Data Security in Human Subjects Research:

New Tools for Qualitative and Mixed-Methods Scholars *

Aidan Milliff †

April 10, 2022

Abstract

Many political science studies use personal information from research participants, but practical discussion about safe ways to handle and store personal information has been piecemeal. As a result, actual security practices vary widely from project to project. In this article, I focus on one common threat to data security—re-identification of respondents/interlocutors who are supposed to be de-identified. First, I discuss the nature of the re-identification threat with special attention to the way it manifests in qualitative and mixed-methods research. Second, I discuss how re-identification threats (and potential solutions) change when political scientists work with teams, from partner organizations to local research assistants or translators. Third, I suggest new data security practices to address the threat of re-identification, and I demonstrate how user-friendly tools can help political scientists manage, but never eliminate, the risks associated with collecting, storing, and sharing personal information.

*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.

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

1 Introduction

Political science research in both qualitative and quantitative traditions frequently uses data that contains 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*[1] 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: keep sensitive data confidential to protect the privacy of human subjects,[2] but also conduct research that meets the method-specific standards of "transparency" 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.[3] At the same time, professional bodies like the Qualitative Transparency Deliberations of 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, identi-

[1] Personally identifiable: the data contains sufficient information to "reasonably infer" the identity of the individual who the data represents, directly or indirectll=y (McCallister et al., 2010).

[2] This essay follows the common rule definitions of "privacy" and "confidentiality": privacy refers to a research participant's desire (and right) to control what other people know about them; confidentiality refers to the way researchers (promise to) handle participants' data, typically focused on protecting their privacy.

[3] This promise is frequently part of the consent forms required by Institutional Review Board (IRB) processes (Fujii, 2012; Zechmeister, 2016), 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, 2016).

fiable information. Some researchers may therefore perceive professional incentives to a) share data as much as possible, and b) maintain copies of *all data* indefinitely.[4]

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. protect sensitive, identifiable data from mis-use, 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—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 section 2, 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 section 3, 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. Section 4 turns to solutions. I propose some practical tools for managing the threat of re-identification in qualitative and multi-

[4]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.

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.

2 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, and, 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 (eg. Identity Theft Resource Center, 2016). 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 increases in growingly-common 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, 2007) or in rare instances, closely followed or questioned (Menoret, 2014). U.S. citizens 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.[5]

Legal threats to data security are often overlooked, but researchers in the United States, for example, lack protection to refuse 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 U.S.-U.K. mutual legal assistance treaty and used to implicate research participants in a murder investigation (McMurtrie, 2014; Radden-Keefe, 2018).

A third threat to data security, re-identification or reverse-engineering personal information from nominally anonymous data, is more amorphous than the first two.[6] 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

[5]Leaving also does too little to protect local colleagues.

[6]Re-identification technically refers to discovering respondent identity in data from which PII has been stripped. De-anonymization refers to inferring respondent identity despite the fact that 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.

in the data. Sparse data structures are less anonymous than researchers expect. As of 2000, 87% of U.S. residents are uniquely identifiable by three attributes—ZIP code, gender, and birth date (Sweeney, 2000)—which would be easy to match with public records.

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. 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.[7] 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, occupation, etc. 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 privacy when sharing these data are not always adequate protection against motivated adversaries.

3 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 schol-

[7] See, for example the Facebook data from Lewis et al. (2008), now un-avialble because it was partially de-anonymized (Zimmer, 2008).

ars work with partners, especially to do "field" research (Kapiszewski et al., 2015). Working with partners including NGOs, governments, companies, research assistants, translators, and enumerators/guides changes 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 network 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.[8] This can 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 have to 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 locally-resident members of the project team. One common academic partnership arrangement, for example, is program evaluation (qualitative or quantitative) for a partner that serves the population

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

that a researcher aims to study. If partners re-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).

4 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.[9] 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.

4.1 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

[9] 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.

are never recorded cannot be stolen, expropriated, or accidentally released.[10] "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 (vs. age group).[11] For information that is unlikely to be included in the final analysis or write up (i.e. 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 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/identifying information are easier to handle safely and easier to prepare for sharing.

There are a number of 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

[10] Un-recorded data can still be inferred by context experts, however.

[11] 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.

is not ideal.[12] Second and related, 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. 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 separate theories about different phenomena, then "minimizing" with respect to those multiple objectives will not necessarily reduce the collection of sensitive information very much. Researchers who need to collect a very 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

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

be attractive to researchers because of its simplicity and relatively strong guarantees of success.

4.2 Preventing Re-Identification

Beyond data minimization, a number of 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 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, propsed 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 (see supplementary information for a demonstration).[13] Second,

[13] https://aidanmilliff.com/publication/data-security-agenda-for-improvement/QMMR$_{A}ppendix.pdf$

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 et al., 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 appendix,[14] 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 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

[14] Available online at: https://aidanmilliff.com/publication/data-security-agenda-for-improvement/QMMR$_{A}ppendix.pdf$

or de-anonymize given basic contextual knowledge. Text data can also 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). 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). 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 supplementary materials.[15] 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: a format which ignores word order, making it difficult to re-assemble the original natural language. For longer documents (multiple sentences containing multiple verbs, multiple subjects, etc.), 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 an "unsupervised learning" tool, researchers usually cannot pre-specify the topics they would like a model to focus on. There is no

[15] https://aidanmilliff.com/publication/data-security-agenda-for-improvement/QMMR$_{A}ppendix.pdf$

[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 re-identification of interview transcripts.

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 that 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.

4.3 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 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

[17] New work by Eshima et al. (2020) may mitigate this downside, allowing researchers to specify keywords for topic formation.

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 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 information. Supplementary materials provide step-by-step instructions.[18] 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 have to be applied carefully and tactfully. It could be really detrimental to a research partnership if partners felt dis-respected

[18] https://aidanmilliff.com/publication/data-security-agenda-for-improvement/QMMR$_{A}ppendix.pdf$

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 and hiring well.

5 Conclusion

This article has proposed new techniques for improving data security in qualitative (and quantitative) political science research. I have argued that 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 research transparent 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 increasing, which means 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, political scientists should not be content to wait for an even larger crisis to prompt 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. So, 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

O. Angiuli, J. Blitzstein, and J. Waldo. How to de-identify your data. *ACM Queue*, 13(8), 2015.

APSA Ad Hoc Committee on Human Subjects Research. Principles and guidance for human subjects research. Technical report, American Political Science Association, Washington, D.C., 2020. 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.

Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, 3(1), Mar 2013.

S. Eshima, K. Imai, and T. Sasaki. Keyword assisted topic models, 2020. URL https://arxiv.org/abs/2004.05964.

E. F. Foundation. An introduction to public key cryptography and pgp. November 2014.

L. A. Fujii. Research Ethics 101: Dilemmas and Responsibilities. *PS: Political Science & Politics*, 45(04):717–723, 2012.

E. Gibney. Internet research triggers scrutiny. *Nature*, 550:16–17, October 2017.

Identity Theft Resource Center. Data breach report 2016. Technical report, Identity Theft Resource Center, 2016.

A. M. Jacobs, T. Büthe, A. Arjona, L. R. Arriola, E. Bellin, A. Bennett, L. Björkman, E. Bleich, Z. Elkins, T. Fairfield, and et al. The qualitative transparency deliberations: Insights and implications. *Perspectives on Politics*, 19(1):171–208, 2021. doi: 10.1017/S1537592720001164.

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

G. King and N. Persily. A new model for industry–academic partnerships. *PS: Political Science amp; Politics*, page 1–7, 2019. doi: 10.1017/S1049096519001021.

C. R. Knerr. *What To Do Before and After a Subpoena of Data Arrives*, pages 191–206. Springer New York, New York, NY, 1982.

K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis. Tastes, ties, and time: A new social network dataset using facebook.com. *Social Networks*, 30(4):330 – 342, 2008.

E. McCallister, T. Grance, and K. Scarfone. Guide to protecting the confidentiality of personally identifiable information (pii). NIST Special Publication 800-122, National Institute of Standards and Technology, Gaithersburg, MD, April 2010.

B. McMurtrie. Secrets from Belfast. *Chronicle of Higher Education*, (January):1–41, jan 2014.

P. Menoret. Repression and Fieldwork. In *Joyriding in Riyadh*. Cambridge University Press, New York, 2014.

F. Mosteller and D. L. Wallace. Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302):275–309, 06 1963.

A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy*, pages 111–25. IEEE, 2008.

P. Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, 58(3):1701–1778, 2010.

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

M. E. Roberts, B. M. Stewart, D. Tingley, and E. M. Airoldi. The structural topic model and applied social science. In *NIPS 2013 Workshop on Topic Models: Computation, Application, and Evaluation*. NIPS, 2013.

P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, 1998.

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

K. Shilton. Emerging ethics norms in social media research, 2016.

S. Summers. *Organising, Storing and Securely Handling Research Data*. UK Data Service, Essex, England, June 2016.

L. Sweeney. Simple demographics often identify people uniquely, 2000.

L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

M. Traynor. Countering the excessive subpoena for scholarly research. *Law and Contemporary Problems*, 59(3):119–48, 1996.

S. Venkatesh. *Gang Leader for a Day*. Penguin, New York, 2008.

E. J. Wood. Field Research. In *The Oxford Handbook of Comparative Politics*. Oxford University Press, Oxford, 2007.

E. J. Zechmeister. Ethics and Research in Political Science: The Responsibilities of the Researcher and the Profession. In S. Desposato, editor, *Ethics and Experiments*. Routledge, London, 2016.

M. Zimmer. More on the 'anonymity' of the facebook dataset - it's harvard college, 2008. URL http://michaelzimmer.org/2008/10/03/moreon-the-anonymity-of-the-facebook-dataset-its-harvard-college/.

Appendices

Supplemental Information: Three Privacy Protection Demos

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 et al., 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-anonymzation.[19]

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*.[20]

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 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

[19] Perturbative 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.

[20] In 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.

2. A computer running `R` 2.10 or newer
3. An installation of the `sdcMicro` package from CRAN—this demo uses version 5.5.1[21]

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.

```
  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
```

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

```
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%)
------------------------------------------------------------------------
```

2. 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.[22]. 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)
```

[22] Counterintuitively, the function `microaggregation()` does something else

```
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.

```
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?
# Varibles 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 seperate 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 mod-

ified 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 %)
----------------------------------------------------------------------
```

5. 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
```

6. 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)
```

7. 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 eignevalues) 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.

Privacy Protection with Qualitative Data: Topic Modeling on Small Corpora

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.

I use structural topic modeling to accomplish this task. 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 et al. (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

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

	Dependent variable:	
	ben_income	
	Pre-modification	Post-modification
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

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.

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). 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.

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.

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 et al. (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 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 attemopt 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 assignations 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 metadata 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

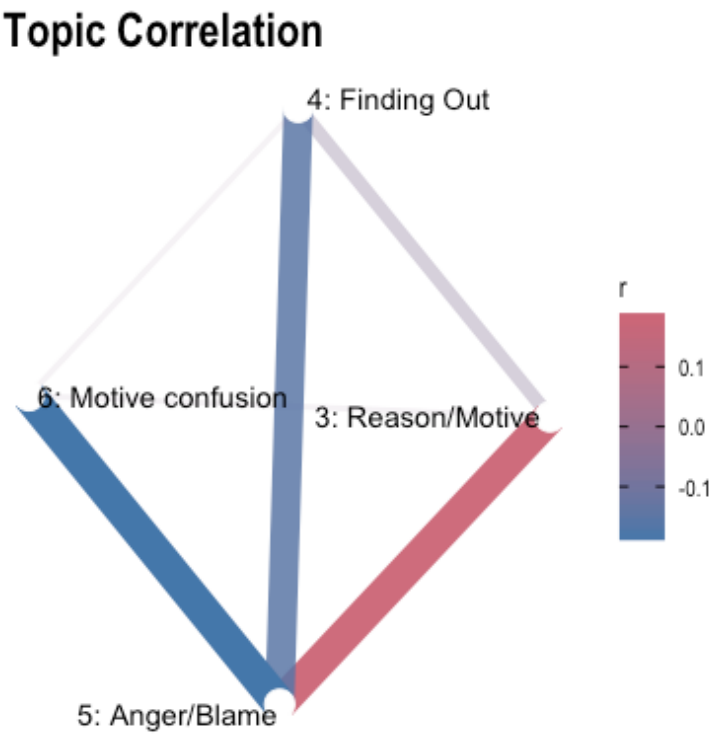

Figure 1: Inter-topic correlation for topics where $r > 0.1$ with Topic 5 (anger, blame).

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.

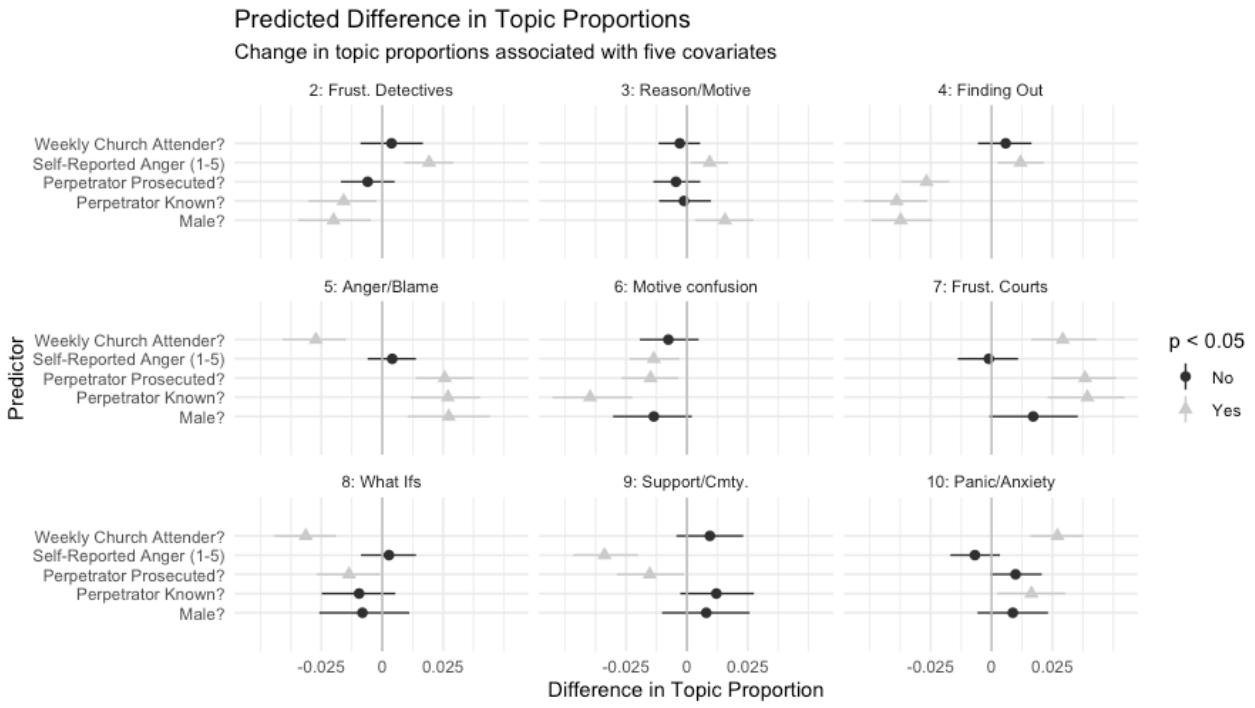

Figure 2: Bivariate associations between respondent characteristics and topic proportions

PGP "LockBox"

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,[23] 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 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, they cannot be decrypted without the private key. After encryption, the researcher destroys the unencrypted data. The encrypted data are then inaccessible until the researcher returns to the physical machine that

[23] A 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.

has the private key.

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.[24]

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.

[24] Instructions 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.

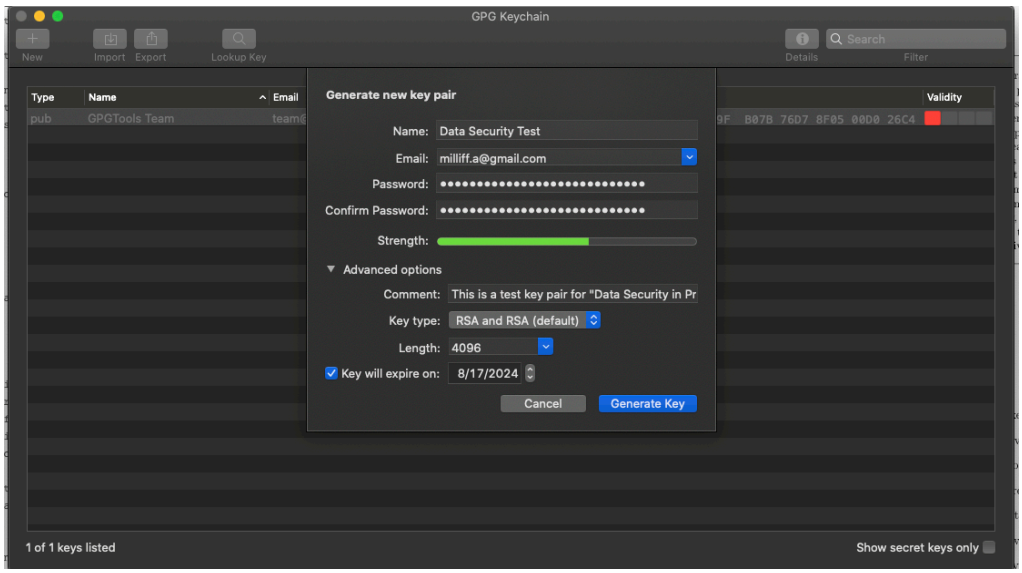

Figure 3: Creating a new keypair in GPG Keychain for Mac OS X

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!"*

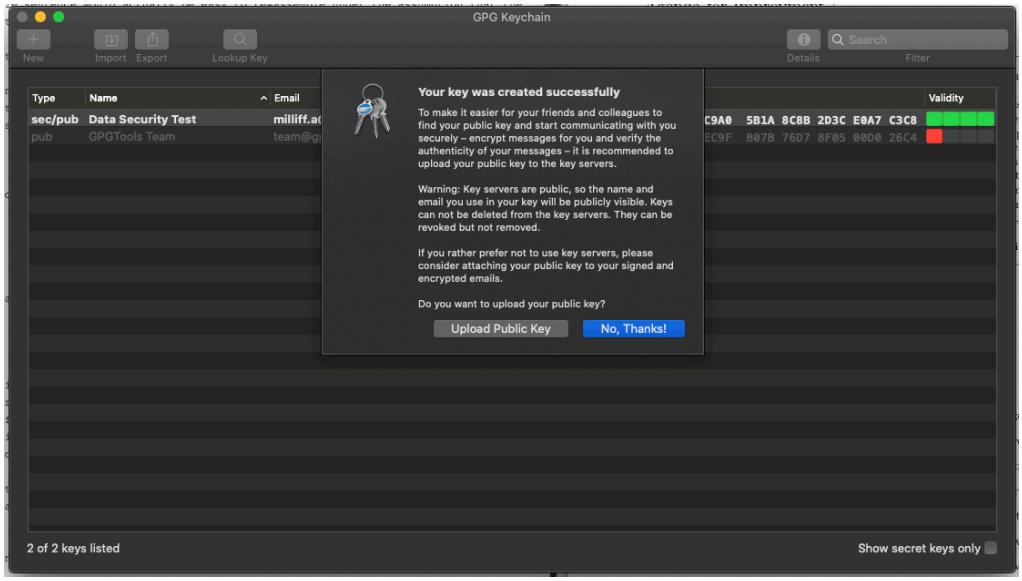

Figure 4: Confirmation of new key pair, GPG Keychain for Mac OS X.

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."

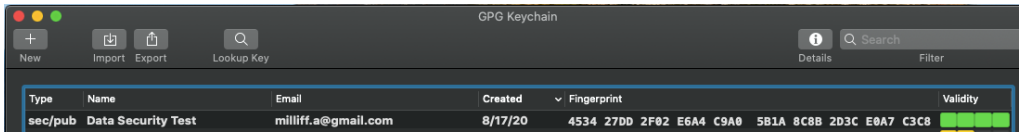

Figure 5: Verifying key pair type on the stay home computer. The top line shows that both secret/private and public keys for "Data Security Test" fingerprint 4534... are stored on this machine.[26]

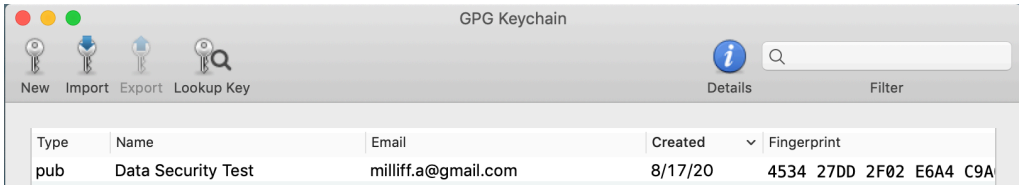

Figure 6: Verifying key pair type on the data collection computer. The top line shows that only the public key for "Data Security Test" fingerprint 4534... is stored on this machine.

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."

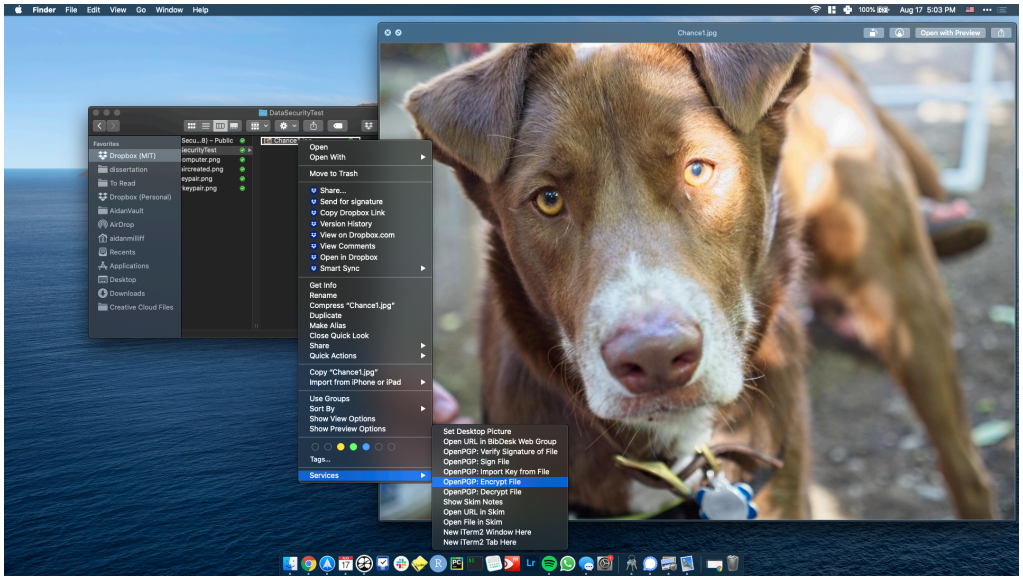

Figure 7: Menus that appear after right-clicking a file in Finder.

(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.

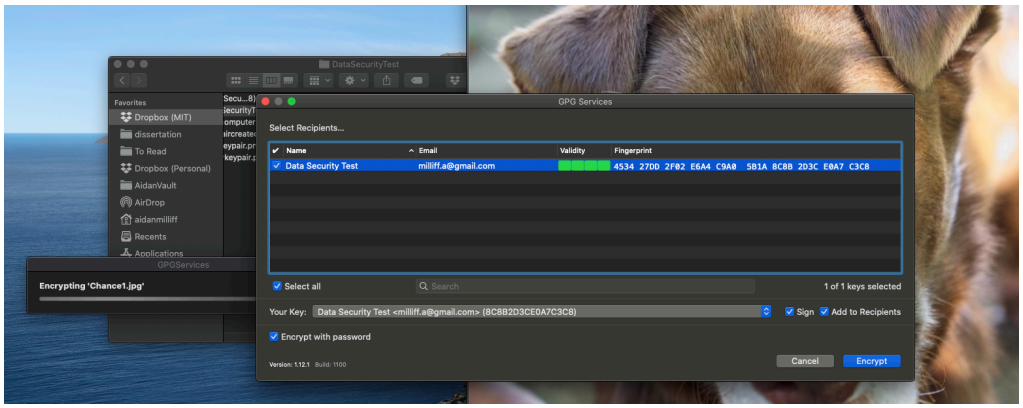

Figure 8: Selecting a key for encryption with a passphrase.

(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.*[27]

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.

[27] As 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).

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.

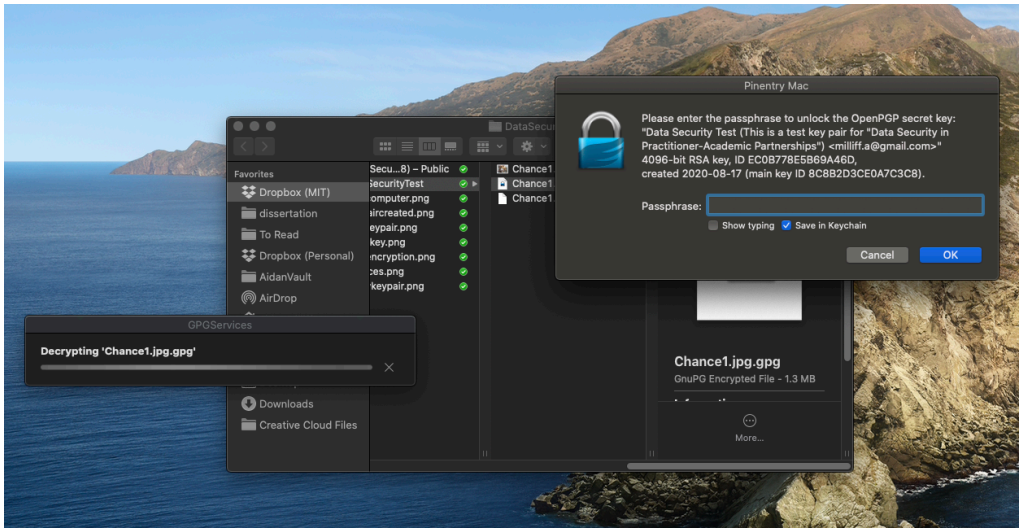

Figure 9: Prompts for decryption.

   (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.

References in Supplementary Information

Blei, D., A. Ng, and M. Jordan (2003). Latent dirichlet allocation. *Journal of Machine Learning Research 3*(993-1022).

Brandom, R. (2014). New documents reveal which encryption tools the nsa couldn't crack. *TheVerge.com*.

Desai, S. and R. Vanneman (2015). India human development survey ii (IHDS-II).

Grimmer, J. and B. M. Stewart (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis 21*(3), 267–297.

Milliff, A. (2020). Facts shape feelings: An information-based framework for emotional responses to trauma. Working Paper, 2019-06, Massachusetts Institute of Technology, Cambridge, MA.

Roberts, M. E., B. M. Stewart, and D. Tingley (2018). stm: R package for structural topic models. *Journal of Statistical Software*.

Templ, M., B. Meindl, and A. Kowarik (2020). Introduction to statistical disclosure control SDC. Technical report, Zurich University of Applied Sciences, Zurich, Switzerland.