Contents

1

Options of Accessing Confidential Data

Lars Vilhuber

Cornell University

CONTENTS

This chapter will rely on and update previous overviews of how researchers, citizens, and administrators can reliably and securely access confidential data, i.e., data that cannot be simply published as "open data". I will discuss various legal, technical, and practical ways of securing access to data that is needed for computations. This obviously depends on the type and complexity of the computations, but also depends on the who, how, and where access is needed.

It might seem appropriate here to define what type of data I will be talking about. Several designations circulate in the various communities that create, handle, and use data: private, sensitive, confidential, proprietary. These terms are not quite interchangeable. Data may be sensitive, because they might contain attributes which could create a risk of harm to individuals or other entities described by the data. However, some sensitive data may be publicly available. In some US states, arrest records will list the name and home address of the arrested individual, but do not necessarily list if the person was ever convicted for the offense for which they were arrested. This might create harm for the individual, but the data themselves are public, and not confidential. Data might also be owned by a specific individual or institution, thus

in private ownership, but not be confidential. Often, such data are referred to as proprietary. For instance, the well-known "S&P 500" data are owned by a company called Standard and Poor's, can be viewed and downloaded by anybody (f.i. [19]), but cannot be redistributed by the downloading user: they are proprietary, but certainly not sensitive or confidential. In general, I will therefore refer to "confidential" data as being the key attribute that data custodians are concerned about, regardless of whether the data are proprietary or sensitive.

I specifically exclude from this discussion mechanisms to obtain or collect confidential data from individual respondents (people and firms). Interested readers should consult [13]. This chapter will focus on access by analysts once data has been collected, with one exception: we will describe some of the foundational aspects of multi-party computing (legal frameworks, consent issues, etc.), though all of the technical aspects of multi-party computing are left to the various chapters in Part 4.

In writing this chapter, I will rely on a variety of publications. [11] touches on a few of these mechanism, and an update is being prepared as of this writing. I have previously written about access options to firm-level data [25], which in turn referenced older summaries such as [26]. I will use framing from [9] and [4]. Astonishingly, many of the access methods in use today are not very different from those implemented nearly 20 years ago, but I will briefly describe several newer approaches, relying on previous work by myself and co-authors [18].

1.1 Framing: Four Regions, Five Safes, or Five Colors

Statistical techniques for disclosure avoidance, and secure computational methods, are conducted with the end goal of reducing the sensitivity of data so that public statistical inferences can be made. The end goal is necessarily data that is publicly available, though that may be a statistical summary as compact as a single number, or even just a qualitative range of numbers, e.g., "positive". In some cases, data is processed in such a way that fairly high-dimensional data can be made publicly available to researchers to use as they wish. In other cases, there may be no statistical protection mechanism that simultaneously allows for detailed analysis and valid inference in the public domain, and other mechanisms must be used. It is clear that to resolve the tension between inference and accessibility of data, there is no single omnibus solution. Data custodians, statisticians, and ethicists have long recognized this.

Conceptually, it is useful to distinguish between raw data, data transformed for disclosure avoidance purposes in such a way as to be amenable for statistical processing by the analyst, and the final statistics used for in-

ference. Consider the transformed data first. If the reduction in sensitivity of the confidential data is such that it can be made available with negligible expected harm, while maintaining analytical validity, it can be published without restrictions. However, if despite extensive application of disclosure avoidance risks, the data could still potentially lead to negative outcomes for some respondents, a data custodian may decide to require certain safeguards or promises by analysts before they can download the data. In the extreme, if the only way to maintain the utility of the data is to retain direct identifiers in the clear, the data custodian may want to retain very tight control over who can access the data, where that access occurs, and for what purpose.

This is the gist of a classification of tradeoffs proposed by [4], a simplified depiction of which is provided in Figure 1.1. In [4], various combinations of identifiability of the transformed data and expected harm are mapped into four levels of 'privacy controls'. Identifiability, which in [4] is tied to specific disclosure avoidance methods, ranges from the "strongly protected" to "not protected at all", whereas expected harm may range from "negligible" to "catastrophic".[1] Level 1 corresponds to the public-use data case - no controls are imposed on the protected data. For Level 2, analysts must subject themselves, or are subject to by virtue of the legal environment, to terms of use, click-through agreements, notices, and the like. For Level 3, the data custodian requires an approval process - not only must the analyst agree to conditions, but this is tied to an approval process, possibly by some third parties (such as institutional review or ethics boards). Finally, for Level 4, much tighter controls are imposed through "data enclaves" (which we will define shortly).

An alternative conceptualization specifically breaks out five dimensions of access control for confidential data - the transformed intermediate data mentioned earlier. It therefore focuses on the various environments associated with Levels 2-4 from Figure 1.1. This is known as the "Five Safes" model [9], and is used by many statistical agencies and data custodians, such as the UK Government [23], Statistics Canada [21], and regional governments such as Canada's province of British Columbia [7].

The "Five Safes" model proposes the following "safes" or dimensions: projects, people, settings, data, and output. *Safe projects* is how the data custodian assesses the overall use of the data as proposed by an analyst. An evaluation may be required by law, as a requirement of an ethics review, or as a matter of policy.[2] By defining what *safe people* are, the data custodian identifies who can be trusted. Thus, whereas a safe project may be defined as a research project, safe people may be defined as researchers affiliated with a recognized academic institution — excluding, for whatever reason,

[1] See the Harvard Data Policy [14] for a framing (and colorization) of these levels of sensitivity in data.

[2] We focus here on project vetting for the purpose of disclosure avoidance, and not for other purposes, such as congruence with policy objectives. An example of the latter might be a firm or a government agency only allowing for projects that show the data provider in a positive light.

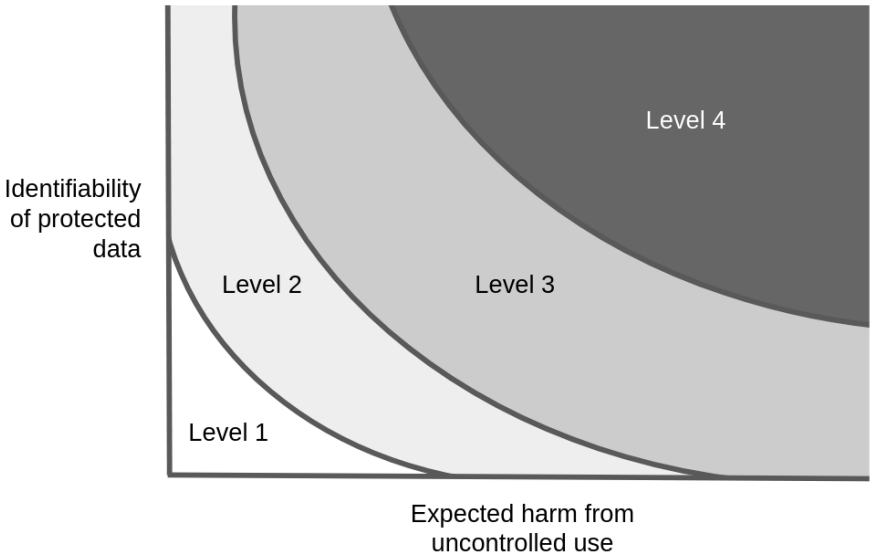

FIGURE 1.1
Regions of equivalent security-access tradeoffs

non-academic researchers. It may also be defined through a legal definition, such as the British Digital Economy Act's "accredited researchers" [23], or by having satisfied some training requirements. The environment in which data access happens is described through "safe settings". This typically involves both technical and physical security of the place where access to and computing on the transformed data occurs.

"Safe data" and "safe outputs" more narrowly cover what the rest of this book will address: How data are transformed, first from the highly confidential raw data to the transformed data accessed by analysts, and subsequently, how the results from the analysis are processed to be able to release them to the general public. These two dimensions cover two distinct parts of the overall research workflow, but do not necessarily entail the use of different methods. For instance, both might use aggregation. Safe data might be created via micro-aggregation, for instance of worker-level data to the firm-level, and safe outputs might be that created by reporting all analyses only highly summary geographic granularity.

It should be clear that these five dimensions intersect with each other to provide a specified (perceived) level of control over the disclosure risk in the underlying raw data. For instance, the choice of safe settings will influence the strength of the disclosure avoidance methods applied to create safe data. The types of projects chosen may affect the methods used for the creation of safe outputs.

1.2 Examples of Access Mechanisms

I start by describing three examples of access mechanisms, which I will subsequently use to illustrate the various dimensions and characteristics of access mechanisms. Additional examples are provided in [18].

First, consider the Norwegian "microdata.no" website.[3] Researchers, attached to specific educational institutions, access confidential microdata through a website with prior registration and application. Registered users must be affiliated with an 'approved research institution' and must be holders of a Norwegian identity card. Computations are run on servers controlled by Statistics Norway, using a programming language similar to, but not identical to Stata. Outputs are protected using an automatic application of disclosure avoidance methods, can simply be downloaded.

Consider then the U.S. **FSRDC!** (**FSRDC!**).[4] Researchers are authorized to access confidential microdata after application and security vetting. Minimum residency requirements apply, at least for university-based researchers. Data are processed on computers housed at the U.S. Census Bureau's computer center. Each project is independently reviewed for compliance with certain criteria. Researchers access these data from thin clients located in secure rooms located at research institutions (universities and other government agencies), access to which is controlled by the U.S. Census Bureau. A second mode of access has gained ground since the COVID pandemic, namely access for those researchers remotely from other authorized locations, such as home offices, using virtual desktop software.[5] In both cases, researchers lack the ability to upload or download any materials from the remote computing facility.

Finally, consider the **NCES!** (**NCES!**) restricted data licenses.[6] Researchers' (US-based) organization sign a licensing agreement with the agency, and submit a security plan. They must also justify the need for access to the confidential data, but need not do so on a per-project basis. Organizations are limited to research institutions, government agencies, and consulting frims. The organization-controlled standalone, secure desktop computer must be located in a separate room (controlled by the university, not the agency). Researchers can upload and download files, but commit to not downloading any of the data.

[3] https://www.microdata.no/en/

[4] Access procedures vary by agency data, https://www.census.gov/about/adrm/ced/apply-for-access.html describes the procedures for data provided by the U.S. Census Bureau.

[5] Virtual or remote desktop software displays a remote computer desktop on the screen of a local computer. In some cases, interactions of the remote desktop with the local computer are inhibited, and only keyboard and mouse control are sent to the remote desktop, practically acting as if it were a thin client.

[6] https://nces.ed.gov/statprog/rudman/j.asp, accessed 2022-04-16.

1.3 Safe People: Exclusions, Inclusions, and Training

Who can be trusted to handle data of varying degrees of sensitivity is one of the key aspects of any access mechanism. The open-access movement of the last decades has made access to non-sensitive data much easier in many countries, building on earlier efforts such as the Canadian 'Data Liberation Initiative' [8], and leading to the widespread adoption of open data licenses [20, 22]. In the United States, of course, public-use micro data have been around for much longer, but as with Canada and the UK, the original files of what are now called public-use microdata samples (PUMS) were sold, not made freely available, and required users to have access to mainframes (U.S. Census Bureau, cited in [5]). De facto, only 85 research institutions [5] purchased these files. Thus, even though anybody could purchase these, there were technical requirements that prevented many from accessing these files. Today, users can simply download the same 1960 PUMS files from IPUMS [**?**] onto a relatively cheap personal computer.

The 1960 PUMS files were subject to relatively simple disclosure avoidance techniques (sampling, coarsening, and de-identification), and whether the limitations on its user base due to the technical requirements were taken into account is unknown. But today, PUMS files are available to any user, without restriction.

For more sensitive files, however, many data providers restrict the user base in various ways, treating some users as "more trusted" then others. Such conditions may be by affiliation (university-affiliated), presumably associating users employed by certain institutions as being more likely to be 'safe'. Conditions can emphasize citizenship or residency requirements. In the Norwegian case, researchers must be affiliated with a limited list of institutions, and must have been a Norwegian resident sometime in the past. For the FSRDC, a three-year minimum residency requirement applies. Such conditions are in practice exclusionary against certain groups. Citizen science becomes nearly impossible to perform, and even journalists may have difficulty in complying with such rules, even though neither may be explicitly excluded.

People may also be considered 'safe' if they have received security and sensitivity training. The IDAN Network lists a comparison for six different European access mechanisms [15]. Users of the FSRDC have yearly IT security and confidentiality training. Users of the Norwegian or NCES data do not appear to have required training, but their host institutions may require their own training.

1.4 Safe Settings and Projects: Legal and Contractual Mechanisms

The users accessing the data do so within certain legal and contractual mechanisms: the legal component of 'safe settings'. These include general laws such as the European Union's General Data Protection Regulation (GDPR) [3], or the U.S. E-government Act of 2002 [2], and more specific laws, such as Title 13, U.S.C., which applies specifically to data collected by the U.S. Census Bureau [1, 24]. Such legal frameworks provide for monetary penalties and potential prison sentences for illegitimate uses of the data, and may define what legitimate uses are.

Other data providers rely not just on legal mechanisms, but may also leverage contractually convened upon penalties. In the NCES, not only are researchers subject to various laws, including the aforementioned E-government Act, but also have contractually agreed-upon penalties. In particular when crossing international borders — which none of the examples I have given allow for — contractual obligations and penalties, enforceable by local legal systems, may be more common. Private providers of confidential data may rely exclusively on such mechanisms.

Physical settings matter, and I discuss these in more detail in the next section, but even beyond the fact that some researchers must go into a *locked room* on a university campus, it may matter that those researchers are *on campus*, and not elsewhere. I speculate that this may affect the particular mindset of the researcher, by possibly being surrounded by other researchers working on similarly sensitive data, magnifying the effect of previously received training.

The selection of projects also interacts in subtle ways with the legal framework and other dimensions of the Five Safes framework. In some cases, legal frameworks require that only certain types of projects be authorized. For instance, certain U.S. states allow access to their data only if the research directly improves program compliance. Often, the expected type of project serves as a filter for the type of expected output. Academic projects often result in very compact model estimates, not vast tables, and thus facilitate the type of output control ('safe outputs') where traditional disclosure avoidance methods can be manually applied.

for methods centered on surveys, and [16] on the limits to ethical data sharing

1.5 Safe Settings: Physical Mechanisms

When considering the 'safe settings' part of the Five Safes, physical environments play an important role, in addition to the legal setting. Physical environments in the broadest sense are those places where data is stored and processed, and where researchers sit when accessing the data. The two places are not necessarily the same, even though inevitably the place where researchers sit also has computers. Generally, one should consider the entire data access mechanism as part of the 'safe settings'.

In [18], we defined five aspects of data access mechanisms, by which we classified data access mechanisms currently in use, or being planned. These aspects are:

- the level of **researcher agency over analysis computers** — how much can a researcher, as opposed to an IT specialist or data custodian, manipulate the computing environment;
- the **location of analysis computers** — the computers on which computations are run, whether they are located in the same space as the researcher, the data provider, or a third party;
- the **location of access computers** — which historically was the same as the analysis computer, but in modern systems is generally separate;
- the **level of access security** over both the physical environments (rooms) and access computers;
- the **range of analysis methods available**.

I will briefly summarize here what each of these aspects implies for the access mechanism, and how they allow to characterize some of the most common access mechanisms; the reader is referred to [18] for more details. I will also point out the implications for the need for safe data, as well how access mechanisms interact with the feasibility of certain safe output mechanisms.

1.5.1 Researcher Agency over Analysis Computers

How much control do researcher have over the analysis computers? In some cases, researchers will have very little influence about most characteristics of the analysis computers, because these computers are managed and controlled by the data custodian, with little to no ability to change the setup. In other cases, researchers may be able to choose or request software, or define specific hardware setups. In the highest researcher agency setting, researchers have full control over the analysis computers, because it may simply be the research workstation or laptop that they use for many other activities as well. Note that this also affects how researchers can interact with those analysis computers

— is it possible to upload researcher-created scripts directly to the analysis computers, or are such insertions of computer code into the analysis computer mediated by semi-automated or manual mechanisms.

Consider the XXXX sample mechanisms outlined earlier. In the Norwegian case, users have no control over the choice of software or hardware capabilities of the analysis computers - they have low agency. In the FSRDC case, they also have no control over analysis computers, but a broad spectrum of software options are available to choose from, and others can be requested. Even when the analysis computers are accessed from home offices, this does not change: the same analysis computers are being accessed as would have been accessed from the secure on-campus room. They have medium agency over the analysis computers. Finally, in the NCES case, researchers have full agency over the analysis computers, except for certain security requirements. Any software can be loaded on the analysis computer, and scripts can be easily uploaded.

1.5.2 Location of Analysis Computers and Data

But where can those analysis computers be located? In many situations, the analysis computers may be located with the data *provider*. This is often the case when private companies, or non-statistical government agencies, provide access to confidential data "on premise". In other cases, the data may be hosted with a secure and trusted third-party, a data *custodian* or intermediary. For most practical purposes, these two cases are indistinguishable for the researcher — they are primarily distinguished from the case where the analysis computer is located with the researcher. In more recent years, many data providers have moved to "offload" data access to a distinct entity, although that entity often resides within the same larger organizations.

Consider again the three examples above. In both the Norwegian and the FSRDC case, analysis computers and data are located with the data provider. However, while the U.S. Census Bureau used to be its own data access provider, it has increasingly expanded use of the **FSRDC!**s to also host data from other agencies. For those data, the U.S. Census Bureau, providing the FSRDC services, acts as a trusted third-party to other agencies. Finally, in the NCES case, the analysis computer and the data are located in premises controlled by the researcher (or the researcher's home institution).

1.5.3 Location of Access Computers

Researchers may traditionally think of their "workstation" or "laptop" as the access computer, when the data are on storage media directly attached. This, in fact, is the likely modus operandi when using most public-use data. However, when processing confidential data, the analysis computer may not be physically accessible to the researcher, and must receive instructions via a separate access computer. Naturally, the researcher must always be able to directly access the access computer, but the location — with the non-

researcher data custodian, a third-party access provider, or with the researcher — plays an important role.

The NCES licensing case still leverages the traditional model of the workstation-as-the-access-computer — access computer and analysis computer are coincident, and in researcher-controlled space. In the case of the FSRDC, the access computer is either a thin client, located in the on-campus secure room, or special software running on a researcher's laptop, from a single authorized location that is typically the researcher's home office. In the Norwegian case, a simple web browser, accessed from any computer anywhere in the world, provides access to the analysis computers.

1.5.4 Security of Access Computers

Regardless of location, there is substantial variation in the implemented security of the access computers and the premises where it is located. Data providers often require the right to approve the security arrangements, conduct audits, or otherwise directly verify that the operator is in compliance with the mandated security requirements. They may require specific third-party operators of facilities or computers. Heuristically, the security of access computers and their rooms can range from highly secure to unsecured (beyond basic computer security).

Consider the Norwegian case. Since access is web-based, even the unsecured hotel lobby computer might conceivably be used to access the system. In contrast, the physical facilities traditionally used by the FSRDC system have building requirements (specifying wall construction, window placement, alarm systems, etc.) and computer restrictions (thin clients with no local storage). While newer remote-access permissions relax some of those restrictions, they still specify locality (home locations, but not the next internet café) and software (only specific virtual desktop software will allow for access, and prevents all interaction with the host computer). In the NCES case, secure rooms are also required, with certain, lighter, requirements – a lock on the door and restricted access appears to be sufficient [17]. Certain security requirements are enumerated for the access computer, but they are likely to be satisfied by any regular university-managed computer.

1.5.5 Range of Analysis Methods Available

While somewhat overlapping with the **Researcher Agency over Analysis Computers**, we considered this to be a distinct aspect, because even when agency is low, there may well be significant differences in the available analysis methods, and these materially affect the researcher's ability to apply and incorporate disclosure avoidance methods, such as those outlined in the other chapters.

Consider the distinction between the Norwegian system and the FSRDC. In both, researchers have no control over the analysis computer, but in the

Norwegian case, there is only one possible programming language, whereas in the FSRDC case, there are many (and because agency is slightly higher, additional ones can be requested). Often, remote submission systems (similar to the Norwegian system) are limited to a single programming language, often with explicitly limited capabilities.

1.5.6 Typical Access Mechanisms

There are a variety of different access mechanisms that researchers typically encounter. I briefly sketch a few here, see [18] for details. The simplest we already outlined for the NCES case: users sign an agreement, obtain the data, and work on a computer they provide, possibly in a locked room. Such a mechanism provides high agency over the analysis computer, a simple IT setup, and allows for a wide range of analysis methods. It is, however, potentially complex to manage at scale, when thousands such agreements are to be monitored. However, it is still quite frequently used for surveys (PSID or HRS confidential data) or some proprietary data (Kilts Center), but is seldomly used nowadays for data provided by national statistical agencies. Those have tended to use designated physical infrastructure — commonly called a "**research data center**" — and historically physical shipment of data to analysis computers located within those spaces [8] — therefore also known as a "physical data enclave". Such a system provides less researcher control over the physical location, and generally less agency over the analysis computers. Yet it also often generates diversity within such a network of research data centers, for instance in computational capacity. More and more data providers have shifted to, or even started out with a variant that incorporates a strong remote component, which may be called "virtual data enclave", "virtual desktop infrastructure (VDI)", or "thin client" (which is a dedicated device used for connecting to a virtual desktop). The access computers may still be in dedicated secure rooms (the traditional FSRDC or the IAB model), in researcher offices (the French Centre d'accès sécurisé de données, CASD), or through dedicated software from researchers' homes (the newer FSRDC model). Some providers of proprietary data will also provide "secure laptops" that de facto use remote connection protocols to connect to corporate infrastructure. The virtual model does not necessarily modify researcher control over locations — in all of the mentioned examples, the data custodian specifies a specific location — but unifies the type of data and software available to hundreds or thousands of researchers. It also increases the sense of security on the data provider side, since access to the data can be terminated quickly and completely, in contrast to the other two mechanisms. Whether that constitutes a de facto increase in actual security is unknown, and may be unknowable. Finally, the Norwegian mechanisms, as well as the Canadian RTRA, are examples of **remote submission** systems, where code is sent to a remote system through a non-interactive mechanism, and researchers then receive results once computations are completed. Computer scientists might

recognize this as a variant on application programming interfaces (APIs), used for instance for geocoding of addresses.

1.6 Discussion

Increase the level of protection applied to 'safe outputs', reducing the emphasis on all other dimensions, and "public-use data" is the result, as we originally pointed out. Shift some protection onto the creation of 'safe data', for instance by creating non-DP synthetic data, and relatively liberal distribution policies might be feasible, relying only on click-through user agreements, but with the downside that inference is uncertain. In other instances, such as Statistics Canada's Real Time Remote Access, the range of analysis methods as well as the set of 'safe people' using it are restricted, but inference validity is maintained and the convenience of the access is increased. Alternatively, by offering "factually anonymous datasets", restricting users to academic researchers at universities may be sufficient, as in the case of the German Institute for Employment Research [12].

In reading the other chapters of this book, readers may want to keep in mind that all levers of protection — the Five Safes — contribute to the overall protection of confidential data, and often, multiple access channels, with different combinations of the level of each of the elements, may be necessary. Almost all data providers that offer public-use data require that applicants demonstrate that the public-use data are insufficient for addressing the research question at hand before considering approval of access to the confidential data. The IAB offers campus data (synthetic data that is not inference-robust), scientific use files (factually anonymized, coarsened, and with a limited number of variables), and files in the research data center of the IAB, all of which have been derived from the confidential internal data of the agency. This would seem to be particularly true when using synthetic data, of either the classical or DP variant, as such data may be valid for some use cases, and yet not valid for others. [6] for instance proposed a mechanism that would signal when inferences might be invalid, though they did not offer an alternate mechanism. [10] illustrated this in the context of interactive usage, restricting access to a subset. The idea could be expanded to a variety of protection mechanisms. In particular, suppose that users have access to a DP-synthetic public-use dataset akin to [6]. Using a **remote submission system**, researchers obtain either an inference-valid result, or a signal that the inference may not be valid. In the latter case, re-submission might happen semi-automatically to a different remote submission system, with access to the same data as [10], and results are made into 'safe outputs' using a different but also DP mechanism. Such a mechanism does not yet exist, but is a

logical consequence of combining various combinations of the Five Safes in a way that reduces both human and computational burden.

Bibliography

[1] Title 13.

[2] E-Government Act, 2002.

[3] General Data Protection Regulation (GDPR), 2016.

[4] Micah Altman, Alexandra Wood, David O'Brien, Salil Vadhan, and Urs Gasser. Towards a Modern Approach to Privacy-Aware Government Data Releases. *Berkeley Technology and Law Journal*, 1967, 2015. tex.ids: altman2015towards.

[5] Margo J. Anderson. *The American census: a social history*. Yale University Press, New Haven, second edition edition, 2015. OCLC: ocn904081530.

[6] Andrés F. Barrientos, Alexander Bolton, Tom Balmat, Jerome P. Reiter, John M. de Figueiredo, Ashwin Machanavajjhala, Yan Chen, Charley Kneifel, and Mark DeLong. Providing Access to Confidential Research Data Through Synthesis and Verification: An Application to Data on Employees of the U.S. Federal Government. *The Annals of Applied Statistics*, June 2018. arXiv: 1705.07872.

[7] BC Ministry of Citizens Services. Privacy, Security and the Five Safes Model. Last Modified: 2019-07-30 Publisher: Province of British Columbia.

[8] R.F. Currie and S. Fortin. *Social Statistics Matter: A History of the Canadian RDC Network*. Canadian Research Data Centre Network = Réseau canadien des Centres de données de recherche, 2015. tex.ids: Currie2015.

[9] Tanvi Desai, Felix Ritchie, and Richard Welpton. Five Safes: designing data access for research. Working Paper, University of the West of England, 2016.

[10] Cynthia Dwork and Jonathan Ullman. The Fienberg Problem: How to Allow Human Interactive Data Analysis in the Age of Differential Privacy. *Journal of Privacy and Confidentiality*, 8(1), December 2018.

[11] FCSM. Report on Statistical Disclosure Limitation Methodology. Technical Report 22 (Second version, 2005), {Federal Committee on Statistical Methodology}, 2005.

[12] Forschungsdatenzentrum der BA im IAB. Scientific Use Files.

[13] Robert M. Groves, Floyd J. Fowler Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. *Survey methodology*. Wiley series in survey methodology. Wiley, Hoboken, N.J, 2nd ed edition, 2009. OCLC: ocn302189175.

[14] Harvard University. Data Security Levels - Research Data Examples.

[15] IDAN Network. IDAN Network - Accreditation.

[16] Michelle N. Meyer. Practical Tips for Ethical Data Sharing. *Advances in Methods and Practices in Psychological Science*, 1(1):131–144, March 2018.

[17] National Center for Education Statistics. Appendix J: Restricted-use Data Security Plan Form. Publisher: National Center for Education Statistics.

[18] Jim Shen and Lars Vilhuber. Physically Protecting Sensitive Data. In Shawn Cole, Iqbal Dhaliwal, Anja Sautmann, and Lars Vilhuber, editors, *Handbook on Using Administrative Data for Research and Evidence-based Policy*, pages 37–84. Abdul Latif Jameel Poverty Action Lab, January 2021.

[19] S&P Dow Jones Indices LLC. S&P 500 [SP500]. Technical report, FRED, Federal Reserve Bank of St. Louis [distributor], June 2020.

[20] Statistics Canada. Statistics Canada Open Licence, February 2012. Last Modified: 2021-10-29.

[21] Statistics Canada. Information on Statistics Canada Privacy Framework. Technical report, November 2018.

[22] UK Government. Open Government Licence for public sector information V3, 2014.

[23] UK Government. Digital Economy Act 2017: Research Code of Practice and Accreditation Criteria, February 2020.

[24] US Census Bureau. Federal Law, October 2021. Section: Government.

[25] Lars Vilhuber. Methods for Protecting the Confidentiality of Firm-Level Data: Issues and Solutions. Technical Report 19, Labor Dynamics Institute, March 2013.

[26] Daniel H Weinberg, John M Abowd, Philip M Steel, Laura Zayatz, and Sandra K Rowland. Access Methods for United States Microdata. Technical Report 07-25, Center for Economic Studies, U.S. Census Bureau, September 2007.