Harvard Data Science Review • Issue 2.1, Winter 2020

Differential Privacy and Social Science: An Urgent Puzzle

Daniel L. Oberski[1], **Frauke Kreuter**[2]

[1]Associate Professor, Department of Methodology & Statistics, Utrecht University,
[2]Director, Joint Program in Survey Methodology, University of Maryland, College Park

Published on: Sep 29, 2020

DOI: 10.1162/99608f92.63a22079

ABSTRACT

Accessing and combining large amounts of data is important for quantitative social scientists, but increasing amounts of data also increase privacy risks. To mitigate these risks, important players in official statistics, academia, and business see a solution in the concept of differential privacy. In this opinion piece, we ask how differential privacy can benefit from social-scientific insights, and, conversely, how differential privacy is likely to transform social science. First, we put differential privacy in the larger context of social science. We argue that the discussion on implementing differential privacy has been clouded by incompatible subjective beliefs about risk, each perspective having merit for different data types. Moreover, we point out existing social-scientific insights that suggest limitations to the premises of differential privacy as a data protection approach. Second, we examine the likely consequences for social science if differential privacy is widely implemented. Clearly, workflows must change, and common social science data collection will become more costly. However, in addition to data protection, differential privacy may bring other positive side effects. These could solve some issues social scientists currently struggle with, such as *p*-hacking, data peeking, or overfitting; after all, differential privacy is basically a robust method to analyze data. We conclude that, in the discussion around privacy risks and data protection, a large number of disciplines must band together to solve this urgent puzzle of our time, including social science, computer science, ethics, law, and statistics, as well as public and private policy.

Keywords: differential privacy, social science, data science, open data, robustness; data protection, confidentiality, GDPR

1. Real and Present Danger

How can we analyze data about people without harming the privacy of the individuals being analyzed? When Swedish statistician Tore Dalenius set out to solve this puzzle in the 1970s, many considered it among the least interesting topics within the already less thought after discipline of official statistics. Dalenius saw it differently. He and others, including Ivan Fellegi, a Hungarian immigrant who would later become Canada's chief statistician, saw a problem that would not just go away by itself. As Fellegi, who experienced government repression first-hand during the Hungarian uprising, wrote in 1972: "the concern is real and the danger is also real" (Fellegi, 1972, p. 8).

Today, Felligi's concerns are reanimated by journalists, privacy activists, academics, and lawmakers everywhere. Do handy medical apps endanger my insurance policy or ability to get a mortgage? Can I be harmed by my census responses? The news media now covers privacy breaches as intensely as

political scandals; digital rights foundations such as the Electronic Frontier Foundation warn against invasions by 'big data'; governments adopt new and far-reaching privacy laws such as the EU's General Data Protection Regulation 2016/679 (GDPR); and large tech firms started to use built-in privacy protections to market their products.

Meanwhile, the statistical techniques that Dalenius helped develop in his long and fruitful career have grown into their own. Statistical Disclosure Limitation, long an obscure specialization within official statistics agencies and not part of regular statistics curricula, is now an active field covering statistics, computer science, and a host of other disciplines. One concept is particularly poised to completely change the way we analyze data about people: 'differential privacy.'

2. Differential Privacy

Differential privacy is a simple mathematical definition that indicates when publishing results or data sets can be considered 'private' in a specific sense. The term, its definition, and many of the modern techniques associated with it, were invented by theoretical computer scientists Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith (see Dwork & Roth, 2014, for key references). These researchers took a step back from the field initiated by Dalenius and Fellegi and rebuilt its foundations on a rigorous definition that could be used to protect data.

2.1 How Differential Privacy Protects Data

Suppose you are asked to participate in a survey on traffic violations. One factor in deciding whether or not you will answer the survey, apart from the usual issues of enjoyment, time constraints, and so forth, might be whether participating in the survey will negatively affect you in some way. For example, you would be affected negatively if you admitted to a traffic violation in a survey and promptly received a fine and a criminal record. To prevent this, confidentiality pledges in research studies state that no personal information will be released, meaning 'personally identifying information' (PII) will be kept separate from the data (the survey answers), and those analyzing the data or getting access to the data will not know who provided the answers. Some confidentiality statements go further and say that information will only be released in the aggregate (e.g., summary statistics on the group level). *From the perspective of differential privacy, neither of the two confidentiality pledges statements is sufficient, and releasing the aggregate is as much a breach of differential privacy as releasing the microdata.*

Removing PII is insufficient because we are now living in a world filled with data. The worry is that some external data might include the PII plus the exact same combination of variables as the researchers' data (e.g., Sweeney, 2002) . In this situation a unique value combination on the string of variables can be used to link the PII back to the remaining data collected by the researcher. A prominent example in the private sector reflecting this problem was the release of the Netflix rental

data on Kaggle: at first this data set seemed anonymized, but on closer inspection it turned out at least two participants could be identified through probabilistic data linkage (Narayanan & Shmatikov, 2008).

Often overlooked is the fact that access to aggregate statistics can also leak information about individuals. For example, if many tables are produced in a way that the combination of the tables reveal information about one person with a specific set of characteristics. If that set of characteristics is known through an outside source, the missing piece of information (traffic violation) can be learned.

Furthermore, privacy researchers worry about the individual data being recreated through database reconstruction. Database reconstruction (Dinur & Nissim, 2003) means that attackers can guess the original data that produced a given set of aggregate statistics. It works because an attacker can consider all hypothetical data sets, and then weed out those data sets that could never have produced the reported aggregate statistics. For example, a thousand-person data set with Bill Gates in it (net worth $107 billion) could never produce an average net worth below $107,000. By considering different aggregates reported from the same data set at the same time, an attacker can, like Sherlock Holmes, rule out the impossible and arrive at the original data set. The only way to prevent database reconstruction is to scramble any reported output from the original data with noise. This added noise prevents the Sherlock Holmes procedure from weeding out impossible data sets, because it leaves uncertainty about what could have produced the (now error-prone) aggregate. If reconstruction of the original data is prevented, so is the identification of people through any kind of linkage to existing or future data.

Statistics or other reported outputs with injected noise are called 'differentially private' if the inclusion or exclusion of the most at-risk person *in the population* does not change the probability of any output by more than a given factor. The parameter driving this factor, referred to as epsilon (ϵ), quantifies how sensitive the aggregate output is to any one person's data. If it is low, the output is highly 'private' in the sense that it will be very difficult to reconstruct anything based on it. If it is high, reconstruction is easy.

Here, outputs can be aggregate statistics, but also synthetic microdata sets based on the original data. Any output can be made differentially private by subjecting it to a randomization 'mechanism,' such as adding random noise or drawing discrete values from a probability distribution. This means that there is no 'one and only' differential privacy algorithm—rather, multiple algorithms can satisfy the definition.

We give a basic version of the differential privacy definition. Consider two possible databases that differ by only a single record (i.e., person), denoted $\mathrm{database}$ and $\mathrm{database}'$, to emphasize that the two are *almost* identical. Also consider a randomization mechanism that generates a *distribution* of

possible aggregate outputs we can produce from the two databases. For example, the mechanism could generate the output by calculating an average on the database, then adding a draw from a random variable to it. Because the mechanism takes a database as input and generates a *distribution* of outputs, we can think about the probability[1] that the mechanism generates some specific outcome. The mechanism is said to be 'differentially private' if:

$$\frac{\Pr(\text{mechanism}\,(\text{database}) = \text{output})}{\Pr(\text{mechanism}\left(\text{database}^{'}\right) = \text{output})} \leq e^{\epsilon},$$

for all databases, all records (in the *population*) and all possible outputs. In other words, the probability of any output should not change by more than a factor depending on epsilon (ϵ) if the databases differ by just one record. For small ϵ, this factor is about $100 \times \epsilon$ percent (because $e^{\epsilon} \approx 1 + \epsilon$). For example, if ϵ is chosen to equal $\epsilon = 0.1$, then the mechanism is 'epsilon-differentially private' if no person's addition or removal from any possible database can change the probability of any output by more than about 10%. The factor e^{ϵ} can also be seen as a Bayes factor for the hypothesis that a specific person is in the database.

The number epsilon ($\epsilon > 0$) must be determined by the person releasing the outputs to the public, and can be tuned through the randomization mechanism. Adding more noise will even out the probability of all outputs, regardless of the underlying data, and therefore make differential privacy more easily attainable. In the extreme, the mechanism is nothing but noise and therefore all outputs are completely impervious to any database changes, making $\epsilon = 0$. This is a different way of recognizing that more noise generates more privacy (smaller ϵ). It also reveals that differential privacy is very closely related to *robustness*. Hansen (2019) gives a practical illustration of differential privacy using a simple mechanism, and Wood et al. (2018) give a conceptual introduction for a non-technical audience. Technical details can be found in the monograph by Dwork & Roth (2014).

2.2 The Privacy–Utility Tradeoff

Now there is one problem. More noise injection means more privacy, but it also means that the published data and results will likely be further from what was found in the original, 'raw' data. Protecting data, with differential privacy or any other means, therefore inevitably reduces its utility. Whether that tradeoff is a good deal will depend on what the data are, how they are used, and for what purposes.

This privacy–utility tradeoff is not new: national statistical institutes have been injecting random noise into published data since Dalenius and Fellegi. Typically, they have protected data with noise by imagining what 'sensitive' values a nefarious attacker might want to glean, and which attack methods

and databases he or she might reasonably use (Elliot et al., 2018; Hundepool et al., 2012). But what is sensitive depends on context, and in the age of digitalization and cheap computing, what is unreasonable today is in your pocket tomorrow. Differential privacy, in contrast with classical approaches, aims to give guarantees for *all* variables and *all* data sets, past and future, including those obtained through data linkage. Where classical approaches have studied *plausible* attacks on privacy, differential privacy studies the *worst possible case*. Its scope is more comprehensive—and more impactful.

The U.S. Census Bureau, for its part, has decided the benefits of guaranteeing high levels of differential privacy outweigh the costs, and recently announced that results from the 2020 census will be published using differentially private mechanisms (Abowd, 2018). The Census Bureau's decision has made many ears prick up, as it has huge potential implications for all traditional uses of the census, including redistricting, subsidies, and some economic analyses (Mervis, 2019), as well as implications for social science research relying on census data and its derivatives.

But the discussion extends well beyond the U.S. Census. There is currently an extreme pressure to increase data protection—as exemplified by new legislation throughout the world's democracies, such as the 2018 European GDPR (European Parliament and Council, 2018), the 2017 Japanese Amended Act on the Protection of Personal Information (AAPI 2016)[2], the 2020 Brazilian General Data Protection Law (LGDP 13.709/2018)[3], the 2020 California Consumer Privacy Act (375/2018)[4], or the 2019 New York SHIELD act (S5575B/2019), or the proposed Personal Data Protection Bill (PDP Bill 2019)[5] in India. Many important players in official statistics, academia, and business see differential privacy as the solution (Abowd & Schmutte, 2018; Mervis, 2019). What's more, differential privacy may be attractive for other reasons than privacy. As we discuss further below, because it prevents too much information from 'leaking out' of the original data set, differential privacy can be leveraged as a device to prevent overfitting such as *p*-hacking (see Dwork et al., 2015). On balance, it appears likely that much human data that are publicly available now will, in the future, be published using differentially private mechanisms—or not at all.

3. Social Science Perspective on Differential Privacy

Differential privacy certainly has its advantages and disadvantages. It guarantees that database reconstruction will be very difficult. The noise mechanism can be safely shared without endangering privacy. And knowing about the amount and type of noise used sometimes allows statisticians to adjust their subsequent data analysis procedures to account for these known sources of random error (e.g., Abowd & Schmutte, 2018).

3.1 The Current Controversy

The main disadvantage of ensuring differential privacy is that it typically requires more noise infusion than traditional techniques. This is a consequence of the fact that traditional techniques only need to prevent linkage, while differential privacy prevents linkage *through* reconstruction. One might expect that in the discussion on how and when differential privacy should be applied, level-headed experts convene to weigh such pros and cons and find a consensus.

But where we might expect a dry weighing of facts, we observe a heated debate, which shows no signs of abating (see Ruggles, Fitch, Magnuson, & Schroeder, 2019; and the 2019 Harvard Data Science Initiative Conference. Why do ordinarily equanimous researchers embroil themselves in such a raging controversy? And, most puzzling, why hasn't evidence and scientific argument been able to adjudicate this apparently scientific disagreement yet?

We believe scientific facts have not been able to end the disagreement, because the disagreement is not about facts. Rather, the parties have different subjective beliefs about risk, and therefore differ in their ideas on how to mitigate such risks. Furthermore, the current debate often overlooks the social science behind data collection and privacy perceptions.

So why are proponents of traditional statistical disclosure limitation (SDL) and differential privacy at odds? At their core, SDL and differential privacy share the same aim: to prevent identification. But they differ in their assessment of the risks of linkage attacks. SDL, as the name implies, tries to limit these risks, and then assesses them using *currently available evidence*. The differential privacy literature, on the other hand, points out that *currently available evidence might be insufficient*, because it is always possible that future data sets and computing power will pose new, currently unforeseeable, risks. Consequently, one should assume that the probability of a linkage attack is 100% and the harm substantial. For this reason, differential privacy focuses on preventing database reconstruction. In other words, SDL postulates that the risk of identification through linkage can be controlled and limited, whereas differential privacy postulates that it cannot. This disagreement cannot be assessed with evidence, because it is exactly the unobserved parts of (future) risk on which beliefs differ.

3.2 Different Protections in Different Circumstances

Our own belief is that different situations will warrant different assumptions—as well as practical concerns—and therefore different approaches. We suggest a way forward by identifying three different types of data of use to social science, which we argue require three different approaches to data protection.

The first type concerns data that are currently widely available as results of small-scale experiments or social surveys. Such data may be available publicly at no cost, as is the case, for example, with well-known publicly funded studies such as the European Social Survey or World Values Survey. For this type of data (comparatively small random samples drawn from the population), our judgement falls on the side of the SDL worldview: the risks, including both the probability of exposure and the potential for harm, appear to have been sensibly assessed. The utility of such data is exactly in their public availability and broad potential to study varying topics, from social inequality to demographics and health. It seems unwise to advocate implementing differential privacy for future releases of data in this category because much of its scientific utility would be diminished at relatively little public benefit.

The second category of data exhibits a clear potential for privacy risks but is available in 'data enclaves.' Examples for such data are confidential microdata from administrative or other processes. Trusted researchers can access the data in a secure computing environment after passing an (often) impressive number of hurdles, including binding agreements to safeguard the participants' confidentiality. Such agreements can also include manual screening of outputs produced from the data for data protection purposes. This is the approach taken by many biobanks containing sensitive genetic data, such as the UK biobank or Estonian biobank, as well as recent initiatives in social science such as the Coleridge Initiative in the United States, and ODISSEI and the System of Social Statistical Datasets at Statistics Netherlands (Bakker, Van Rooijen, & Van Toor, 2014, pp. 418–419).

In this second category, it may be possible to liberate some of the information that is currently—rightly—under lock and key by imposing differential privacy guarantees on the outputs produced from the data. A difficulty with this approach is that the choice of output itself could potentially reveal sensitive information; Dwork & Ullman (2018) discuss some potential solutions to this conundrum. Allowing trusted researchers full data access while controlling the publicly produced outcomes could allow more rapid discovery of important medical correlations, for example, by removing red tape. But that is only useful if the original, raw data remain equally accessible, curated, and protected. Differential privacy can therefore be beneficial here, but only if it happens while also preserving the integrity of the data enclave approach. This approach would require additional funding.

Third and finally: data that are generated in large 'Google-sized' quantities as part of our interactions with (mostly) private sector platforms and devices. These are the data that are just too sensitive to share in any form at current, sometimes due to privacy but often equally due to competition: geolocations, purchases, ad viewing, clicking, and many other human behaviors. Moreover, because these data were not generated for research, they were collected without explicit consent by the data subjects for research purposes, and should be held to higher privacy standards. While several companies that collect such data have provided them in limited form to the research community in the

past, most of this provision has been discontinued amid privacy concerns (Bruns, 2019). Although the raw data will remain out of most researchers' reach for the foreseeable future, by applying the principles of differential privacy, much of the useful social-scientific information in them can be rescued. For example, Social Science One aims to use differential privacy to allow researchers access to Facebook data (King & Persily, 2019). Recent developments suggest that many of the challenges we have outlined above are also encountered in this project (Alba, 2019).

3.3 Privacy—The Multidisciplinary Responsibility

In addition to arguing we should employ different standards of risk for different datatypes, we argue that the current debate can and should benefit from insights from the social sciences.

Privacy is a social issue. It involves norms, expectations, trust, understanding, and relationships. Therefore, we see an important role for social scientists themselves in the science of privacy. Differential privacy guarantees that participants are unlikely to be harmed by participating in a survey or randomized experiment (compared to not participating)—think back to the traffic survey example above. But will study participants actually understand and act upon such guarantees? For example, we know from studies that validate "randomized response" (Warner, 1965)—a method of asking sensitive questions closely related to differential privacy—that many respondents do not understand, or do not trust, randomization (Kirchner, 2014). With randomized response, many act as though no privacy protections were in place (Coutts & Jann, 2011).

One of the possible benefits of differential privacy is that it would improve survey participation, by guaranteeing people's privacy (up to a factor epsilon). Social scientists already know quite a bit about peoples' participation in (scientific) surveys and official statistics; unfortunately, the evidence so far would not suggest implementing and informing respondents about differential privacy would have an overall positive effect. A strong emphasis on privacy might not have the desired effect of gaining participation in surveys or census data collection (Singer & Couper, 2010). In previous studies, informing potential respondents of lower privacy risks did not change their willingness to participate, and emphasizing possible harms *reduced* their willingness to participate (Couper, Singer, Conrad, & Groves, 2008, 2010). In addition, telling respondents that their answers are guaranteed to change the outcome only in the very slightest of ways—the core of differential privacy—may further lower, rather than improve, participation. After all, a cornerstone of convincing people to participate in surveys has been to emphasize the importance of their contribution (Dillman, Smyth, & Christian, 2014). Schnell (1997, p. 174) characterizes responses to surveys as a situation of low-cost helping behavior. Consequently, emphasizing the absence of making a difference, can trigger a form of 'bystander effect,' with potential participants less likely to help because others are already doing so, as has been found for crowdfunding campaigns (Kuppuswamy & Bayus, 2017).

As another example of why social science is relevant here, consider the main premise of differential privacy and other disclosure limitation guarantees: that people will prefer other parties to have only noisy measures of their data, rather than accurate ones. Will this premise hold for all people in all situations? Not according to studies of randomized response (RR), for example. Randomized response is a set of techniques to ask respondents whether they engaged in a sensitive behavior indirectly, so as to preserve privacy (Warner, 1965). In one version of the technique (the 'forced response design'), the respondent is asked to secretly throw a die; if it comes up six, he should always answer 'yes' (the sensitive behavior), regardless of the truth—otherwise, he should answer the question truthfully (see Blair, Imai, & Zhou, 2015, for an overview of RR techniques). Because the researcher cannot see the dice throw, she will never know whether the respondent engaged in the behavior, but she can still estimate the overall proportion of people who engaged in it (by solving the equation: $observed\ proportion = \frac{5}{6} \times true\ proportion + \frac{1}{6}$.)

Although this procedure is completely private, several studies have found that many people refuse to answer 'yes' when the die command them to do so—destroying the validity of the procedure (Coutts & Jann, 2011; Edgell, Himmelfarb, & Duchan, 1982; Kirchner, 2015). One explanation for this finding is that people do not trust or understand the procedure (Landsheer, Van Der Heijden, & Van Gils, 1999); a similar problem may occur with differential privacy. Alternatively, people may understand RR perfectly well, but still rationally prefer their true values to their noisy ones to be publicly available—one need only imagine a dictatorship persecuting political dissidents based on whatever information might be available, true or noisy. Less dramatically, differentially private outputs could sometimes lead to erroneous but apparently sensible reconstructions that have unintended consequences. Say, for example, noise is added to a data set including a classification into drug user vs. non–drug user. A reconstructed data set can now turn true nonusers into users and vice versa, risking possible harm if employers' agents investigate drug use.[6] Interestingly, in RR research, the converse to avoidance of 'yes' answers has also been found: respondents preferring to report the sensitive behavior (Höglinger & Diekmann, 2017). Continuing the dictatorship example, from a moral perspective one might even consider that such obfuscation is required, in order to protect those whose true values are in fact 'yes' (Brunton & Nissenbaum, 2016).

In other words, while the protections afforded by differential privacy should put the public at ease, for psychological and social reasons, the effect may sometimes turn out opposite to that intention. We do not currently know under which circumstances such situations might occur; exactly for this reason, privacy issues should be studied not only from a mathematical and computational, but also from a social perspective. Leaving this perspective out of the debate on differential privacy will prevent it from moving forward in the much-needed interest of public data protection.

4. How Will Social Science Be Affected by Differential Privacy?

In the previous section, we suggested some ways in which social science can inform the debate on data protection with differential privacy. In this section, we will show that social science itself will be affected by differential privacy. If data about people—whether this is from the census, a survey or some other source—are made increasingly noisy this may well cause a sea-change in social science. Some of the change will be negative, partially hindering our ability to draw clear conclusions of import. But not all consequences will be bad; differential privacy may help social scientists to more clearly see and communicate the limits of what can be learned about humans from data. To understand the effects of differential privacy on the social sciences it is useful to gain some understanding of key elements of quantitative social science research.

4.1 What Do Social Scientists Do With Data?

Social scientists want to understand how humans think, feel, and behave. *Quantitative* social scientists do this by analyzing data—traditionally administrative records, economic time series, surveys, and lab experiments; and, more recently, new data sources, including location data from mobile phones, accelerometers in smart watches, social media, and mass online experiments (Salganik, 2018). The social sciences are many and various, with highly distinct subdisciplines such as economics, psychology, sociology, political science, anthropology, communication studies, social geography, and public health. Equally diverse are the statistical methods encountered across these fields, ranging from *t* tests and ANOVA, to (linear) regression, factor analysis, multilevel (hierarchical) models, and complex Bayesian approaches—as well as simpler descriptive measures.

But despite this large diversity in data and methods, the approaches taken by these fields often share two commonalities. First, after a model is fit to the data, interest usually focuses on the *values* model parameters take, as well as their statistical (sampling) variation. For example, Fetzer (2019) estimated a regression coefficient measuring the effect of austerity in British constituencies on votes for Brexit[7]; the size of this coefficient tells us whether Remain could have won this referendum, if it had not been for austerity. Another example is Van de Rijt, Kang, Restivo, and Patil's (2014) study of 'success-breeds-success' dynamics. Van de Rijt et al. randomly gave financial contributions to nascent Kickstarter projects to see whether these arbitrarily lucky recipients subsequently also received more funding from third parties; a χ^2 test then determined how plausible the observed difference in funded projects would have been if there were no effect. In both examples, the focus is not on prediction but on aspects of the model itself—a characteristic we believe to be typical of many social science studies. Social scientists often tend to focus on interpreting parameter values, especially relationships among variables, as well as their sampling variability and inference rather than prediction.

Second, social science studies are often exploratory and adaptive, given the complex nature of human behavior. Therefore the analyses needed in many social science studies are difficult to pin down exactly in advance (Fiedler, 2018). For example, standard models of the macro-economy or human collaboration exist, and they can dictate which variables should be predictive of which others. However, a host of alternative choices, including control variables, form of the model, and subgroup analysis, may also be reasonable (Leamer, 1983) and they can give substantially different and scientifically interesting results. To illustrate this point, Silberzahn et al. (2018) asked 61 data analysts to investigate whether dark-skinned players get more red cards in soccer; the different teams reported odds ratios between 0.9 (10% *lower* odds) and 2.9 (290% *higher* odds). The situation is different when social scientists implement randomized experimental tests of theories developed from prior observation, with validated and commonly accepted measurement instruments: such studies can specify analyses in advance and can be preregistered (Nosek, Ebersole, DeHaven, & Mellor, 2018). However, the same does not hold for the prior observations that prompted such studies in the first place. Overall, in addition to predetermined goals and algorithms, there is a culture of data exploration in addition to predetermined goals and algorithms. Preregistration of analyses will therefore never be the only mode of social scientific study (see also Szollosi et al., 2019, for a provocative argument to this effect).

4.2 How Differential Privacy Might Affect Traditional Social Science

First, differential privacy, through its necessary randomization and censoring of the data, usually creates bias in estimates of relationships, just as random measurement error and selection bias do. This is easy to understand when you realize that the method relies on not knowing for certain to which category of, for example, sex, a person belongs; any differences in averages of other variables between the sexes will be blurred. Just as with random measurement error, however, knowing the problem is solving it: if the data provider tells us the differentially private algorithm that was used, it is possible to extend current statistical models to correct this blur. For example, if we know a random 100 men were intentionally misclassified as women and vice versa, then we can easily adjust the final table to account for this fact (Bakk, Oberski, & Vermunt, 2014; Di Mari, Oberski, & Vermunt, 2016; Fuller, 1987). The data provider only needs to tell us the total number moved, and what their chances of being misclassified were—without having to reveal *which* people's values were changed, guaranteeing both privacy and unbiasedness. In other words, bias is, fortunately, a solvable technical problem—albeit one that deserves more attention from the statistical community. It may also imply that *social scientists will need to routinely employ measurement error corrections to obtain unbiased estimates.*

Second, differential privacy necessarily adds a layer of nonsampling error. Propagating this additional error into the final analysis is, again, a technical problem. Computational theorists and statisticians are currently well underway to providing solutions to this problem. However, it does mean that the

uncertainty about parameters of interest will suffer a 'privacy effect,' similar to the 'sampling design effects' perhaps familiar to users of surveys. In other words, after privacy protections, the *effective sample size* will be lower—potentially much lower—than the original sample size (see Evans, King, Schwenzfeier, & Thakurta, 2019, for a calculation of effective sample size in differential privacy). In cases where sample sizes are currently sufficient for the intended statistical usage, they will have to be increased by the privacy effect. As explained above, the privacy effect is a measurement error effect, and its size will depend on the situation; it may be small, as found by Chetty & Friedman (2019); or, it may be devastatingly large, as found by Meng (2018). It has also been argued that many current sample sizes in social science are *already* inadequate for the intended usage (Button et al., 2013; Open Science Collaboration, 2015); in those cases, sample size will need to be increased even more. *Social scientists will likely need to drastically increase the number of people in their samples to achieve both privacy and acceptably powerful tests of their theories.*

Third, to release data for general usage with differential privacy guarantees, the party that releases the data must weigh the privacy requirements against the foreseen usage of the data. For example, if we know in advance that practitioners will only require linear correlations, then a lot of information in the raw data can be thrown away to the benefit of privacy, while preserving the correlation structure. But what if the releaser does not know what will be done with the data? Is it possible to protect privacy and also allow for any and all potential analysis to yield an accurate answer? Unfortunately, the answer is a resounding 'no'; after all, if the system can accurately answer any question at all, this will also include questions about individuals—something the mechanism is explicitly designed to prevent. In fact, the "database reconstruction theorem," which states that releasing too much information will always allow an attacker to reconstruct the original database accurately, was the original reason why differential privacy was invented (Dinur & Nissim, 2003). The upshot is that it will no longer be possible to publish data sets that are fit for any number of different purposes: *Social scientists will have to much more severely limit the type, scope, and/or number of queries they perform on any given data set, ahead of time.*

Fourth and finally, some subdisciplines will be more affected than others. Researchers that study small groups may find their current methods no longer suffice. For example, if one were to apply differentially private mechanisms to the 2014 Polish European Social Survey, any result involving the small group of 12 ethnic minority respondents will likely be substantially changed. 'Mixed-method' researchers have suggested probing quantitatively outlying survey respondents with in-depth qualitative interviewing (Gibbert, Nair, & Weiss, 2014). Differential privacy would prevent this, since outliers—and qualitative research in general—are by definition privacy-sensitive. *Social scientists in these fields will likely require new research designs with increased costs,* such as oversampling of minorities or two-step mixed-method approaches that use protected quantitative data to learn about 'typicality,' followed by a qualitative study of atypical cases. An alternative might be sought in *new*

infrastructural designs with increased costs, such as allowing third-party researchers to contact outlying respondents based on consent.

In short, researchers will need to use more complex statistical methods to account for nonsampling errors in their data; they will often need to drastically increase their sample sizes; they will have to severely limit the scope and complexity of their research questions to some extent; and they will sometimes need to target their data collection much more precisely to their questions.

4.3 Differential Privacy and Changing Traditions in Social Science

Social science is not static, and in the past few years has undergone at least three rapid changes. First, unprecedented detail about individuals is available. The traditional questionnaire and experimental data are increasingly linked to data from smartphones and online behavior, as well as other measures, including genome, eye tracking, video, audio, biomarkers, and fMRI brain scans—often by following a group of people longitudinally, that is, over time (Salganik, 2018). Second, the open science movement has gained traction within and beyond social science. Funders, journals, and employers of social-scientific researchers increasingly require open access to papers, open and reproducible analysis code, and FAIR (findable, accessible, interoperable, and reusable) data publication (European Commission & Directorate-General for Research and Innovation, 2018; National Institutes of Health, 2020; Wilkinson et al., 2016). Third, wherever randomized experiments are popular, open science is often linked to a call for 'preregistration': the practice of publicly announcing as exact a description of the intended analysis steps as possible, in advance of the actual data collection (Nosek et al., 2018).

In short, social science is rapidly become *more open, less exploratory* (in some subdisciplines)**,** and *more complex.* These developments have clear implications for (differential) privacy. The clamor for reproducibility[8]—involving the original data and code —clearly provides a strong justification for differential privacy, since the research community's calls to share data in open repositories can only be met by providing some form of privacy protection. Preregistration is another clear win for the pairing of differential privacy and modern social science: by prespecifying exactly what the data will be used for, differential privacy can be achieved by a straightforward application of existing principles. Collected data can be shared in a differentially private manner that also affords full reproducibility of any fully prespecified analyses, as well as power calculations. At the same time, more complexity in data analysis poses a challenge, because it requires more detailed personal information—threatening privacy if these data are to be shared. Social scientists are put between the rock of nonreproducibility and the hard place of limiting the complexity of their data, and, by extension, their research questions.

4.4 Are These Consequences Bad or Good?

Though privacy protections provide a benefit to the data subjects, they may be detrimental to the researcher without additional funding, since collecting more data may be expensive. They will inherently limit what can be learned from census data, since there the sample size can't be increased. And they might limit the types of research questions that can be answered. However, at its heart, differential privacy is about limiting the sensitivity of one's conclusions to the presence or absence of any one person in the analysis. In this sense, it serves simply as a reminder of the importance of robustness. And lack of robustness is precisely the property that dubious statistical practices, such as *p*-hacking, HARKing, data peeking, and other overfitting activities share with one another (Dwork et al., 2015). When using robust methods to analyze data, we already limit the effect an individual observation can have on the result. In this situation less noise needs to be added in order to meet the criteria of differential privacy. Likewise, when noise is added in a differential privacy context, repeated analyses of the data are constrained by the privacy budget and thus make it hard to engage in *p*-hacking. Working in a differential privacy context could therefore hold social science to account, not only for enforcing privacy, but also for enforcing statistical rigor. And that might not be such a bad idea, after all.

5. Conclusion

In the past decades, social science has become more quantitative, and has expended considerable effort and public funding to collect new data that might improve society. We now face the obligation to both protect the data subjects' privacy and leverage the utility of these data for social good. To work out the best way forward, we will need computer scientists, IT specialists, ethicists, statisticians, mathematicians, lawyers, managers, governments, private companies, public research institutes, policymakers, and lawmakers —and, yes, social scientists—to work together. We foresee an important role for differential privacy: as one piece of a difficult, and fascinating, puzzle that society desperately needs us to complete.

Disclosure Statement

Part of this paper was written during the semester on privacy at the Simons Institute for the Theory of Computing. We would like to thank the semesters' participants for inspiring conversations that shaped the paper, in particular Cynthia Dwork, Helen Nissenbaum, and Kobbi Nissim. Daniel Oberski was supported by the Netherlands Organisation for Scientific Research, NWO VIDI grant [VI.Vidi.195.152]. Frauke Kreuter received additional support for this work from the Mannheim Center for European

Social Research, from the SFB 884 (DFG 139943784 and DFG 139943784), and from the Volkswagen Foundation.

Acknowledgments

We thank Samantha Chiu, Jörg Drechsler, Frederic Gerdon, Konstantin Kakaes, Arthur Kenickell, Florian Keusch, Gary King, Julia Lane, Katrina Ligett, Xiao-Li Meng, Marcel Neunhoeffer, Patrick Schenk, Rainer Schnell, and Adam Smith for their comments on previous versions of this article.

References

APPI (2016) Kojin jōhō no hogo ni kansuru hōritsu [Act on the Protection of Personal Information (APPI)], Act No. 57 of 2003 (May 30, 2003), *last amended by* Act No. 51 of 2016, English translation as amended by Act No. 65 of 2015 http://www.japaneselawtranslation.go.jp/law/detail_main?re=02&ia=03&vm=02&id=2781, *archived at* https://perma.cc/SD3Z-UU8T

Abowd, J. M. (2018). The U.S. Census Bureau adopts differential privacy. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining—KDD '18*, 2867–2867. https://doi.org/10.1145/3219819.3226070

Abowd, J. M., & Schmutte, I. M. (2019). An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices. *American Economic Review*, *109*(1), 171–202. https://doi.org/10.1257/aer.20170627

Alba, D. (2019, September 30). Facebook Fails to Part With Data. *The New York Times*, Section B, page 1, (New York edition). URL: https://www.nytimes.com/2019/09/29/technology/facebook-disinformation.html (accessed 2020-01-28).

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2014). Relating latent class assignments to external variables: Standard errors for correct inference. *Political Analysis*, *22*, 520–540.

Bakker, B. F., Van Rooijen, J., & Van Toor, L. (2014). The system of social statistical datasets of Statistics Netherlands: An integral approach to the production of register-based social statistics. *Statistical Journal of the IAOS*, *4*, 411–424. https://doi.org/10.3233/SJI-140803

Blair, G., Imai, K., & Zhou, Y.-Y. (2015). Design and analysis of the randomized response technique. *Journal of the American Statistical Association*, *110*, 1304–1319. https://doi.org/10.1080/01621459.2015.1050028

Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447.

Brunton, F., & Nissenbaum, H. (2016). *Obfuscation: A user's guide for privacy and protest*. Cambridge, MA: MIT Press.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365–376. https://doi.org/10.1038/nrn3475

Chetty, R., & Friedman, J. N. (2019). A practical method to reduce privacy loss when disclosing statistics based on small samples. *AEA Papers and Proceedings*, *109*, 414–420. https://doi.org/10.1257/pandp.20191109

Couper, M. P., Singer, E., Conrad, F. G., & Groves, R. M. (2008). Risk of disclosure, perceptions of risk, and concerns about privacy and confidentiality as factors in survey participation. *Journal of Official Statistics*, *24*, 255–275.

Couper, M. P., Singer, E., Conrad, F. G., & Groves, R. M. (2010). Experimental studies of disclosure risk, disclosure harm, topic sensitivity, and survey participation. *Journal of Official Statistics*, *26*, 287–300.

Coutts, E., & Jann, B. (2011). Sensitive questions in online surveys: Experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT). *Sociological Methods & Research*, *40*, 169–193. https://doi.org/10.1177/0049124110390768

Di Mari, R., Oberski, D. L., & Vermunt, J. K. (2016). Bias-Adjusted Three-Step Latent Markov Modeling with Covariates. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(5), 649–660. https://doi.org/10.1080/10705511.2016.1191015

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method* (4th ed.). Hoboken, NJ: Wiley.

Dinur, I., & Nissim, K. (2003). Revealing information while preserving privacy. *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 202–210. https://doi.org/10.1145/773153.773173

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., & Roth, A. (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science*, *349*, 636–638.

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, *9*(3–4), 211–407.

Dwork, C., & Ullman, J. (2018). The Fienberg problem: How to allow human interactive data analysis in the age of differential privacy. *Journal of Privacy and Confidentiality*, *8*(1). https://doi.org/10.29012/jpc.687

Edgell, S. E., Himmelfarb, S., & Duchan, K. L. (1982). Validity of forced responses in a randomized response model. *Sociological Methods & Research*, *11*, 89–100. https://doi.org/10.1177/0049124182011001005

Elliot, M., O'Hara, K., Raab, C., O'Keefe, C. M., Mackey, E., Dibben, C., … McCullagh, K. (2018). Functional anonymisation: Personal data and the data environment. *Computer Law & Security Review*, *34*, 204–221. https://doi.org/10.1016/j.clsr.2018.02.001

European Commission, & Directorate-General for Research and Innovation. (2018). *Turning FAIR data into reality: Final report and action plan from the European Commission expert group on FAIR data.* Retrieved from http://publications.europa.eu/publication/manifestation_identifier/PUB_KI0618206ENN

European Parliament and Council. General Data Protection Regulation, Pub. L. No. Regulation (EU) 2016/679, Https://eur-lex.europa.eu/eli/reg/2016/679/oj (2018).

Evans, G., King, G., Schwenzfeier, M., & Thakurta, A. (2019). *Statistically valid inferences from privacy protected data*. Retrieved from: https://gking.harvard.edu/files/gking/files/udp.pdf

Fellegi, I. P. (1972). On the question of statistical confidentiality. *Journal of the American Statistical Association*, *67*(337), 7–18. https://doi.org/10.1080/01621459.1972.10481199

Fetzer, T. (2019). Did austerity cause Brexit? *American Economic Review*, *109*, 3849–3886. https://doi.org/10.1257/aer.20181164

Fiedler, K. (2018). The creative cycle and the growth of psychological science. *Perspectives on Psychological Science*, *13*, 433–438. https://doi.org/10.1177/1745691617745651

Fuller, W. A. (1987). *Measurement error models*. New York, NY: Wiley.

Gibbert, M., Nair, L. B., & Weiss, M. (2014). Ooops, I've got an outlier in my data—What now? Using the deviant case method for theory building. *Academy of Management Proceedings*, *2014*, 12411. https://doi.org/10.5465/ambpp.2014.12411abstract

Hansen, M. (2019, January 2). Differential privacy, an easy case. Retrieved January 5, 2019, from https://accuracyandprivacy.substack.com/p/differential-privacy-an-easy-case

Höglinger, M., & Diekmann, A. (2017). Uncovering a blind spot in sensitive question research: False positives undermine the crosswise-model RRT. *Political Analysis*, *25*, 131–137. https://doi.org/10.1017/pan.2016.5

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Gießing, S., Schulte Nordholt, E., Spicer, K., & De Wolf, P.-P. (2012). *Statistical disclosure control*. Chichester, West Sussex: Wiley.

King, G., & Persily, N. (2019). A New model for industry—Academic partnerships. *PS: Political Science & Politics*, 1–7. https://doi.org/10.1017/S1049096519001021

Kirchner, A. (2014). *Techniques for asking sensitive questions in labour market surveys*. Bielefeld, Germany: Bertelsmann.

Kirchner, A. (2015). Validating sensitive questions: A comparison of survey and register data. *Journal of Official Statistics*, *31*, 31–59. https://doi.org/10.1515/jos-2015-0002

Kuppuswamy, V., & Bayus, B. L. (2017). Does my contribution to your crowdfunding project matter? *Journal of Business Venturing*, *32*, 72–89. https://doi.org/10.1016/j.jbusvent.2016.10.004

Landsheer, J. A., Van Der Heijden, P., & Van Gils, G. (1999). Trust and understanding, two psychological aspects of randomized response. *Quality and Quantity*, *33*, 1–12. https://doi.org/10.1023/A:1004361819974

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, *73*, 31–43.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, *12*, 685–726. https://doi.org/10.1214/18-AOAS1161SF

Mervis, J. (2019, January 4). Can a set of equations keep U.S. census data private? *Science*. https://doi.org/10.1126/science.aaw5470

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. *2008 IEEE Symposium on Security and Privacy (Sp 2008)*, 111–125. https://doi.org/10.1109/SP.2008.33

National Academies of Sciences, Engineering, and Medicine (2019). *Reproducibility and replicability in science*. Washington, DC: National Academies Press. https://doi.org/10.17226/25303.

National Institutes of Health (NIH). (2020). *Draft NIH policy for data management and sharing*. Retrieved from https://osp.od.nih.gov/draft-data-sharing-and-management/

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*, 2600–2606. https://doi.org/10.1073/pnas.1708274114

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Ruggles, S., Fitch, C., Magnuson, D., & Schroeder, J. (2019). Differential privacy and census data: Implications for social and economic research. *AEA Papers and Proceedings, 109*, 403–408. https://doi.org/10.1257/pandp.20191107

Salganik, M. J. (2018). *Bit by bit: Social research in the digital age*. Princeton, NJ: Princeton University Press.

Schnell, R. (1997). *Nonresponse in Bevölkerungsumfragen: Ausmaß, Entwicklung und Ursachen*. Retrieved from https://doi.org/10.1007/978-3-322-97380-1

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., … Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science, 1*, 337–356. https://doi.org/10.1177/2515245917747646

Singer, E., & Couper, M. P. (2010). Communicating disclosure risk in informed consent statements. *Journal of Empirical Research on Human Research Ethics, 5*(3), 1–8. https://doi.org/10.1525/jer.2010.5.3.1

Strandburg, C. J. (2014). Monitoring, datafication, and consent: Legal approaches to privacy in the big data context. In J. Lane, V. Stodden, S. Bender, & H. Nissenbaum (Eds.), *Privacy, big data, and the public good: Frameworks for engagement* (pp. 5–43). https://doi.org/10.1017/CBO9781107590205

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

Szollosi, A., Kellen, D., Navarro, D., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2019). *Preregistration is redundant, at best*. https://doi.org/10.31234/osf.io/x36pz

van de Rijt, A., Kang, S. M., Restivo, M., & Patil, A. (2014). Field experiments of success-breeds-success dynamics. *Proceedings of the National Academy of Sciences, 111*, 6934–6939. https://doi.org/10.1073/pnas.1316836111

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*(309), 63–69. https://doi.org/10.1080/01621459.1965.10480775

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

Wood, A., Altman, M., Bembenek, A., Bun, M., Gaboardi, M., Honaker, J., … Vadhan, S. (2018). Differential privacy: A primer for a non-technical audience. *Vanderbilt Journal of Entertainment & Technology Law*, *21*, 209–276.

Footnotes

1. For continuous-valued outcomes, such as regression coefficients or averages: the probability density ↩

2. See here for an English Translation https://www.ppc.go.jp/files/pdf/Act_on_the_Protection_of_Personal_Information.pdf ↩

3. See here for an English Translation https://iapp.org/media/pdf/resource_center/Brazilian_General_Data_Protection_Law.pdf ↩

4. https://oag.ca.gov/privacy/ccpa ↩

5. https://meity.gov.in/writereaddata/files/Personal_Data_Protection_Bill,2018.pdf ↩

6. Strandburg (2014) discusses privacy risks and examples of potential harm through investigations. ↩

7. Brexit is the withdrawal of the United Kingdom from the European Union. ↩

8. National Academies of Sciences, Engineering, and Medicine (2019, p. 7) on need for code and original data for reproducibility as opposed to replicability, which could be achieved without the original data. ↩