

Will Differential Privacy Transform Social Science?

Data privacy concerns are all around us and social scientists need to pay attention.

Daniel L. Oberski

Utrecht University

Frauke Kreuter

University of Maryland, University of Mannheim

Institute for Employment Research

Corresponding author:

Frauke Kreuter, 1218 Lefrak Hall, College Park MD 20742

fkreuter@umd.edu ; 202 390 0143 ; @fraukolos

Abstract

Accessing and combining large amounts of data is important for social scientists. Increasing amounts of data also increase privacy risks. There is significant pressure from society to increase data protection. Several important players in official statistics, academia, and business see differential privacy as the solution. In this opinion piece, we put differential privacy in a larger context and discuss, from a social science research perspective, pros and cons for adapting differential privacy. It becomes clear that the social science research workflow must change if differential privacy is implemented. It also becomes clear that common social science data collection will become more costly. However, there are a series of positive side effects in implementing this approach, in addition to preserving privacy, that could solve some issues social scientists currently struggle with. We conclude with an assessment of a seemingly reasonable approach in the short-term, given current technology, and point out why we think that collecting data with the promise of using it in a differentially private way will likely not change the participation decision of the public, but may help in sharing data across institutions.

Keywords:

Differential Privacy; Social Science; Data Science; Open Data; Robustness

Acknowledgements: Arthur Kenickell, Rainer Schnell, Katrina Ligett, Jörg Drechsler, Xiao-Li Meng, Patrick Schenk, Florian Keusch. The paper was written during the semester on privacy at the Simons Institute for the Theory of Computing. The first author was supported by the Netherlands Organisation for Scientific Research, NWO VIDI grant [VI.Vidi.195.152].

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 answer this question in the 1970's, 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).

Today in 2019, Fellegi'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 such 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 GDPR; and large tech firms started to use built-in privacy protections to market its 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, in particular, is poised to completely change the way we analyze data about people: differential privacy.

Differential privacy is coming

"Differential privacy" is a simple mathematical definition of when publishing results or datasets 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.

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 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 personal 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. **From the perspective of differential privacy researchers neither one of**

the two confidentiality pledges statements is sufficient, and releasing the aggregate is the same as releasing the microdata.

Removing personal identifiable information is insufficient because we are now living in a world with many different data out there. The worry is that some of the external data might include the PII plus the exact same combination of variables as the researchers' data (e.g. Sweeney et al. 2018). 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 (Narayanan & Shmatikov, 2008).

Often overlooked is the fact that access to aggregate statistics can (under certain circumstances) 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 mentally consider all hypothetical datasets, and then “weed out” those datasets that could never have produced the reported aggregate statistics. For example, a thousand-person dataset 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 dataset at the same time, an attacker can, like Sherlock Holmes, rule out the impossible and arrive at the original dataset. The only way to prevent database reconstruction is to scramble any reported output from the original data with noise. The added noise prevents the Sherlock Holmes procedure from weeding out impossible datasets, 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 linkage.

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 (usually referred to as epsilon) 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, algorithms can satisfy the definition. 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 and Roth (2014).

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; and whether that tradeoff is a good deal will depend on what the data are, and how they are used.

This privacy-utility tradeoff is not, in itself, 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 she might reasonably use (Hundepool 2012; Elliot et al. 2018). But what is “sensitive” depends on context, and in the age of datafication 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* datasets, 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 United States Census Bureau, for its part, has decided the benefits of guaranteeing high levels of differential privacy outweigh the costs, and recently announced that all 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 the traditional uses of the Census, including redistricting, subsidies, and some economic analyses (Mervis, 2019).

Can the science of people peer through the noise?

But the consequences of adopting differential privacy extend well beyond the Census: all of social science will be affected. There is currently an extreme pressure from society to increase data protection - as exemplified by new legislation such as the GDPR (European Parliament and Council, 2018), the 2020 [California Consumer Privacy Act](#), or the 2019 [New York SHIELD act](#). At the same time, many important players in official statistics, academia, and business see differential privacy as the solution (Abowd & Schmutte, 2018; Mervis, 2019). So it seems likely that much human data that is publicly available now will, in the future, be published using differentially private mechanisms - or not at all.

If data about people 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 actually help social scientists to more clearly see and communicate the limits of what can be learned about humans from data.

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 in spite of 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; 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, & Patil's (2014) study of “success-breeds-success” dynamics. Van de Rijt et al. randomly gave crowdfunding contributions to beginning 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.

Second, the analyses needed in many social science studies would be difficult to pin down exactly in advance (Fiedler, 2018). For example, even though standard models of the macro-economy or human collaboration exist, and dictate which variables should be predictive of which others, a host of alternative choices, including control variables, form of the model, and subgroup analysis, may also be reasonable (Leamer, 1983), and 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). An important exception is randomized experimental tests of theories developed from prior observation, with validated and commonly accepted measurement instruments: such studies can be preregistered to scientific advantage (Nosek et al., 2018). Of course, the same does not hold for the prior observations that prompted the theory in the first place. Preregistration can therefore never be the only mode of social scientific study (see Szollosi et al., 2019 for a provocative argument to this effect).

How will traditional social science be affected by differential privacy?

We have seen that quantitative social science currently tends to focus on interpreting parameter values, especially relationships among variables, as well as their sampling variability - rather than prediction - and has a culture of data exploration - rather than strictly predetermined goals and algorithms.

So how do these traditional characteristics of social science play with differential privacy measures?

First, differential privacy, through its necessary randomization of the data, sometimes creates bias in estimates of relationships, just as random measurement error does. 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 particular 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 (e.g. Bakk et al., 2014; Di Mari et al., 2016). 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 non-sampling 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. 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 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 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 a reasonably 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 datasets that are fit for every purpose: **Social scientists will have to explicitly limit the type, scope, and/or number of questions they ask of any given dataset, ahead of time.**

Fourth and finally, some sub-disciplines 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 search for 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.

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, eyetracking, video, audio, biomarkers, and fMRI brain scans - often by following a group of people longitudinally over time (Salganik, 2018). Second, the open science movement, which originated in psychology, 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 “open, unless” FAIR data publication (Wilkinson et al., 2016; European Commission & Directorate-General for Research and Innovation, 2018; NIH, 2020). 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 and open data clearly provides a strong reason 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 this data is 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.

Differential privacy: just what the doctor ordered?

It seems clear that differential privacy, if implemented widely, will substantially transform social science. 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 explicitly 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.

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 increasing sample sizes 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 analyse 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.

Differential privacy - the controversy

Differential privacy has been readily embraced by some as an attractive mathematical framework to protect privacy and prevent overfitting; others have reported churned stomachs when they ponder its effects on data utility and data analysis.

Differential privacy certainly has its advantages and disadvantages. It guarantees 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). 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 *and* reconstruction.

But where we might expect a dry weighing of facts, we actually observe a highly heated debate, which shows no signs of abating. Why do ordinarily level-headed 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?

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. At their core, traditional statistical disclosure limitation (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 datasets 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, whereas differential privacy postulates that it cannot. By definition, this disagreement cannot be assessed with evidence, because it is exactly the unobserved parts of risk on which beliefs differ.

What is the way forward for social data protection?

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. 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, differential privacy can only hurt, since there is no evidence of privacy risks and therefore no known benefit from implementing privacy guarantees. Of course, there may be unknown risks. But, assuming that organizations such as the European Social Survey have carefully weighed these risks, it seems unwise to us to advocate implementing differential privacy for this category.

The second category of data entails clear privacy risks, but is available in "data enclaves". Trusted researchers can access the data in a secure computing environment after passing an often impressive number of hurdles, including binding agreements to honor the participants' confidentiality. Such agreements can include disclosure protections, such as differential privacy, of any publicly released outputs. 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 to link administrative records to surveys, such as the [Coleridge Initiative](#) in the US, and [ODISSEI](#) and the System of Social Statistical Datasets at Statistics Netherlands (Bakker et al., 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 requires 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](#). 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](#).

Privacy - the multidisciplinary responsibility

Privacy is a social issue. It involves norms, expectations, trust, understanding, and relationships. So 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 it is not clear that potential study participants actually understand and act upon such guarantees. For example, we know from studies that validate “randomized response” - a method of asking sensitive questions closely related to differential privacy - that many respondents do not understand, or do not trust, randomization. With randomized response, many act as though no privacy protections were in place. Worse: telling respondents that their answers are guaranteed to change the outcome only in the very slightest of ways may actually lower, rather than improve, participation. We also know that people do not respond to surveys based on objective cost-benefit analyses; if this were the case, five-dollar incentives to do a two-hour survey would not work, and a response of more than 10 participants would be a small miracle. This phenomenon is well known in the study of voter turnout in democratic elections, where it is called the “paradox of voting”; here, we propose a “paradox of survey participation”.

As another example of why social science is relevant here, consider the 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. This preference may not actually hold for all people in all situations. For example, as large advertising companies such as Google have long argued, people who

accept the need for advertisements, may prefer interesting, accurately targeted, ads to inaccurately targeted ones. In other cases, adding noise to data may lead to uncomfortable situations. For example, past randomized response research has found that people who have *not* engaged in socially undesirable behavior often refuse to falsely report such behavior - even when they know the researcher has no way of discovering their true value. Similarly, adding noise in a differentially private setting could lead to erroneous but apparently sensible reconstructions that have unintended consequences. In other words, while the protections afforded by differential privacy are intended as a comfort to the public, 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.

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 for privacy, we will need computer scientists, IT specialists, philosophers, statisticians, mathematicians, lawyers, managers, governments, private companies, public research institutes, policy- 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.

References

- 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. (2018). *An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices* (SSRN Scholarly Paper No. ID 3232398). Retrieved from Social Science Research Network website: <https://papers.ssrn.com/abstract=3232398>
- Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2014). Relating Latent Class Assignments to External Variables: Standard Errors for Correct Inference. *Political Analysis*, 22(4), 520–540. <https://doi.org/10.1093/pan/mpu003>
- 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>
- 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(5), 365–376. <https://doi.org/10.1038/nrn3475>
- Chetty, R., & Friedman, J. (2019). *A Practical Method to Reduce Privacy Loss when Disclosing Statistics Based on Small Samples* (No. w25626; p. w25626). <https://doi.org/10.3386/w25626>
- 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>
- 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(6248), 636–638. <https://doi.org/10.1126/science.aaa9375>
- Dwork, Cynthia, & 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>
- 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).
- Fetzer, T. (2019). Did Austerity Cause Brexit? *American Economic Review*, 109(11), 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(4), 433–438. <https://doi.org/10.1177/1745691617745651>
- Gibbert, M., Nair, L. B., & Weiss, M. (2014). Oops, I've Got an Outlier in My Data – What Now? Using the Deviant Case Method for Theory Building. *Academy of Management Proceedings*, 2014(1), 12411. <https://doi.org/10.5465/ambpp.2014.12411abstract>
- King, G., & Persily, N. (2019). A New Model for Industry–Academic Partnerships. *PS: Political Science & Politics*, 1–7. <https://doi.org/10.1017/S1049096519001021>

- Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, 73(1), 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(2), 685–726. <https://doi.org/10.1214/18-AOAS1161SF>
- Mervis, J. (2019). 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>
- 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(11), 2600–2606. <https://doi.org/10.1073/pnas.1708274114>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716–aac4716. <https://doi.org/10.1126/science.aac4716>
- Salganik, M. J. (2018). *Bit by bit: Social research in the digital age*. Princeton: Princeton University Press.
- 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(3), 337–356. <https://doi.org/10.1177/2515245917747646>
- Stigler, S. M. (2002). *Statistics on the table: The history of statistical concepts and methods* (3. printing). Cambridge, Mass.: Harvard Univ. Press.
- Sweeney, L., Loewenfeldt, M. von, & Perry, and M. (2018). Saying it's Anonymous Doesn't Make It So: Re-identifications of “anonymized” law school data. *Technology Science*. Retrieved from <https://techscience.org/a/2018111301/>
- 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(19), 6934–6939. <https://doi.org/10.1073/pnas.1316836111>
- 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(1). <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(1), 209–276.