Commission on Evidence-Based Policymaking
February 24, 2017

Open Meeting

Commissioners Present:
Katharine G. Abraham, Chair
Ron Haskins, Co-Chair
Sherry Glied (via phone)
Robert M. Groves
Robert Hahn (via phone)
Hilary Hoynes
Jeffrey Liebman
Bruce D. Meyer
Paul Ohm
Kathleen Rice (via phone)
Robert Shea
Latanya Sweeney
Kenneth R. Troske
Kim Wallin

The open meeting was called to order at 10:00 AM. Chair Katharine Abraham provided an introduction.

Presentation 1: Legal Standards for Achieving De-Identification

Alexandra Wood, Berkman Klein Center for Internet & Society, Harvard University

- Ms. Wood said that open data initiatives are encouraging agencies to make more data available but that these data carry privacy risks. Many agencies de-identify data but privacy science is evolving and the Federal government needs new practices to protect privacy. The current practices won't sufficiently protect privacy over the long term.
- Ms. Wood said that the U.S. generally takes a sectoral approach to privacy laws and standards that depend on the context. Some standards focus on method while others focus on objectives.
- Ms. Wood discussed variations in selected laws. The Health Insurance Portability and Accountability Act (HIPAA) uses expert determination to document that risk of re-identification is small; it also has a "safe harbor" provision. The Federal Education Rights and Privacy Act (FERPA) establishes a standard that allows release of de-identified data without explicit consent after removal of personally-identifiable information provided that the agency has made a reasonable determination that individuals cannot be re-identified. She added that this provision is interesting because it requires consideration of whether other data releases could be used in a re-identification attack. The Confidential Information Protection and Statistical Efficiency Act (CIPSEA) protects data in identifiable form when identity could be "reasonably inferred" directly or indirectly.
- Ms. Wood discussed some gaps in the existing privacy protection framework, including that personally-identifiable information is not precisely defined so must be interpreted, guidance

on application of legal standards is limited and not clear to practitioners (Working Paper 22 from the Federal Committee on Statistical Methodology focuses on de-identification techniques but is hundreds of pages long), and standards for re-identification focus on microdata not aggregate data yet aggregate releases also carry privacy risks that should be taken into account.

Discussion and Questions:

- Commissioner Troske asked whether tiered access helps with the variation in how agencies release data. Ms. Wood suggested looking at her slide on "a modern approach" about selecting combinations of privacy and security controls. She said that agencies should conduct a systematic analysis of their data that points to the controls that are appropriate for each case. Tiered access allows different access to different data for different purposes. The same data can be made available in different versions with different restrictions. Tiered access may help agencies achieve more standardization though it is unlikely to be perfect. A systematic approach could help agencies converge in the middle.
- Commissioner Wallin asked whether privacy laws impede the standardization of tiered access. Ms. Wood responded that a common standard for privacy would be beneficial because many laws apply to many different sectors and settings and how each is interpreted differently depends on the institution. Standardization about which requirements apply to which settings would be helpful.
- Commissioner Groves asked how to judge which data goes in which tiers. Ms. Wood responded that there are Institutional Review Board (IRB) practices and guidance that address this issue. She also recommended that agencies develop guidance on assessing the sensitivity and potential harm from different types of data. She added that there is a real need for more guidance in this area.

Panel: Technologies for Minimizing Risks to Data Security and Privacy, Part 1

Presentation #1: *Daniel Goroff, Alfred P. Sloan Foundation*

- Mr. Goroff told the Commission that he is "in the public good business." He said everyone would like a free ride, and the solution is taxes or philanthropy. He added that data themselves are not a public good but open data is a public good. But he cautioned that using data for policymaking means "using it up" for other purposes.
- Mr. Goroff described the difference between data and evidence. Data alone can't give you the conclusions needed in "if/then" statements; you need modeling as well. There are spillover effects in using data because every query can affect privacy and validity.
- Mr. Goroff gave the example of testing multiple hypotheses against a p-value of .05—sometimes called hypothesis fishing, data mining, or p-hacking. Solutions are to limit access (but that hurts reproducibility), pre-register hypotheses, or use differential privacy.
- Mr. Goroff indicated that there are new ways to protect privacy including differential privacy solutions and synthetic datasets. He suggested that researchers could explore using synthetic data, decide on their hypotheses, then get access to the original data. This makes a distinction between exploratory and confirmatory research. Facilitating that kind of research reduces the transaction cost to privacy of each query on the data.

Presentation #2: *Lars Vilhuber, Cornell University*

- Mr. Vilhuber spoke about his experiences accessing microdata and about physical safeguards for data. He said that one tradeoff is between access and loss of detail. Public use microdata, for example, is easy to access and convenient to use. Generations of researchers have had access to public use files. Whey researchers need confidential data, they have become used to going to secure rooms. The old style way of doing things was to lock the researchers in a data enclave.
- Today there are virtual data enclaves with thin client connections to secure servers. Virtualization is a reality. The Federal Statistical Research Data Centers (FSRDCs) have been virtual since the early 2000s; the data never leave the server in Bowie, Maryland. European models of secure data access started out virtual.
- The basic question is where and how can researchers access data. The answer to "where" can be anywhere or in a secure environment like a mini pod at a university or at a data enclave controlled by the Census Bureau.
- Mr. Vilhuber discussed how disclosure review works today. He suggested that authorized researchers could do their own disclosure review based on established rules. Denmark does this now as do some Census Bureau researchers. Adherence to the rules is occasionally verified. The final firewall, he noted, is in the minds of the people who are accessing the data.

Discussion and Questions

- Commissioner Shea asked Mr. Vilhuber to describe a situation where there was a damaging release from a physical or a virtual data enclave. Mr. Vilhuber said no breach that he knew of had ever occurred. There had been minor procedural violations but no serious disclosure breaches. He added that there needs to be a credible stick. There are rumors but no evidence that universities have lost access to data due to violations. Mr. Goroff stated that if you use differential privacy you can't tell if any one person is in the data set, so you don't need disclosure review at all. Mr. Vilhuber replied that his colleagues in Europe are concerned about eyeballing today. There is no control except whom you let in. There are stories from other countries about losses of laptops or sale of tax data, so creating a culture among researchers about protecting data is important. It's important to distinguish between breaches by researchers versus breaches by another outside source.
- Co-Chair Haskins asked if there had ever been a breach of confidential data by researchers in Europe and Mr. Vilhuber responded that he had never heard of such a breach and would have if it had been significant.
- Commissioner Meyer asked if the logic of differential privacy implies that the government should prioritize the most important research because every use of data uses up the privacy budget. Mr. Goroff said that differential privacy is a theorem that says some validity and privacy is used up each time you use the data so it's good to slow down that use while allowing exploratory research. Synthetic data allows for that.
- Commissioner Ohm asked if Mr. Goroff was recommending the Commission see differential privacy as a lynchpin or as one of the many tools to consider. Mr. Goroff replied that differential privacy has a precision that means he doesn't care if the government uses his data because researchers won't know if he's in the dataset let alone what his data points are. He added that differential privacy protects data subjects now and regardless of how the data are used in the future.

- Commissioner Hoynes and Chair Abraham asked whether brick and mortar data enclaves have advantages over virtual models. Mr. Vilhuber replied that even physical enclaves don't prevent eyeballing. He added that training is important. The FSRDCs did a survey about how much researchers knew about data protection rules before and after training and there was a marked improvement. Mr. Goroff said that even with differential privacy you can still have the eyeballing problem because someone has to curate and clean the data. You have to establish trust and culture and procedures for a small number of people to access the original data.
- Commissioner Wallin asked presenters to elaborate on the pros and cons of centralizing data infrastructure. Mr. Vilhuber said that, in practice, virtualization means that the data resides at the statistical agency and everything needs to be done on those computers. There could be additional data enclaves, but there is a complaint that it takes a long time to do analysis at the FSRDC that could be done quickly on a laptop. He said that the smallest computer node is 14 times more powerful than the FSRDC system as a whole. Virtualization, he added, just means that you trust the data center that houses the data.
- Chair Abraham asked Mr. Goroff for the pros and cons of having a network of facilities in the Commission's context. Mr. Goroff replied that to do curation and cleaning of data and to serve researchers it is important to involve the people who know the data best. He suggested making data owners part of a network that agrees on standards.

Panel: Technologies for Minimizing Risks to Data Security and Privacy, Part 2

Presentation #1: *Jerome Reiter, Duke University*

- Mr. Reiter spoke about giving researchers access to confidential social science data. He said there are three components: giving researchers unrestricted access to synthetic data, using a verification server with synthetic data, and giving approved researchers access to the real data.
- Mr. Reiter said that modeling makes the results of synthetic data for wide aggregates the same as for real data. The verification server performs an analysis on both the synthetic and the real data and returns measures of how close the results are. Even if the researcher then decides to seek approval to access the real data, they are ahead on their analysis and their initial data exploration frees up staff and the privacy budget.
- Mr. Reiter said that, today, developing synthetic data is labor intensive but that researchers are developing routines to make it easier. He described a project that is creating a fully synthetic version of Federal personnel files with longitudinal work histories for the Office of Personnel Management.

Presentation #2: *Bradley Malin, Vanderbuilt University*

- Mr. Malin said that his presentation would zoom in a very specific component of secure computation—secure multiparty computation (SMC). Mathematicians can now take this theory and put it into practice. Advances in cryptography and secret sharing take advantage of higher order mathematical functions and create complex statistical models.
- Mr. Malin gave an example of a project that used three servers to combine over ten million tax records with over 500,000 higher education events. It got the correct answer but it took 384 hours to calculate. It would have taken months to calculate five years ago.

- Microsoft has now taken this process down to seconds. It is possible to take models and put them into a trusted cloud environment, then someone can ask a prediction question and get the answer back from the encrypted system without any data being disclosed.
- Mr. Malin said that the basic mathematics are there to do SMC. The software is available for special projects, but not for regular production uses.
- There are some challenges in the use of SMC. It is not a panacea. Calculations can be secured but the queries will still use up the privacy budget. There is a need for good key management, authentication, and trust in the system and in the data. There's also a need for secure hardware.

Discussion and Questions

- Commissioner Troske asked whether the results from synthetic data would be close enough for policymaking or would satisfy journal editors. Mr. Reiter replied that the really important analyses should be done on the real data. He added that researchers should decide for themselves what percentage of variation from the real results is acceptable. He said that he doesn't have a lot of experience to judge how journal editors would react to using synthetic data with a verification server, but public use data is always mucked up and we close our eyes to that in publishing results. This way is a little more honest.
- Commissioner Groves asked how the Commission could recommend changes that would engender trust. Mr. Malin said he was working on that question with the precision medicine initiative. Trust requires a comprehensive model for determining and assigning protections, including technical approaches and appropriate oversight. He said that you have to consider data utility versus the risk you are willing to tolerate; at no point is the risk zero. Mr. Reiter added it is important to make the case to data subjects that there are real benefits to using their data and asking them about acceptable risk. Mr. Goroff said that the tradeoff is the "epsilon" measure in differential privacy.
- Commissioner Ohm said that the message from these presentations is that we are on the cusp of extraordinary advances. If the Commission hitches its wagon to these techniques for the most sensitive data, there would be a mechanism for more research as the techniques develop. Mr. Malin responded that what is missing is the transition from technical specifications into the real world. He said that convincing people that it is safe to share their data is nontrivial. There needs to be a holistic approach to linking technological experts with legal scholars to see what an actual solution would look like in practice.
- Commission Wallin asked about block chain technology. Mr. Malin said it will help with authentication.

Commission Chair Abraham adjourned the meeting at 12:00 PM.