition should be widely applicable. A researcher designing an interview guide might ask themselves, for example: Can I articulate an analysis for which I will need this information? before asking respondents for personally-identifying information like their ZIP code, exact address, or date of birth.[13] For information that is unlikely to be included in the final analysis or write-up (i.e., where the researcher is more likely to list city or neighborhood than home address when quoting an interview subject), I argue that researchers would often do well to shed a "just in case" attitude about collecting additional information.

Data minimization comes with both benefits and costs. The most important benefit, I argue, is the potential to reduce risk to research participants. Even if other steps are taken to reduce the chance of data security failures like theft and expropriation, limiting the collection of sensitive or personally identifying data might mitigate some harm to participants if theft or expropriation were to happen. A second, smaller benefit accrues to the researcher: data that contain less sensitive or identifying information are easier to handle safely and easier to prepare for sharing.

There are several important costs associated with data minimization, though. For one, data minimization reduces a researcher's freedom to conduct exploratory analyses or find things the researcher was not expecting. If minimization makes the utility of a given data collection effort more narrow, one could say it means that researchers are spending participants' time less efficiently, which is not ideal.[14] Second and relatedly, data minimization reduces the re-usability of data. Conducting data collection is time and resource intensive, so many researchers try to use a single set of interviews, a single ethnographic site, or a single survey to produce multiple works. Data minimization might decrease the possibility of serendipitous spin-offs. Third, there might be professional costs to data minimization because having less information limits the researcher's ability to respond to comments or conduct additional analyses. The severity of this downside in practice likely depends on early adoption by more senior researchers, and integration of data minimization into already accepted norms like pre-registration.

With these costs and benefits in mind, when can researchers pursue a data minimization strategy? Three characteristics seem important for it to be feasible.

11 Though the other threats discussed above—theft and expropriation—are also important, the ways to address them are much less generalizable because they vary so much with political and legal context.

12 Un-recorded data can still be inferred by context experts, however.

13 The intuition may be different in the special case of elite interviews, where potentially identifying information like specific job title might be a necessary part of the published analysis. In this special case, I would argue it is important to treat interviews as essentially "on the record," and affirmatively seek participants' consent to reprint identifiable quotes.

14 This effect would hopefully be limited if data minimization decreases the length of participation by cutting questions/topics.

First, to accrue the harm mitigation benefits of data minimization, the data collection project needs to be more-or-less single purpose. If a single set of interviews (or an omnibus survey) seeks to test multiple theories about different phenomena, then "minimizing" with respect to those multiple objectives will not necessarily reduce the collection of sensitive information. Researchers who need to collect a wide range of information from the same participants may need to adopt other strategies for data security. Second, data minimization is probably only feasible for deductive, hypothesis-testing data collection. Adopting a data-minimization mindset for exploratory or inductive fieldwork (likely including a lot of critical and interpretive research) could impinge on a researcher's ability to find things they are not expecting. Third, data minimization will not be useful for projects where sharing identifying information like job title (with permission!) is important for establishing the credibility of the speaker. Minimizing other collection will not pay dividends for scholars conducting "on the record" elite interviews, for instance. Where the limitations of data minimization are tolerable, though, I argue it should be attractive to researchers because of its simplicity and relatively strong guarantees of success.

Preventing Re-Identification

Beyond data minimization, several methods are available to guard against re-identification specifically. Preventing re-identification is typically a priority when data are shared (in a manuscript or other public product), but as I discuss in a subsequent section, researchers can also take steps to prevent partners from re-identifying or misusing sensitive data before public release. I describe two techniques for preventing re-identification here: statistical disclosure control/*k*-anonymity and topic modeling for privacy protection.

Statistical Disclosure Control and *k*-anonymity:

Statistical Disclosure Control (SDC) and *k*-anonymity are concepts that come from the quantitative data security literature, but I argue that their shared, underlying intuition is also extremely useful for scholars analyzing, presenting, or sharing qualitative data. The idea behind *k*-anonymity, as proposed by Samarati and Sweeney (1998), is to modify data such that no value of any identifying attribute in the data is shared by fewer than *k* records (see also Sweeney 2002). If no individual value for "age" appears for fewer than three records, the dataset has 3-anonymity for age. This principal is more commonly implemented with respect to "quasi-identifier tuples," or combinations of attributes that could collectively lead to identification—for example, age-gender-ZIP code. K-anonymity is manufactured by suppressing values of identifiable attributes, or by generalizing values (i.e., converting birth years to birth decades).

K-anonymization has drawbacks. First, adversaries can still learn about individuals they know to exist *somewhere* in a dataset. Adversaries trying to learn the HIV status of "Steve"—male, age 35, ZIP Code 60637, known survey respondent—can look at HIV status for all records that match Steve's quasi-identifier tuple and infer the probability that Steve is HIV positive. Recent improvements at least make this risk easier to measure.[15] Second, *k*-anonymization is hard to implement in high-dimensional data, where the unicity of quasi-identifier tuples is remarkably high (de Montjoye et al. 2013). Finally, *k*-anonymization can change the distributional characteristics of data (Angiuli, Blitzstein, and Waldo 2015). K-anonymity is an attractive solution, though, because it is intuitive, relatively easy to implement, and widely used. A related tool, part of the broader research area around Statistical Disclosure Control (SDC), focuses on aggregation, limiting both the geographic and quantitative resolution at which data are reported. Like *k*-anonymity, aggregation eliminates unique records in data. This increases security at the cost of analytical value or informativeness. Aggregation necessarily obliterates high-leverage observations which may be major drivers of the results of statistical analysis.

How can the intuition behind these tools be applied to qualitative research? The intuition and the actual tools behind *k*-anonymity and statistical disclosure control can be a helpful rubric for deciding how to report the demographic identity of interlocutors in a variety of types of qualitative analysis, especially interviews and ethnography. Using tools demonstrated in the online appendix, scholars can empirically measure the relative identification risk of describing an interview participant as "female, age 45, from XYZ village" against the risk of describing that same participant as "female, in her 40s, from ABC district." Researchers trying to weigh the costs and benefits of providing more specificity in descriptions of the people they quote can simply make a spreadsheet containing the demographics they want to describe and then apply tools to measure and increase *k*-anonymity to find a privacy-preserving but still informative way to identify participants.

Maintaining Anonymity in Text and Other Qualitative Data:

Qualitative researchers often analyze sensitive data

15 For a demonstration, see the online appendix: https://aidanmilliff.com/publication/data-security-agenda-for-improvement/QMMR_Appendix.pdf

that are either naturally represented in text (historical or legal documents, social media data), or can be coerced into text (interviews). Text data are often very easy to re-identify or de-anonymize given basic contextual knowledge. Text data can also be uniquely identifying in its pragmatics (context, implication) even if identifying data have been removed from the semantics (words) and syntax (organization of words). An increasing number of text studies use data that are semi-public (like tweets), or clearly private (like longer transcripts of interviews, which are traditionally analyzed qualitatively (but see Milliff, 2021)). For these applications, researchers need to pay attention to de- anonymization concerns when sharing data in manuscripts or in replication files. One novel method for privacy-protecting analysis of sensitive text, building on the user-friendly Structural Topic Model by Roberts et al. (2013), is demonstrated in the online appendix. Topic models are typically used for comparing documents in corpora of text that are too large to read. This new approach uses topic modeling to compare documents in a corpus that is quite small, but for which presentation of raw, high-dimensional data threatens the privacy of the speakers represented in the text.

Topic modeling helps here because it focuses exclusively on *morphologic* patterns (words and their meanings). The data format that topic models ingest (data that would be shared for replication) is a document-term matrix (DTM): a format which ignores word order, making it difficult to re-assemble the original natural language. For longer documents (such as multiple sentences containing multiple verbs or multiple subjects), re-assembling the original document from a DTM is practically impossible. A document-term matrix, so long as no terms are themselves identifiers, is hard to connect to a particular individual.[16]

Topic modeling, however, is not a silver bullet for portraying patterns in qualitative data. Three downsides are worth noting. First, because topic modeling is an "unsupervised learning" tool, researchers usually cannot pre-specify the topics they would like a model to focus on. There is no ironclad guarantee, in other words, that a topic model will return topic clusters that are relevant to the research question at hand.[17] Second, if raw text data contains identifying terms (i.e., proper names), the topic model will contain them as well. Researchers who want to use topic models for privacy preservation need to ensure before modeling that directly identifying terms are censored or replaced. Third, topic modeling is time intensive. Using this technique for interview data, for example, requires text transcripts that are either time consuming or expensive to make. Cleaning the data to get rid of identifiers is likewise time consuming (or computationally intensive). If researchers can produce clean, non-identifying text from their qualitative data, though, topic models offer an interesting new way to present privacy-preserving summaries of sensitive information.

Mitigating Threats from Partners

As noted above, working with research partners changes the threat of re-identification in both qualitative and quantitative data. As such, I argue that additional techniques to preserve data security might be necessary or useful when a researcher is trying to prevent disclosure or re-identification by partners *before* data are shared publicly. I describe two techniques here, both of which are aimed at "keeping honest partners honest" and erecting modest barriers to the misuse of data after it is collected. Neither is a substitute for up-front work to vet partners and ensure that research collaborators share a strong commitment to treating participants with respect and dignity.

One intuitive way to reduce the risk that partners re-identify respondents in non-public data is to guard against over-sharing. Partners, in many cases, only need access to a specific subject of project information in order to participate in a project. Sharing *necessary* rather than *complete* versions of information like lists of participants, interview notes/tapes/transcripts, or recruitment blasts will limit the ability of partners to use contextual knowledge to re-identify research participants. With some partners, negotiating an agreement that limits sharing of re-identifiable data is not difficult because practitioner partners are primarily interested in finished products, like internal reports created by the researcher, rather than raw data. If social scientists work proactively to identify products that the partner wants, they may be able to avoid sharing sensitive data. When the structure of a partnership requires sharing PII or sensitive data with a partner, sharing via cloud storage is a good way to keep honest partners honest. Cloud storage platforms like Dropbox allow file owners to monitor access and downloads, so that researchers can make sure raw data aren't being misused.

A second way to reduce the risk of re-identification is to practice a "hand tying" strategy when working with partners, simply taking the possibility of data sharing off

16 Mosteller and Wallace (1963) find that it is sometimes possible to identify authors based on the rate at which they use common words. Unless adversaries are searching for a known author in a corpus analyzed using STM and have a substantial amount of "labeled" reference material, this seems like an unlikely vector for the re-identification of interview transcripts.

17 New work by Eshima, Imai, and Sasaki (2020) may mitigate this downside, allowing researchers to specify keywords for topic formation.

the table. This strategy is likely more useful in situations where the partner has some leverage over the researcher. One new, simple technique uses PGP (pretty good privacy) encryption software to set up a "vault" for sensitive in- formation. Supplementary materials in the online appendix provide step-by-step instructions. Once researchers "deposit" information into the PGP vault and delete unencrypted copies, the information is inaccessible until the researcher can access the key. If the key is left in another location and is not internet accessible, the researcher has effectively *tied her hands*: she cannot access the data herself. Other methods, like mailing physical media, could theoretically serve the same purpose without using computer encryption. Hand-tying is fundamentally a short-term solution—the researcher will have to access the private key eventually in order to unlock the data.

These tools, which provide simple ways to manage the risk of re-identification by research partners, also have some downsides. Both tools, for one, are additional work and make collaboration less smooth. The researcher takes on something like a systems administrator role in order to structure and manage data access—this could consume a lot of time. Second, these tools must be applied carefully and tactfully. It could be detrimental to a research partnership if partners felt disrespecte by the systems a researcher put in place to ensure data security. This is especially a risk with hand tying. If a researcher took steps to be unable to comply with a request for data, it would likely jeopardize future work with the requesting partner. Finally, neither of these tools prevent people from knowing what they saw with their own eyes. Research assistants and translators especially will still be able to identify research participants because they will be present at data collection. None of the techniques here can supplant good leadership, communication of clear ethical standards, hiring well, and vetting employees.

Conclusion

This article has proposed new techniques for improving data security in qualitative (and quantitative) political science research. I have argued that the re-identification of individual research participants is a particularly important threat to researchers' ability to fulfill the promises they often make to participants and have identified some simple technical solutions that should help researchers fulfill their promises while still responding to professional imperatives to make qualitative researchtransparent when possible. The article has tried to show that it is eminently possible to reduce the risk of data security failures when gathering and storing sensitive data. Whether or not better practices are ultimately adopted, though, depends on whether social science disciplines incentivize good practices and tolerate the compromises that good security requires.

Ensuring the security of sensitive data is an evolving challenge that researchers will have to revisit regularly throughout their careers. By ignoring data security, researchers are allowing the (admittedly small) likelihood of failure to increase over time. As political scientists adopt new technology for collecting and storing data, new threats to the security of that data will arise as well and may catch researchers unprepared. Contemporary data security practices are not "future proof" in any meaningful sense, so it is important for researchers to update their knowledge and use of relevant data security tools regularly to prevent the pile of un-addressed threats from growing too large. As the likelihood of data security failure appears to increase, the expected consequences of failure are surely growing: The popularity of collecting and analyzing large, identifiable data is in- creasing, which means the ethical and professional consequences of a potential data breach grow as well. Examples from the academy in the last two decades (e.g., Venkatesh 2008; McMurtrie 2014) already hint at the grave consequences that the release of sensitive data can have for research subjects. With these examples in mind, political scientists should not be content to wait for an even larger crisis to prompt the re-examination of data security practices in their own research.

Taking more systematic steps to guard respondent privacy is important, but not without trade- offs and fundamental limitations. Researchers should be mindful of these limitations as they adopt new tools. First, increasing privacy via more robust data security impinges on transparency. Even in the best-case compromise, rigorous data security protocols might make it harder to detect dishonesty in research by limiting the amount of data that a curious reviewer can demand to see. Second, good data security practices are sure to vary widely across the incredible range of methods and contexts in empirical political science. It is up to scholars to weigh the risks and benefits of specific data security techniques before deciding what strategy is most appropriate for their work. Third, using new and more complex data security techniques increases the difficulty researchers face in explaining their security precautions to research participants, who need to be adequately informed about the privacy risks of participating in political science research. Finally, there is a risk that promoting new tools for privacy protection incentivizes riskier behavior to begin with. To end with a warning: none of the technical solutions presented here are as ironclad as simply declining to collect and store sensitive data. Because the data security challenge is fundamentally political and social, technical fixes can help, but are naturally incomplete.

References

Angiuli, Olivia, Joe Blitzstein, and Jim Waldo. 2015. "How to De-Identify Your Data." *Communications of the ACM*, 58, no. 12 (December): 48-55.

APSA. 2020. "Principles and Guidance for Human Subjects Research." Ad Hoc Committee on Human Subjects Research, American Political Science Association. https://www.apsanet.org/Portals/54/diversity%20and%20inclusion%20prgms/Ethics/Final_Principles%20with%20Guidance%20with%20intro.pdf?ver=2020-04-20-211740-153

Eshima, Shusei, Kosuke Imai, and Tomoya Sasaki. 2020. "Keyword Assisted Topic Models." Last revised March 10, 2021. https://arxiv.org/abs/2004.05964.

Fujii, Lee Ann. 2012. "Research Ethics 101: Dilemmas and Responsibilities." *PS: Political Science & Politics* 45, no. 4 (October): 717–23. https://doi.org/10.1017/S1049096512000819

Gibney, Elizabeth. 2017. "Ethics of Internet Research Triggers Scrutiny." *Nature* 550 (7674):16–7. https://doi.org/10.1038/550016a

Identity Theft Resource Center. 2017. *Data Breach Reports: 2016 End of Year Report*. El Cajon, CA: Identity Theft Resource Center. https://www.idtheftcenter.org/wp-content/uploads/images/breach/2016/DataBreachReport_2016.pdf.

Jacobs, Alan M., Tim Büthe, Ana Arjona, Leonardo R. Arriola, Eva Bellin, Andrew Bennett, Lisa Björkman, et al. 2021. "The Qualitative Transparency Deliberations: Insights and Implications." *Perspectives on Politics* 19 (1): 171–208. https://doi:10.1017/S1537592720001164.

Kapiszewski, Diana, Lauren M. MacLean, and Benjamin L. Read. 2015. *Field Research in Political Science: Practices and Principles*. Cambridge: Cambridge University Press.

King, Gary, and Nathaniel Persily. 2019. "A New Model for Industry–Academic Partnerships." *PS: Political Science & Politics* 53, no. 4 (October): 703-09. https://doi.org/10.1017/S1049096519001021

Knerr, Charles R. Jr. 1982. "What To Do Before and After a Subpoena of Data Arrives," In *The Ethics of Social Research: Surveys and Experiments*, edited by Joan E. Sieber, 191–206. New York: Springer.

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, no. 4 (October): 330–42. https://doi.org/10.1016/j.socnet.2008.07.002

McCallister, Erika, Tim Grance, and Karen Scarfone. 2010. "Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)." National Institute of Standards and Technology (NIST) Computer Security Resource Center SP 800-122. https://doi.org/10.1016/j.socnet.2008.07.002

McMurtrie, Beth. 2014. "Secrets from Belfast." *Chronicle of Higher Education*. January 26, 2014. https://www.chronicle.com/article/secrets-from-belfast/.

Menoret, Pascal. 2014. "Repression and Fieldwork," In *Joyriding in Riyadh: Oil, Urbanism, and Road Revolt*, 21-60. New York: Cambridge University Press.

Milliff, Aidan. 2021. "Facts Shape Feelings: Information, Emotions, and the Political Consequences of Violence." *Political Behavior*. https://doi.org/10.1007/s11109-021-09755-1.

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 (1376). https://doi.org/10.1038/srep01376.

Mosteller, Frederick, and David L. Wallace. 1963. "Inference in an Authorship Problem." *Journal of the American Statistical Association* 58, no. 302 (June): 275–309. https://doi.org/10.1080/01621459.1963.10500849

Narayanan, Arvind, and Vitaly Shmatikov. 2008. "Robust De-anonymization of Large Sparse Datasets," In *2008 IEEE Symposium on Security and Privacy*, 111–25. Oakland: IEEE.

Ohm, Paul. 2010. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization." *UCLA Law Review* 57:1701–1778.

Radden Keefe, Patrick. 2018. *Say Nothing: A True Story of Murder and Memory in Northern Ireland*. New York: Penguin Random House.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, and Edoardo M. Airoldi. 2013. "The Structural Topic Model and Applied Social Science." Working paper, prepared for the 2013 NIPS Workshop on Topic Models: Computation, Application, and Evaluation.

Samarati, Pierangela, and Latanya Sweeney. 1998. "Protecting Privacy when Disclosing Information: *k*-Anonymity and Its Enforcement through Generalization and Suppression." *Technical Report SRI-CSL-98-*04 Computer Science Laboratory, SRI International.

Scarce, Rik. 2005. *Contempt of Court: A Scholar's Battle for Free Speech from Behind Bars*. Lanham: Rowman and Littlefield.

Shilton, Katie and Sheridan Sayles. 2016. " 'We aren't all going to be on the same page about ethics:' Ethical practices and challenges in research on digital and social media." In *Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS 2016),* 1909–1918. Kauai, HI: IEEE.

Summers, Scott. 2016. "Organising, Storing and Securely Handling Research Data." PowerPoint presentation, UK Data Service, Essex, England, June 15. https://dam.ukdataservice.ac.uk/media/604451/2016-06-15_storing_data.pdf

Sweeney, Latanya. 2000. "Simple Demographics Often Identify People Uniquely." Working paper, Carnegie Mellon Univeristy Data Privacy Working Paper Series.

Sweeney, Latanya. 2002. "k-Anonymity: A Model for Protecting Privacy." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10 (5): 557–70. https://doi.org/10.1142/S0218488502001648

Traynor, Michael. 1996. "Countering the Excessive Subpoena for Scholarly Research." *Law and Contemporary Problems* 59, no. 3 (Summer): 119–48.

Venkatesh, Sudhir. 2008. *Gang Leader for a Day*. New York: Penguin Press.

Wood, Elisabeth J. 2009. "Field Research," In *The Oxford Handbook of Comparative* Politics, edited by Carles Boix and Susan C. Stokes. Oxford: Oxford University Press, Oxford.

Zechmeister, Elizabeth J. 2015. "Ethics and Research in Political Science: The Responsibilities of the Researcher and the Profession," In *Ethics and Experiments*, edited by Scott Desposato. New York: Routledge, London.

Zimmer, Michael. 2008. "More on the 'Anonymity' of the Facebook Dataset—It's Harvard College." Blog post. October 3, 2008. https://michaelzimmer.org/2008/10/03/more-on-the-anonymity-of-the-facebook-dataset-its-harvard-college/.

QMMR Qualitative & Multi-Method Research

Symposium:

Author-Meets-Critic: James Mahoney, 2021. *The Logic of Social Science.* Princeton, NJ: Princeton University Press.

Qualitative and Multi-Method Research Fall 2021 - Spring 2022, Vol. 19.2 / 20.1 https://doi.org/10.5281/zenodo.6448059

Applying A New Approach to Knowing the Social World

Jennifer Cyr
Universidad Torcuato di Tella

"*[M]ainstream social science methods depend on the assumed truth of essentialism."* (Mahoney 2021, 5)

The Logic of Social Sciences is a tour de force. The book and its author are advocating for revolution—a revolution in the social sciences. I admire the author greatly for writing it.

I am also rather overwhelmed by this book. The need to *un-learn* how we undertake research and think about causality in the social sciences, in order to *learn* it all once more, is daunting. Indeed, the book sets out myriad tasks for us as potential teachers and practitioners of the kind of social sciences it promotes. At times I wondered if the book was more aspirational than applicable.

In this intervention, I consider what we must do to put into action the kind of social science that this book promotes. I consider the central arguments of the text