The Five Safes as a Privacy Context

James Bailie jamesbailie@g.harvard.edu Harvard University Cambridge, Massachusetts, USA Ruobin Gong ruobin.gong@rutgers.edu Rutgers University Piscataway, New Jersey, USA

Abstract

The Five Safes is a framework used by national statistical offices (NSO) for assessing and managing the disclosure risk of data sharing. This paper makes two points: Firstly, the Five Safes can be understood as a specialization of a broader concept – contextual integrity – to the situation of statistical dissemination by an NSO. We demonstrate this by mapping the five parameters of contextual integrity onto the five dimensions of the Five Safes. Secondly, the Five Safes contextualizes narrow, technical notions of privacy within a holistic risk assessment. We demonstrate this with the example of differential privacy (DP). This contextualization allows NSOs to place DP within their Five Safes toolkit while also guiding the design of DP implementations within the broader privacy context, as delineated by both their regulation and the relevant social norms.

Keywords

contextual integrity, differential privacy, the Five Safes framework

1 Introduction

As a supplier of official data, a national statistical office (NSO) is an integral part of a well-functioning democratic state. Its data are essential for informing government policy, business strategy and academic research, thereby advancing society and driving economic growth [20]. Yet an NSO's ongoing value depends upon maintaining its social license to collect and share data. It is therefore necessary that NSOs balance their social and economic utility with the privacy of their data providers. With the recent increase in resources available to malicious actors and the growth of competing data vendors, this trade-off is increasingly difficult to manage [5].

The Five Safes is one tool that assists NSOs in balancing the utility derived from sharing their data with the consequent privacy risks. Originally developed in 2003 to enable researchers' access to Office of National Statistics (ONS) microdata [13], its use has since expanded to all forms of statistical dissemination [4]. It has been employed by NSOs in the UK, Australia, and New Zealand to guide design decisions and risk assessments for statistical disclosure control (SDC) [4, 19, 25–27]. In the USA, the Five Safes have been used by the Coleridge Initiative in the context of data sharing within and across states, government agencies and researchers [17]. Further, the Advisory Committee on Data for Evidence Building recently recommended that the use of the Five Safes in the US federal bureaucracy be expanded [2].

Broadly speaking, the Five Safes is a framework for designing and assessing modes of statistical data sharing in ways which maintain the privacy and confidentiality of data providers. More specifically, it is both a theory that decomposes disclosure risk into five dimensions, or 'safes,' as well as a loose set of guidelines for how an NSO

can manage risk by appropriate choices in each of these five dimensions. These two parts of the Five Safes have been conflated in previous work – often along with additional hypotheses and judgments concerning disclosure risk – which has led to considerable confusion surrounding what exactly the Five Safes comprises [18].

In this work, we bring the perspective of contextual integrity (CI) to isolate and explicate the first – and most central – part of the Five Safes, its theory of disclosure risk. We make two main points. Firstly, the Five Safes theory may be viewed as a reparametrization of CI in the situation where the information flow is a statistical dissemination. Section 3 provides a brief overview of the CI theory and describes the translation between the Five Safe parameters (people, projects, settings, data and outputs) and the CI parameters (sender, recipient, subject, information type and transmission principles), where the latter set of parameters collectively determine the appropriateness, or privacy, of the information flow in question. Viewed this way, the Five Safes provides specialized guidance as to how NSOs can satisfy contextual information norms. In the reverse, by placing Five Safes within the CI theory, NSOs benefit from the extensive CI literature in justifying and understanding the Five Safes.

Secondly, as a framework for controlling the disclosure risk of statistical dissemination, the Five Safes provides a natural context to understand the advantages and limitations of formal privacy criteria. To make its case, this work focuses on differential privacy (DP), the predominant and state-of-the-art formal privacy criterion. As Section 4 explains, DP is a broad collection of technical standards which all measure in various ways how a statistical dissemination can depend on the response of a data provider - or, in other words, how a data provider can influence the data being shared. Importantly, we argue that DP accounts for some - but, crucially, not all - of the dimensions relevant for assessing the contextual integrity of statistical data sharing. We use the Five Safes, as a holistic risk assessment tool, to situate DP within the dimensions it partially measures (safe data and safe outputs) and to explicate the dimensions to which DP is agnostic. The contextualization of DP within the Five Safes is important for two reasons. Any implementation of DP requires choosing its various components (see Section 4). This choice depends on the broader context of the implementation, which the Five Safes explicate. Moreover, by placing DP within the Five Safes, NSOs can see how DP could be used to partially control safe data and safe outputs, and how DP can be traded-off against the other safes.

To recap, this work seeks to, firstly, situate the Five Safes framework within the broader concept of CI and, secondly, to contextualize DP via the Five Safes. This explains the dual meaning of this paper's title: the Five Safes is a context for narrow, technical notions of privacy and security, like DP; at the same time, it is also a specific context within the broader theory of CI.

2 The Five Safes and the information flows they govern

This section provides a brief review of the Five Safes framework, concentrating on the two information flows with which it is concerned. The major thesis of the Five Safes is that five dimensions of data access - people, projects, settings, data, and outputs - collectively characterize the disclosure risk in statistical dissemination. These dimensions are related yet conceptually independent from one another, in that a change alongside any of the safety dimensions within a data dissemination paradigm need not compel a change in other dimensions. In principle, the safety of each of these dimensions can be measured on a continuous scale. In the design of a statistical dissemination paradigm, a data custodian strives to ensure data confidentiality by promoting safety of these five dimensions. Doing so usually entails active monitoring and decision-making on the custodian's part, including vetting, supervision, and higher degrees of infrastructure security and compliance monitoring. Viewing these five dimensions under a joint framework allows a data custodian with a fixed amount of resources to focus on the safety of a subset of these dimensions, while maintaining control of the overall disclosure risk.

To explicate the meaning of the Five Safes, we begin with an examination of the two types of information flow that it is concerned with:

$$data \rightarrow people (researcher),$$
 (1)

outputs
$$\rightarrow$$
 people (general public). (2)

The two types of information flow are neither independent nor mutually exclusive to one another. Flow (1) is the process through which the researcher learns from the data in the possession of the data custodian. Typically, the researcher takes the initiative to access the data, conducts analyses based on the data, and publishes a set of scientifically significant findings. These published results, alongside any open information required to support the verification of these results, make up the outputs that reach the general public as captured in Flow (2). Alternatively, Flow (2) may occur when the data custodian directly shares information derived from their database with the general public, without involving researchers as an intermediary.

The term 'researcher' here refers to a person or an entity whose identity has been subject to some degree of vetting by the data custodian. This distinguishes a researcher from a member of the general public in our current discussion, even though in practice the two identities are not well separated. As such, a 'safe' researcher is someone who has demonstrated a good scientific standing as well as a commitment to data confidentiality and research ethics, including compliance with any stated requirements of privacy protection pertaining to the data that they use.

The 'safe projects' dimension concerns whether the intended use of the data is appropriate, ethical, and compliant with relevant legislation or regulations. The use of certain sensitive data may be restricted by law to support independent scientific research only [17, Section 12.3]. The data custodian would often also ascertain that their data is used toward the advancement of science, with clear and positive social benefits and in a manner consistent with modern scientific norms, including standards of reproducibility and

knowledge sharing. In addition, the safety of 'settings' refers to the security of the environment in which data access and sharing takes place, be they physical or virtual. Lastly, a 'safe output' is a disseminated statistical result that is sufficiently non-disclosive when judged against pertinent standards.

We illustrate how the two information flows interact under the Five Safes framework with three examples of data dissemination paradigms.

Example 1 (Public use data files for access by the general public and researchers alike. The agencies do not vet people who seek access because, by design, any person or entity without abusive intentions should be able to access the resource. A high level of scrutiny is placed on the data, which doubles as the output, to ensure that they are safe. On the other hand, since the data custodian cannot supervise the use of the data once it is made open access, no scrutiny is possible regarding the safety of the projects nor the settings in which the projects will be conducted.

Public use data files are frequently in the form of tabular data, which are highly aggregated from underlying microdata to ensure adequate confidentiality. Public use microdata exist too, but they are often heavily subsampled. The U.S. Census Bureau curates the Public Use Microdata Sample (PUMS) based on a small sample (1% and 5%) of responses from the American Community Survey [29]. The PUMS files are available on the Census Bureau's website and may be accessed via the file transfer protocol (FTP), a microdata analysis tool, or through an API provided by the Bureau.

EXAMPLE 2 (DATA ENCLAVES). Data enclaves are secure access environments through which authorized researchers can query the custodian's database. Data enclaves provide a highly secure setting for data access. Both the people and the projects seeking access are heavily vetted: only researchers who demonstrate legitimate scientific purposes of their inquiry and compliance with research ethics are allowed access. The outputs that the researcher is allowed to obtain and bring to outside of the data enclave is subject to various degrees of scrutiny. As a result, the data accessible through data enclaves can be detailed and comprehensive.

Data enclaves may be physical or virtual. A physical enclave is synonymous with a research data center (RDC), such as the Federal Statistical Research Data Center (FSRDC) of the U.S. Census Bureau and the Canadian Research Data Centre Network (CRDCN) of Statistics Canada. A virtual data enclave allows authorized researchers to access restricted-use data by logging into a secure, remote server. The DataLab of the Australian Bureau of Statistics (ABS) is an example of a virtual data enclave. The reader is referred to [4] for further illustrations of dissemination paradigms discussed in Examples 1 and 2 and an analysis of important safety considerations that pertain to them.

EXAMPLE 3 (SYNTHETIC DATA WITH VALIDATION SERVERS). The data custodian releases a synthetic dataset that resembles the underlying confidential dataset. Researchers who hold permission to access the synthetic dataset may use it to compose their desired statistical

¹Due to a lack of full oversight on the data access setting compared to physical data enclaves, statistical agencies debate the safety of virtual data enclaves [see e.g. 24].

analysis including its code implementation. Then, they may validate the results with the data custodian who will run the analysis on the restricted-use dataset, and release the results to the researcher if they are deemed safe.

A prominent case of Example 3 is the Survey of Income and Program Participation (SIPP) Synthetic Beta (SSB) [28]. The SSB is synthesized by the U.S. Census Bureau through integrating nine annual SIPP panels between 1984 and 2008, together with the W-2 records from the Social Security Administration (SSA) and Internal Revenue Service (IRS). Researchers whose proposed analysis is deemed appropriate and feasible by the Census Bureau are permitted to access the SSB. Prior to October 2022, access could be obtained via the Synthetic Data Server (SDS) hosted by Cornell University.² Once the researcher composes a functional and correct statistical analysis program, they submit the code to the Bureau, which in turn performs the validation on the researcher's behalf on the Gold Standard File (GSF) which is internal to the Bureau and confidential. The output is subject to a stringent level of disclosure review similar to those applicable to the FSRDCs.

Example 3 is an interesting mode of statistical dissemination, through which we see a juxtaposition of safety levels pertaining to the two information flows. On the level of Flow (2) where the relevant people are the general public, the outputs are strictly scrutinized according to a high safety standard, rendering this setting similar to the open data mode discussed in Example 1. On the level of Flow (1), data consists of two distinct components, the SSB and the GSF, where the former commands a level of safety higher than the latter due to its synthetic nature. The people, here referring to the researchers, are placed under a moderate level of scrutiny. The setting, the Cornell SDS, is effectively a virtual enclave. These elements render this setting analogous to the data enclave mode discussed in Example 2.

3 The Five Safes as a privacy context for statistical dissemination

Contextual integrity (CI) defines an information flow as private if it conforms with contextual informational norms [23]. There are five parameters that define contextual informational norms: the *sender*, the *recipient*, the *subject*, the *information type*, and the *transmission principles* [22]. To understand these parameters requires that we take a tailored approach to their explication by first situating this discussion in the context of statistical dissemination.

CLAIM 1. In the context of statistical dissemination, the Five Safes is an instantiation of a set of informational norms that govern privacy protection.

CLAIM 2. The Five Safes can be viewed as a (faithful albeit imperfect) reparametrization of CI, when the information flow in question is a statistical dissemination.

Table 1 outlines in shorthand the meanings of the privacy contextual informational norm parameters as they apply to statistical dissemination. These meanings are explicated with reference to the elements of the Five Safes framework.

Of the five CI parameters, the first two are straightforwardly understood. The 'sender' is the entity sending the information, that is, the data custodian or the NSO initiating the statistical dissemination. The 'recipient' is the entity receiving the information, here being either the researcher or the general public. The latter three parameters require more explication. Viewed through the lens of the Five Safes, the notions of subject and information type are interrelated. Subject refers to the individuals about whom the potentially private information concerns. These are the data contributors: often persons, sometimes businesses, and can possibly be other entities. Information type refers to the nature of information involved in the transmission: demographics, medical history, financial records, to name a few possibilities. For the purpose of statistical dissemination, subject is a component of 'data' because data subjects are both the source of the information that the custodian collects and the target entity that the collected information describes. On the other hand, information type is a component of both 'data' and 'outputs.' It is a component of 'data' because the nature of the information affects both the structure and the substance of the data. It is also a determinant of 'outputs' because the custodian's decision about what to disseminate to their recipients turns on the information type.

Among all the CI parameters, the *transmission principle* is the most complex to analyze. This parameter captures the constraints placed upon the other dimensions of informational norms, but is treated as a separate dimension due the distinct theoretical utility of doing so [22]. For statistical dissemination, transmission principles encompass multiple dimensions of the Five Safes, including (though are not limited to) 'projects' and 'settings.' As alluded to in Section 2, to assess whether a project is safe is to ask whether the dissemination, whatever form it may take, is used for an appropriate purpose, be it to support scientific research, to inform the public, to enable evidence-based policy making, or to ensure compliance with legal and regulatory mandates. The intended use of the dissemination delineates the scope of the project, and in this capacity, the purpose defines the project and in turn determines the appropriateness of the chosen transmission principle.

Similarly, the safety of settings is concerned with the type of data access that the sender offers to the recipient. Can the setting enable permissible access by legitimate recipients, while keeping out impermissible access by unauthorized parties? The answer to these questions most straightforwardly depends on the mode of transmission. Is the transmission accomplished via physical or virtual data enclaves (Example 2)? Via synthetic data and validation (Example 3)? Or is the data or output openly shared for public access (Example 1)? It would nonetheless be a mistake to confine the reach of transmission principles to the mode of data access. Whether a statistical dissemination ultimately constitutes an appropriate flow of information as specified by the transmission principles also turns on other factors, such as the nature of the data, the identities of the subject, and the intended use by the recipient, some of which have been discussed previously. The transmission principles may also entail aspects that are not easily captured within the five safety dimensions. An example may be the distinct reason or authority by which the custodian collects certain data or disseminates certain outputs. Take for example the U.S. Census Bureau and its constitutional mandate to enumerate the population

 $^{^2\}mathrm{Unfortunately},$ the Cornell server was shut down on September 30, 2022.

every ten years for the purpose of apportionment. Every person living in the U.S. on Census day has an obligation to respond to the Decennial Census. The public dissemination of state population totals is exempt from any statistical disclosure control protection, even though exact statistics may carry increased risk of disclosure. The Five Safes framework may have to acknowledge the dissemination of state population totals as a case in its own right, but the apparently risky dissemination is entirely appropriate when viewed under the pertinent transmission principle, namely, the Bureau's charge to enable a constitutionally sanctioned political process via quantitative evidence.

Transmission principles are recognized as a malleable dimension of informational norm in the CI literature. The malleability results in a dense mapping from the CI parameter to the five safety dimensions. At the same time, however, transmission principles capture nuanced distinctions between appropriate and inappropriate transmissions as well as in novel circumstances where both guidance and intuition may be lacking as to whether a statistical dissemination can be deemed 'safe' within the meaning of the Five Safes. The possibility of reasoning via the reparametrized CI dimensions is precisely the value for situating the Five Safes within the CI context.

4 Differential privacy in the context of the Five Safes

Differential privacy is a state-of-the-art technical formulation of privacy associated with statistical data dissemination. It has been adopted by data agencies and intermediaries. Since the proposal of ε -DP (or pure DP) [16], a multiplicity of flavors of differential privacy has emerged, including probabilistic DP [15], approximate DP [21], zero-Concentrated DP [11], f-DP [14], to name a few. There are bounded versus unbounded versions of DP to suit the scenarios with known versus unknown dataset sizes. The TopDown Algorithm [1], the U.S. Census Bureau's differentially private disclosure avoidance system (DAS) for the 2020 Decennial Census Redistricting Data (P.L. 94-171) Summary and Demographic and Housing Characteristics Files, introduced the concept of invariants, which are exact statistics of the confidential data that are nevertheless released without any privacy protection. The state population totals discussed in Section 3, which are disseminated exactly as enumerated, are one type of invariants.

To gain conceptual clarity amidst the plurality of choices for DP definitions, we employ the unified construction proposed by [7–9], which explicitly spells out the necessary components of a *differential privacy specification*, some of which are often overlooked:

- Who is eligible for protection, as defined by the domain X, the set of all possible datasets;
- What is the granularity of protection, as conceptualized the input "distance" d_X(x, x') between any two datasets x and x';
- How to measure protection, as captured by the output "distance" d_{Pr}(P⁽¹⁾, P⁽²⁾) between any two probability distributions P⁽¹⁾ and P⁽²⁾;

• *How much* protection is afforded, as quantified by the protection loss budget $\varepsilon_{\mathcal{D}}$ for each universe $\mathcal{D} \in \mathcal{D}$.

A data release mechanism can be most generally defined as differentially private as follows.

Definition 4.1 (Bailie et al. [7]). A data-release mechanism T satisfies a differential privacy specification $(X, \mathcal{D}, d_X, d_T, \varepsilon_D)$ if

$$d_{\Pr}[P_{\mathbf{x}}(T \in \cdot), P_{\mathbf{x}'}(T \in \cdot)] \le \varepsilon_{\mathcal{D}} d_{\mathcal{X}}(\mathbf{x}, \mathbf{x}'), \tag{3}$$

for all \mathbf{x}, \mathbf{x}' in every data universe $\mathcal{D} \in \mathcal{D}$. Here $\mathsf{P}_{\mathbf{x}}(T \in \cdot)$ denotes the probability distribution of the data release mechanism output $T(\mathbf{x})$ upon given input data \mathbf{x} .

In this work we will not explicate Definition 4.1 in further detail, other than remark that it illustrates how the aforementioned elements of a differential privacy specification come together. Notably, the first four components -X, \mathcal{D} , d_X and d_{Pr} – confer a quantitative description of the privacy guarantee (its *flavor*), whereas the protection loss budget $\varepsilon_{\mathcal{D}}$ serves as a quantitative measurement (its *strength*).

In statistical dissemination, the Five Safes delineate a context in which differential privacy can be understood. What we mean by this is the following:

CLAIM 3. Differential privacy can be viewed as a quantitative standard of safety pertaining to aspects of the outputs and the data components of the Five Safes.

A quantified measurement of safety levels for aspects of the outputs and the data is helpful in the Five Safes framework, because it allows for the modulation of the various elements that collectively contribute to disclosure risk. The modulation may be achieved in multiple ways, two of which we discuss here.

First, differential privacy acts as a "screen" between the data and the people who access the data. By construction, differential privacy constrains specific probabilistic properties of an output by a certified mechanism. This may be employed by the statistical agency to compose open data as well as to restrict researcher release from data enclaves or validation servers [6]. Differential privacy may also directly restrict how the researcher may interact with the data in the first place. One example is differentially private synthetic data [see e.g. 10]. In the case of local privacy, the measurements taken from individual data contributors are privatized (e.g. perturbed or infused with noise) as soon as they leave the end device prior to arriving at the data custodian. In each of these cases, the differential privacy guarantee ascertains, in a mathematically rigorous way, a level of difficulty for adversarial agents to deduce the value of the confidential data. This enables, at least in a heuristic way, less scrutiny to be placed on the people.

Second, differential privacy enables precise privacy accounting by statistical agencies. For privacy definitions of the same *flavor* (i.e. the same choices for X, \mathcal{D} , d_X and d_{Pr}), the protection loss budgets of two mechanisms may be *composed* to yield an overall, combined budget.³ What this means is that the data custodian with a fixed amount of privacy loss budget may choose to divide the budget across a number of *projects*, evenly or unevenly according to their

³Composition is not possible within approaches to privacy that are not formal. It may well be the case that the combined disclosure risk of two data products is infinite, even though either carries a finite risk.

Informational norm parameters	Their meanings in statistical dissemination
sender	statistical agencies/NSOs/data custodians
recipient	people : researchers (1) and general public (2)
subject	is a component of data (1)
information type	is a component of data (1) and outputs (2)
transmission principles	encompass projects , settings , and more

Table 1: The privacy contextual informational norm parameters and their meanings in statistical dissemination, with a reference mapping to the Five Safes (in bold).

significance, modulating the quantity-quality tradeoff in ways that the custodian sees fit.

We also observe two limitations of differential privacy as a quantitative standard of safety.

REMARK 1. Differential privacy is silent on the safety of certain aspects of the outputs and the data.

Differential privacy specifies the flavor, as it does the strength, of the privacy protection. It is, however, agnostic to the *nature* of the data at hand. A differential privacy mechanism will treat two datasets identically so long as they possess identical mathematical structures, even though one may be highly sensitive in nature (e.g. records of patients suffering from a socially stereotyped disease) while the other is not (e.g. a log of dairy preferences of customers at a coffee shop). Therefore, differential privacy should not be taken as a comprehensive quantification for the safety of the outputs and the data.

Remark 2. Differential privacy does not purport an assessment of safety for people, projects, or settings.

Differential privacy is a property of the output rather than the process through which it is generated. As such, it is agnostic to the settings in which privatization and data sharing takes place. Indeed, one of the celebrated feature of differential privacy is that the privacy mechanism can be entirely transparent without sabotaging the privacy guarantee.

For similar reasons, differential privacy does not directly measure the safety of the people and the projects. When used as a standard to quantify the safety of aspects of the data and the outputs, however, it may serve as an indirect guidance on the safety tuning for the people and for the projects. In fact, such tuning often constitutes a balancing act between privacy and *utility*.

5 Discussion

Disclosure control in statistical dissemination poses a "wicked problem" [3]. Two defining characteristics of a wicked problem are that 1) the stakeholders have conflicting interests in the situation, and 2) the problem is so complex that there may not be consensus as to what the problem actually is, let alone a solution to it. The oft-stated mantra of "privacy-utility tradeoff" bears witness to both issues. On the one hand, data custodians have a dual mandate to procure high quality data for its users while protecting the privacy of data contributors. The data users' demand for finer and more accurate information comes into direct conflict with the data contributors' expectation for confidentiality and anonymity. On the other hand,

the fact that the tradeoff remained a mantra demonstrates how successful it has resisted attempts to operationalize it into a set of workable guidance devoid of human judgment, not one that commands wide agreement anyway.

The Five Safes framework makes a valuable contribution by articulating, and hence clarifying, an otherwise intractable problem. It reduces the "wickedness" of disclosure control decisions by modularizing a statistical dissemination into five manageable dimensions, along which an NSO may assess their risk tolerance and utility preference, subject to the pragmatic constraints they face.

It is one thing to articulate a question, and a whole other to devise a solution, though. A data custodian looking to the Five Safes and the Five Safes only for an answer to their disclosure control question may be disappointed. The five safety dimensions don't come equipped with yardsticks that provide quantitative measures of safety for a given dissemination regime. As skeptics of the Five Safes point out (see e.g. [12]), they provide qualitative narratives that may be inherently unfit as a quantitative standard.

All is not lost in our view. The Five Safes have demonstrated its value in instructing the NSOs' decision processes by placing, as discrete and enumerated case studies, the existing major statistical dissemination paradigms such as those reviewed in Section 2 relative to one another along the safety dimensions. The framework is best viewed as a set of guiding principles to reason with the potential factors that affect the disclosure risks of particular statistical dissemination choices. We may easily concede that these efforts in no way provide a total ordering for all conceivable dissemination choices, nor a watertight guarantee of the absence of counterexamples - such are the distinct advantages of technical criteria of privacy by virtue of their mathematical rigor. We believe, however, and as our analysis in Section 4 shows, that while narrow constructions of the disclosure control problem afford the apparent comfort with concrete and precise answers, it sidesteps the heart of the challenge that is the fostering of a consensus about what the disclosure control problem really is. To this end, we reiterate our viewpoint that the CI theory, as a reparametrization of the Five Safes, provides a distinct conceptualization of the privacy-utility tradeoff for statistical dissemination, hence a viable alternative angle for analysis.

Before closing, we briefly discuss three important questions that remain unanswered in this work. First, we view the Five Safes as a reparametrization of contextual integrity when the information flow in question is statistical dissemination. As Section 3 discusses, this reparametrization is faithful but imperfect, in the sense that the mappings between the two frameworks can seem either narrow or capacious. In future work, we aim to supplant the meanings that are lost in this translation, in particular by expanding the discussion in Section 3 to detail aspects of data and outputs (from the Five Safes) that go beyond the subject and the information type (from the privacy norms), as well as aspects of transmission principles that go beyond the dimensions of the Five Safes.

Second, as we have already argued, differential privacy offers a strong technical notion for assessing certain aspects of the 'safe data' and 'safe output' dimensions. Does there exist more comprehensive technical notions to these dimensions, and does there exist technical notions for the other dimensions of Five Safes? We surmise that a pursuit towards technicalization may not make sense universally. After all, some of the safety dimensions (such as safe people) may be too complex to allow mathematical tractability. However, others – such as safe settings – may benefit from advances in fields such as information security.

Third, we look to further explore the interaction between the Five Safes and the legal frameworks governing the operations of the NSOs. For starters we ask: what configurations of the Five Safes best correspond to the current legal framework for a specific NSO? And how can the Five Safes be used to update the legal framework in the future?

Acknowledgments

An earlier version of this paper was presented at the 5th Annual Symposium on Applications of Contextual Integrity on September 22, 2023 in Toronto, Canada. We thank John Eltinge and Jeremy Seeman for helpful feedback on that version. Nevertheless, all errors and shortcomings in this work are solely attributable to ourselves. JB gratefully acknowledges partial financial support during this project from the Australian-American Fulbright Commission and the Kinghorn Foundation.

References

- [1] John Abowd, Robert Ashmead, Ryan Cumings-Menon, Simson Garfinkel, Micah Heineck, Christine Heiss, Robert Johns, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, Brett Moran, William Sexton, Matthew Spence, and Pavel Zhuravlev. 2022. The 2020 Census Disclosure Avoidance System TopDown Algorithm. Harvard Data Science Review Special Issue 2 (2022). doi:10.1162/ 99608f92.529e3cb9
- [2] Advisory Committee on Data for Evidence Building. 2022. Year 2 Report. Technical Report.
- [3] John Alford and Brian W Head. 2017. Wicked and less wicked problems: a typology and a contingency framework. *Policy and society* 36, 3 (2017), 397–413.
- [4] Australian Bureau of Statistics. 2021. Five Safes Framework. https://www.abs.gov.au/about/data-services/data-confidentiality-guide/five-safes-framework
- [5] James Bailie. 2020. Big Data, Differential Privacy and National Statistical Organisations. Statistical Journal of the IAOS 36, 4 (2020), 1067–1074. doi:10.3233/SJI-200685
- [6] James Bailie and Chien-Hung Chien. 2019. ABS Perturbation Methodology Through the Lens of Differential Privacy. In Work Session on Statistical Data Confidentiality, UN Economic Commission for Europe. 13. https://jameshbailie. github.io/Papers/papers/UNECE2019.pdf
- [7] James Bailie, Ruobin Gong, and Xiao-Li Meng. 2025+. A Refreshment Stirred, Not Shaken (I): Five Building Blocks of Differential Privacy. In preparation (2025+).
- [8] James Bailie, Ruobin Gong, and Xiao-Li Meng. 2025. A Refreshment Stirred, Not Shaken (II): Invariant-preserving Deployments of Differential Privacy for the US Decennial Census. arXiv:2501.08449 [cs] doi:10.48550/arXiv.2501.08449
- [9] James Bailie, Ruobin Gong, and Xiao-Li Meng. 2025. A Refreshment Stirred, Not Shaken (III): Can Swapping Be Differentially Private? arXiv:2504.15246 [cs] doi:10.48550/arXiv:2504.15246

- [10] Claire McKay Bowen, Victoria Bryant, Leonard Burman, John Czajka, Surachai Khitatrakun, Graham MacDonald, Robert McClelland, Livia Mucciolo, Madeline Pickens, Kyle Ueyama, et al. 2022. Synthetic Individual Income Tax Data: Methodology, Utility, and Privacy Implications. In Privacy in Statistical Databases: International Conference, PSD 2022, Paris, France, September 21–23, 2022, Proceedings. Springer, 191–204.
- [11] Mark Bun and Thomas Steinke. 2016. Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds. In Theory of Cryptography (Lecture Notes in Computer Science), Martin Hirt and Adam Smith (Eds.). Springer, Berlin, Heidelberg, 635–658. doi:10.1007/978-3-662-53641-4_24
- [12] Chris Culnane, Benjamin I. P. Rubinstein, and David Watts. 2020. Not Fit for Purpose: A Critical Analysis of the 'Five Safes'. arXiv:2011.02142 [cs] doi:10. 48550/arXiv.2011.02142
- [13] Tanvi Desai, Felix Ritchie, and Richard Welpton. 2016. Five Safes: Designing Data Access for Research. Working Paper 1601. University of the West of England, Bristol. 27 pages.
- [14] Jinshuo Dong, Aaron Roth, and Weijie J Su. 2022. Gaussian differential privacy. Journal of the Royal Statistical Society Series B 84, 1 (2022), 3–37.
- [15] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. 2006. Our Data, Ourselves: Privacy Via Distributed Noise Generation. In Advances in Cryptology - EUROCRYPT 2006 (Lecture Notes in Computer Science), Serge Vaudenay (Ed.). Springer, Berlin, Heidelberg, 486–503. doi:10.1007/11761679 29
- [16] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*. Springer, 265–284.
- [17] Ian Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter, and Julia Lane (Eds.). 2021. Big Data and Social Science: Data Science Methods and Tools for Research and Practice (second edition ed.). CRC Press, Boca Raton, FL.
- [18] Elizabeth Green and Felix Ritchie. 2023. The Present and Future of the Five Safes Framework. Journal of Privacy and Confidentiality 13, 2 (2023). doi:10.29012/jpc. 831
- [19] Craig Jones, Anna McDowell, Vince Galvin, and Dorothy Adams. 2022. Building on Aotearoa New Zealand's Integrated Data Infrastructure. Harvard Data Science Review (2022). doi:10.1162/99608f92.d203ae45
- [20] Lateral Economics. 2019. Value of the Australian Census. Technical Report. Australian Bureau of Statistics. 63 pages.
- [21] Ashwin Machanavajjhala, Daniel Kifer, John M Abowd, Johannes Gehrke, and Lars Vilhuber. 2008. Privacy: Theory meets practice on the map. In Proceedings of the 2008 IEEE 24th International Conference on Data Engineering. IEEE Computer Society, 277–286.
- [22] Helen Nissenbaum. 2019. Contextual Integrity Up and Down the Data Food Chain. Theoretical Inquiries in Law 20, 1 (2019), 221–256.
- [23] Helen Fay Nissenbaum. 2010. Privacy in Context: Technology, Policy, and the Integrity of Social Life. Stanford Law Books, Stanford, Calif.
- [24] J. Neil Russell. 2022. National Center for Health Statistics Virtual Data Enclave (VDE) Project Overview. https://www.cdc.gov/nchs/data/bsc/bsc-pres-j-neil-russell-5-26-2022.pdf.
- [25] Statistics New Zealand. 2022. How We Keep Integrated Data Safe. https://www.stats.govt.nz/integrated-data/how-we-keep-integrated-data-safe/.
- [26] Peter Stokes. 2017. The 'Five Safes' Data Privacy at ONS. https://blog.ons.gov.uk/2017/01/27/the-five-safes-data-privacy-at-ons/.
- [27] UK Data Service. [n.d.]. What Is the Five Safes Framework? https://ukdataservice.ac.uk/help/secure-lab/what-is-the-five-safes-framework/.
- [28] U.S. Census Bureau. 2022. Synthetic SIPP Data. https://www.census.gov/programs-surveys/sipp/guidance/sipp-synthetic-beta-data-product.html.
- [29] U.S. Census Bureau. 2023. American Community Survey 2017-2021 5-Year PUMS User Guide and Overview. https://www2.census.gov/programs-surveys/acs/ tech_docs/pums/2017_2021ACS_PUMS_User_Guide.pdf.