Privacy Protective Research: Facilitating Ethically Responsible Access to Administrative Data

Daniel Goroff
Jules Polonetsky
Omer Tene[*]

"A secret isn't invalidated by its disclosure, it's defined by its disclosure. What makes a secret a secret is simply the operating instructions that accompany its movement from one person to the next."

– Malcolm Gladwell writing in *The New Yorker* (December 19, 2016) about the work of sociologist Beryl Bellman.

[*] Daniel Goroff is Vice President and Program Director, Alfred P. Sloan Foundation. Opinions or errors expressed here are his own rather than those of the Foundation or its grantees. Jules Polonetsky is CEO of the Future of Privacy Forum (FPF), where Omer Tene is a Senior Fellow. Work on this project benefited from generous support from the Rockefeller Foundation in the form of a Bellagio Center Writer's Residency and from a grant to FPF from the Bill & Melinda Gates Foundation.

Introduction

Researchers in academic institutions and private sector businesses seek to repurpose a variety of sources of organizational data ("administrative data") to pursue projects that promise great societal benefits. Companies collect massive amounts of administrative data through the Internet, mobile communications, and a vast infrastructure of devices and sensors embedded in healthcare facilities, retail outlets, public transportation, social networks, workplaces and homes. They use administrative data to test new products and services, improve existing offerings, conduct research and foster innovation. For example, in an article recently published in the *Journal of Oncology Practice*, a group of Microsoft scientists demonstrated that by analyzing a large sample of search engine queries, researchers could in some cases identify Internet users who were suffering from pancreatic cancer even before those users were diagnosed with the disease.[1]

Increasingly, the collection and analysis of large-scale administrative data drives advances in the measurement and tracking of economic activities. While compelling from a societal standpoint, such data-intensive research projects face formidable transaction costs, including real and perceived legal and ethical challenges. Too often, data remain locked in corporate coffers weighed down by concerns about individuals' privacy, data security, and re-identification risk, as well as by corporate incentives to protect trade secrets and intellectual property and a general inclination to avoid risk by keeping data close. These challenges affect all links of the data chain, including access to administrative data, its protection, analysis, sharing and linking. Yet as Malcolm Gladwell suggests, secrets are best addressed with appropriate information disclosure controls, not with absolute limits on use of information.

The lack of a clear legal framework and ethical guidelines for use of administrative data jeopardizes the value of important research, either because the public perceives it as ethically tainted or because research outputs remain hidden from public view. Concerns over legal impediments and ethical restrictions threaten to diminish productive collaboration between researchers and private sector businesses, restricting funding opportunities and potentially locking administrative data and research projects behind corporate walls. Further complicating matters, companies struggle to define the line between scientific research projects and A/B testing for marketing or product improvement.

Similarly, data held by government agencies as well as entities such as universities, school districts, or other quasi-governmental institutions, are also often inaccessible for important analysis by researchers. Data use benefits can include everything from improving official

[1] John Paparrizos, Ryen W. White & Eric Horvitz, *Screening for Pancreatic Adenocarcinoma Using Signals from Web Search Logs: Feasibility Study and Results*, JOURNAL OF ONCOLOGY PRACTICE, June 7, 2016, doi: 10.1200/JOP.2015.010504.

government statistics using corporate transactions records to rigorously testing causal hypotheses about policy improvements by using randomized controlled experiments. Despite important steps to make data available, which have been advanced by open data movements, by efforts to use data to encourage accountability in education systems, by smart city efforts and other programs, the ability of researchers to access significant government data sets is often limited by a range of concerns, in large part consisting of privacy and security objections.

This paper provides strategies for organizations to minimize risks of re-identification and privacy violations for individual data subjects. In addition, it suggests privacy and ethical concerns can be most effectively managed by supporting the development of administrative data centers. These institutions will serve as centers of expertise for de-identification, certify researchers, provide state-of-the-art data security, organize ethical review boards and support best practices for cleansing and managing data sets. Such centers have demonstrated success in Europe and are emerging in various sectors in the United States.

Given the multiple issues that need to be addressed to manage privacy concerns and the general trust deficit that strains society when research is viewed skeptically and risks are magnified, well-resourced and trusted centers can provide a path forward for different sectors of the economy. Such centers would be part of a data environment with the requisite data governance, accountability and enforcement mechanisms to ensure accuracy, privacy and efficacy of data driven research. Ideally, a network of these centers would develop expertise solving data sharing privacy and ethics problems, which could be leveraged across different centers, further enhancing their capabilities. In an age where data is used as currency and as a raw material of production, establishing new data intermediaries follows logically from the existence of financial intermediaries that have streamlined and regulated market practices for many years.

I. Privacy, Ethics and Evidence-based Policymaking

Policymakers have long grappled with the need for evidence-based decision making. In the words of House of Representatives Speaker Paul Ryan, a lead sponsor of the bipartisan National Commission on Evidence-based Policymaking, "You always hear people in Washington talk about how much money was spent on a program, but you rarely hear whether it actually worked. That has to change."[2]

A policy is "evidence-based" when it is informed and created by a process relying on data-driven research. In many cases, evidence-based policies rely on data about individuals, known in law and policy circles as personally identifiable information (PII). Often the data used by researchers contains private or sensitive information about individuals, including about their health,

[2] Speaker Paul Ryan Press Release, *Evidence-Based Policy Commission Gets to Work*, July 26, 2016, http://www.speaker.gov/press-release/evidence-based-policy-commission-gets-to-work.

education, financial condition, race or ethnicity. How can researchers continue to access and use the data necessary to support evidence-based policymaking while at the same time ensuring individuals' privacy? This paper addresses challenges and opportunities associated with balancing the benefits of evidence-based policymaking against the costs potentially imposed on individual data subjects.

Such challenges arise in particular when researchers access organizational data for a purpose different than the one for which the data were originally collected. We refer to such repurposed organizational data as "administrative data."[3] Research into administrative data is fundamentally different than traditional research on "made" data, typically assembled with hard opt-in consent from the individuals who agreed to participate in an experiment. Over the past few years, the great explosion in volume, velocity and variety of administrative data has led commentators to herald the arrival of "big data" encompassing government, company and other organizational records. Big data of this sort holds great promise for evidence-based policymaking, but at the same time raises privacy and ethical concerns. Often performed without affording individuals with meaningful choice, research should only take place when the benefits of evidence-based policymaking outweigh the costs to individuals.

II. Concerns Limiting Research Access to Administrative Data

This section introduces the main roadblocks impeding the sharing of administrative data between public and private sector organizations and researchers. Organizations considering sharing administrative data about citizens, consumers, patients and employees face privacy concerns and ethical challenges that strain the existing regulatory frameworks and require new, innovative solutions.

Privacy Risk

Traditionally, de-identification was the primary method for enabling researcher access to administrative data while protecting individuals' privacy. Organizations viewed de-identification as a silver bullet allowing them to reap data benefits while at the same time avoiding operational risks and legal requirements. The applicability of legal frameworks governing privacy and data management in the U.S. and EU turned on whether the data involved was personally identifiable or not.[4]

[3] Other common terms are transactional, observational, or "found" data.

[4] Paul Schwartz & Dan Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, 86 NYU L. REV. 1814 (2011); *also see* Jules Polonetsky, Omer Tene & Kelsey Finch, *Shades of Gray: Seeing the Full Spectrum of Practical Data De-identification*, 56 SANTA CLARA L. REV. 593 (2016).

However, in recent years, scientists have repeatedly demonstrated that data sets that were claimed to be de-identified were in fact vulnerable to re-identification attacks. Critics argued that real-world demonstrations of re-identification undermined the validity of de-identification techniques.[5] They cast doubt on the extent to which de-identification remained a credible method for using and deriving value from large datasets while protecting privacy. Importantly, they claimed that de-identification methodologies lacked rigor, failed to reliably distinguish between identifiable and de-identified information and lacked formal proof.[6]

Other experts, particularly practitioners who have long implemented de-identification of health data on the ground, vigorously disputed the claim that anecdotal demonstrations of successful de-identification undermined the validity of de-identification techniques writ large.[7] They noted that in some of the cases, the attacked data sets were not anonymized in a credible manner to start with or that the de-identification attacks were limited in scope, proving only the possibility of re-identifying public figures for whom extensive exogenous data was available. They claimed that despite the theoretical and demonstrated ability to mount such attacks, the likelihood of re-identification for most data sets remained minimal.

The de-identification debate continues, unresolved, and introduces obstacles to research access to critical data. Clearly, re-identification risks will continue to grow as computing technologies become ever faster and the data economy emits more and more data for linkage and analysis over time. Section IV below discusses a number of de-identification solutions and their relative attributes for different types of research.

Ethical Uncertainty

The ethical framework applying to human subject research in the biomedical and behavioral research fields dates back to the Belmont Report, which was drafted in 1976[8] and adopted by the United States government in 1991 as the Common Rule.[9] The Belmont principles were geared towards a paradigmatic controlled scientific experiment with a limited population of human subjects interacting directly with researchers and manifesting their informed consent.

[5] Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1717-23. (2010).

[6] CYNTHIA DWORK & AARON ROTH, THE ALGORITHMIC FOUNDATIONS OF DIFFERENTIAL PRIVACY, FOUNDATIONS AND TRENDS IN THEORETICAL COMPUTER SCIENCE, 9(3-4), pp.211-407. See especially Chapter 8.

[7] Khaled El Emam & Cecilia Alvarez, *A Critical Appraisal of the Article 29 Working Party Opinion 05/2014 on Data Anonymization Techniques*, INT'L DATA PRIVACY LAW (Dec. 13, 2014), http://idpl.oxfordjournals.org/content/early/2014/12/12/idpl.ipu033.full.pdf?keytype=ref&ijkey=K8xdZaj1rw3EzDx.

[8] NATIONAL COMM'N FOR THE PROT. OF HUMAN SUBJECTS OF BIOMEDICAL AND BEHAVIORAL RESEARCH, BELMONT REPORT: ETHICAL PRINCIPLES AND GUIDELINES FOR THE PROTECTION OF HUMAN SUBJECTS OF RESEARCH *(1979), available at http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html.*

[9] HHS, FEDERAL POLICY FOR THE PROTECTION OF HUMAN SUBJECTS ('COMMON RULE'), http://www.hhs.gov/ohrp/humansubjects/commonrule/.

These days, researchers in academic institutions as well as private sector businesses not subject to the Common Rule analyze a wide array of data sources, from massive commercial or government databases to individual tweets or Facebook postings publicly available online. In doing so, researchers have little or no opportunity to engage human subjects directly to obtain their consent or even inform them of the research. The challenge of fitting the round peg of data-focused research into the square hole of existing ethical and legal frameworks may end up determining whether society can reap the tremendous opportunities latent in administrative data of governments and cities, health care institutions and schools, social networks and search engines, while at the same time protecting privacy, fairness, equity and the integrity of the scientific process.

These difficulties weigh down the application of the Belmont Principles to even the academic research that is directly governed by the Common Rule. In many cases, the scoping definitions of the Common Rule are strained by new data-focused research paradigms. For starters, it is not clear whether research on large datasets collected from public or semi-public sources even constitutes human subject research, as data driven research often leaves little or no footprint on individual subjects, such as in the case of search query analysis or automated testing for security flaws.[10]

Unlike experiments with "made" data, which required active recruitment of participants and direct engagement with them to obtain their informed consent, research can be performed on administrative data without any prior communication with data subjects. Indeed, one of the hallmarks of big data research is its promise to find previously hidden and unanticipated correlations in large unstructured datasets.

Not only the definitional contours of the Common Rule but also the Belmont principles themselves require reexamination in this context. The first principle, *respect for persons*, is focused on individual autonomy and its derivative application, informed consent. While obtaining individuals' informed consent may be feasible in a controlled research setting involving a well-defined group of individuals, such as a clinical trial, it is untenable for researchers experimenting on databases that contain the footprints of millions, or indeed billions, of data subjects. The second principle, *beneficence*, requires a delicate balance of risks and benefits that not only respects individuals' decisions and protects them from harm but also secures their wellbeing. Difficult to deploy even in traditional research settings, such cost-benefit analysis becomes daunting in a data research environment where benefits could be probabilistic or incremental, and where the definition of harm subject to constant wrangling between minimalists who reduce privacy to pecuniary terms and maximalists who view any collection of data as a dignitary infringement.

[10] *See, e.g.*, Arvind Narayanan & Bendert Zevenbergen*, No Encore for Encore? Ethical Questions for Web-Based Censorship Measurement*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2665148.

III. Competing Objectives

Before assessing possible solutions for privacy and ethics concerns, it is necessary to discuss on what basis a data-access system for researchers would be judged successful. We suggest three criteria:

a. **Accuracy Objective**: The system should enhance utility by promoting *research reliability*. That is, the results should be robust, unbiased, fully documented, and easily contestable or validated by other researchers. Judgement on this criterion rests with academic editors, referees, and other gatekeepers.

b. **Privacy Objective**: The system should *protect privacy*. That is, the studies undertaken should be publicly defensible with respect to privacy protection. Judgement on this criterion rests with the public generally, but especially with anyone whose data may be under study.

c. **Efficacy Objective**: The system should promote *practical sustainability*. That is, the framework's demands should not be too burdensome in terms of the financial, legal, or bureaucratic support necessary to maintain smooth operations. Judgement on this criterion rests with system funders and managers as well as with data donors and users.

Of course with multiple objectives, pursuing any one objective could adversely affect the others. For example, there is a fundamental trade-off between utility, measured by research accuracy, and privacy. Increasing privacy protection necessarily entails additional data obfuscation or reproducibility obstructions. These two goals are inherently at odds. While technological advances may change and improve the options available, they will not entirely obviate the need for some trade-offs. One consideration when deciding how to supply data for a given research project is, therefore, the appropriate balance between the privacy and utility/accuracy objectives. This will depend on the nature of the study and the type of administrative data along with other factors. Another kind of consideration involves familiar constraints on time, funding and attention. Because supplying data entails certain monetary and non-monetary costs, policymakers must establish priorities among projects.

How should policymakers handle such competing objectives? When considering a request to access data for research purposes, there may be many ethical, practical, or other considerations to take into account. But to the extent privacy *per se* is the focus, we suggest the following three steps:

Step 1: **Admissibility**. First, compile a menu of all the protocols available that would govern how administrative data are released and used. Options could range from de-identification to nondisclosure agreements and data management plans. Next, rank each protocol twice, once according to the privacy it affords and separately according to the accuracy it affords. A protocol will be admissible if: a) there is no other protocol available that would provide at least as good privacy, at least as good accuracy, and do better in one respect or the other (*i.e.*, there is no *pareto superior* solution); and b) both the privacy protection and the accuracy exceed a minimum acceptable threshold.

Step 2: **Appropriateness**. From among the admissible protocols, some might be more appropriate for certain kinds of data than others. For example, Internal Revenue Service information, which is multidimensional and highly sensitive, would typically be studied only under the strictest privacy protections. In contrast, scanner card data about individuals' grocery purchases is already sold and studied by corporations. Before making such datasets available for academic research, there may be little point in manipulating the information other than replacing names with random identifiers.

Step 3: **Affordability**. While the marginal cost of providing data to one more researcher may be small, there are also high fixed costs associated with infrastructure, legalities, bandwidth and documentation. Besides weighing down budgets, each additional research study necessarily leaks some privacy and accuracy, so it is neither practical nor desirable to make a particular dataset containing sensitive information accessible to everyone who might ever be interested. Priority should be given to projects that stand to provide the greatest net-benefit to society by advancing scientific knowledge and understanding. Of course, this cannot be predicted with certainty ahead of time—if we knew what the answers would be in advance of conducting a study, we would not call it research—but academic panels, funders, and researchers already make judgments about such matters regularly.

The minimum admissibility levels may vary from time or place depending on prevailing societal norms. We offer that with regard to privacy, any protocol that could be overcome in a few days by a skilled adversary should be ruled out as providing substandard protection. For example, "white hat" teams could be charged with randomly mounting re-identification attacks to test a system's resilience.

As for minimal utility, some critics argue that exploratory research projects are so inaccurate and unreliable that they should never warrant *any* sacrifice of privacy.[11] Where this is feasible, such preliminary work could be performed on synthetic datasets rather than sensitive ones. Access to actual data should be reserved for confirmatory research that tests a specified hypothesis using a pre-registered analysis plan. This would prevent "p-hacking," that is, dredging through data until a researcher sooner or later finds coincidences that can masquerade as statistical significance.[12]

[11] Andrew Gelman & Eric Loken, *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*, Department of Statistics, Columbia University, 2013, http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.

[12] Uri Simonsohn, Leif Nelson & Joseph Simmons, *P-curve: a key to the file-drawer*, 143(2) J. EXPERIMENTAL PSYCHOLOGY: GENERAL 534 (2014).

IV. Protecting Private Administrative Data

Common research protocols sometimes purport to deliver accuracy or privacy that in fact they do not. One useful way to classify proposed protocols is according to the stage in the research process that data obfuscation occurs:[13] during input, computations, or output. Applying or relaxing restrictions at each stage gives rise to eight possible states (see table). Illustrations of each state follow, starting with the traditional methods whose shortcomings motivated more recent approaches.

Table: At which of three research stages do protocols impose restrictions

	Input	Computation	Output	State / Protocol
1				Open Data
2			X	Data Enclave
3	X		X	Nondisclosure Agreement
4		X		Anonymization
5	X			Randomized Response
6	X	X		Multiparty Computation
7	X	X	X	Fully Homomorphic Encryption
8		X	X	Differential Privacy

1. No Restrictions: **Open Data**

Consider, for example, a researcher who wishes to study state university faculty wages. Some states publish names, salaries, and other information about university employees in downloadable formats. In this case, there are no restrictions on data collecting or sampling, linking or analyzing, or release and reuse.

This is the ideal supported by "open data" advocates. It facilitates accuracy but, of course, not confidentiality. Individuals who care about keeping their salaries private will at least know about such a disclosure policy before they decide to accept a position.

2. Restricted Output: **Federal Data Enclaves**

Consider, for example, a researcher who wishes to study U.S. wage and employment trends more broadly. The most comprehensive datasets are compiled from state and federal administrative

[13] Daniel Goroff, *Balancing privacy versus accuracy in research protocols*, 347(6221) SCIENCE 479-480 (2015).

records under the Longitudinal Employment and Household Dynamics Program (LEHD). Academics can apply for access to personal information at one of several "Research Data Centers" run by the U.S. Census Bureau.[14] If approved, a researcher's "Special Sworn Status" makes him or her subject to prosecution for misuse of private information under the same terms as a government official. Computations typically take place on site, in a "data enclave" that is both physically and digitally disconnected from the rest of the world. To protect against improper disclosures, the resulting research papers must be approved by the Census Bureau before they can be released.

Typically, the Census Bureau checks that any information reported is aggregated so as to obfuscate the identity of individuals. This is akin to pixelating faces in photographs to hide identities, a strategy that works well for unfamiliar people but not so much for those who an adversary has additional information about. Such procedures have produced no known security breaches to date and are gradually becoming less cumbersome. At the same time, this strategy prevents the replication of research results obtained at a data enclave.

3. Restricted Input and Output: Commercial Non-Disclosure Agreements

Companies often draw inferences about online users' salaries and other characteristics based on their web behavior. This method can involve inaccuracies at the input stage because of its indirect, obscure, and irremediable nature, as well as potential sample bias. Researchers rarely gain access to such datasets without signing "non-disclosure agreements" that give a company control over what private or proprietary details may be released.

This arrangement usually precludes replication of the results or reuse of the data. The *American Economic Review*, a premier academic journal whose authors are required to post the data they use, reported having to waive this requirement for nearly half the empirical papers published in 2014 because of non-disclosure agreements.[15]

4. Restricted Computation: Anonymization

In 2014, New York City released "anonymized" data about every taxi ride taken in the city in 2013. Hackers quickly re-identified the data by exploiting weak techniques used to encode the information and by linking with other publicly available datasets. Not only is it now possible to track the earnings of each taxi driver by name, but a researcher can also trace the times, fares, and tips of trips made by celebrities, or map all the precise GPS coordinates on the other end of trips to or from The Hustler Club, for example.[16]

[14] United States Census Bureau, Center for Economic Studies (CES), https://www.census.gov/ces/rdcresearch/.

[15] *See* Liran Einav & Jonathan Levin, *Economics in the age of big data*, 346(6210) SCIENCE 715 (2014).

[16] Neustar Research, *Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset*, Sept. 15, 2014, http://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset

This privacy fail joined several other famous examples of datasets that were initially released with assurances of de-identification, yet were nevertheless linked with other public information to leak private confidences. More generally, Latanya Sweeney has claimed that 87% of the U.S. population can be uniquely identified by just three pieces of information: gender, zip code, and date of birth.[17] Once this information is out, an adversary can search for or purchase the names, social security numbers as well as additional categories of sensitive, sometimes intimate data about individuals in a given dataset.[18]

In most of these cases, while re-identifying individuals from a de-identified dataset alone would be extremely difficult, privacy violations arose by linking the anonymized data with other publicly available datasets. The possibility of such linkage attacks makes safety guarantees of de-identification dependent on knowledge of all the current and future datasets that could ever conceivably emerge to be used to re-identify individuals. Clearly, this is not a practical approach to rigorous privacy protection. That is one of the reasons why—despite creative efforts to aggregate, average, edit, adjust, or otherwise impose obfuscation at the computational stage—experts are now concluding that “sanitizing data doesn’t” and “de-identified data isn’t.”[19]

5. Restricted Input: **Randomized Response**

Consider, for example, a researcher who wants to estimate what percentage of individuals in a group have incomes below the poverty line, yet do so without compromising any individual’s confidentiality. One approach yielding a privacy protective response would be to give each individual in the group a coin to flip without anyone else seeing the outcome. If the coin lands on heads, the individual must truthfully answer whether or not their income level is below the poverty line. If the coin lands on tails, they must flip the coin again. If the second toss is a head, the individual again should answer truthfully; but if the second toss is a tail, he or she should lie—that is, respond that their income level is above the poverty line if it really is not – and vice versa. Under this method, to arrive at a good estimate of the fraction of the group that is *actually* below the poverty line, the researcher would compute twice the number of yes responses minus 0.5.[20]

Even if the researcher knows who answered what, he or she cannot tell who is poor without seeing the result of the coin toss. The usefulness of this technique depends on having participants

[17] http://aboutmyinfo.org/

[18] CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY (Crown Publishing Group, New York, 2016).

[19] Cynthia Dwork, *Differential Privacy: A Cryptographic Approach to Private Data Analysis*, in PRIVACY, BIG DATA, AND THE PUBLIC GOOD: FRAMEWORKS FOR ENGAGEMENT (J. Lane, V. Stodden, S. Bender, and H. Nissenbaum, eds, Cambridge University Press, 2014). *Cf.* KHALED EL EMAM, GUIDE TO THE DE-IDENTIFICATION OF PERSONAL HEALTH INFORMATION (CRC Press, 2013).

[20] If *r* is the reported fraction of yes responses and *p* is the true fraction, then the expected value of *r* is

$$E(r) = \frac{p}{2} + \frac{p}{4} + \frac{(1-p)}{4} \quad .$$

who are all willing to follow instructions and do not cheat or collude. Other privacy-preserving variants are more efficient estimators, but some accuracy is sacrificed in any case.

6. Restricted Input and Computation: Multi-Party Computation

Consider, for example, a researcher who would like to calculate the average salary of a group of individuals without any individual ever communicating his or her own salary. Surprisingly, this can be done quite precisely as long as everyone cooperates. If, for example, there are three individuals in the group, each individual would generate two random numbers and give one number to each of the other two participants. Next, each individual would add the two random numbers he or she generated to his or her own salary, subtract the two numbers he or she was given, and report the result. Adding all of those results and dividing by three would give the average salary, without revealing any individual's number.[21]

Although no one's salary was ever communicated, this protocol does not necessarily protect the privacy of those who participate. If, for example, all but one of the participants collude by using the same method to compute their average salary, that new group could easily deduce the salary of their additional colleague.

Applications of blockchain technology have been proposed to make deviations from agreed protocols more easily detectable.[22] Similarly, more complicated computations can secretly carry out operations beyond just taking averages.[23]

7. Restricted Input, Computation, and Output: Fully Homomorphic Encryption

Banking and other sensitive information routinely travels over the internet without interception. Suppose that an individual could not only send salary information to a researcher using similar or stronger encryption, but that the researcher could perform computations and return the

[21] In other words, let S_i denote the secret salary of person *i,* and let R_{ij} denote the random number generated by person *i* and given to person *j*. Then person *i* reports the result X_i where

$$X_1 = S_1 + (R_{12} + R_{13}) - (R_{21} + R_{31})$$
$$X_2 = S_2 + (R_{21} + R_{23}) - (R_{12} + R_{32})$$
$$X_3 = S_3 + (R_{31} + R_{32}) - (R_{13} + R_{23})$$

When the three equations are added up, all the random numbers cancel out:

$$X_1 + X_2 + X_3 = S_1 + S_2 + S_3 \ .$$

This means that the sum of the reported numbers equals the sum of the salaries.

[22] Guy Zyskind, Oz Nathan & Alex Pentland, *Enigma: Decentralized computation platform with guaranteed privacy*, arXiv preprint arXiv:1506.03471 (2015).

[23] MANOJ PRABHAKARAN & AMIT SAHAI, SECURE MULTI-PARTY COMPUTATION (IOS Press, 2013). For an application to collecting regulatory data, *see* Emmanuel Abbe, Amir Khandani & Andrew Lo, *Privacy-Preserving Methods for Sharing Financial Risk Exposures*, 102:3 AMER. ECON. REV. 65-70 (2012).

encrypted results to the individual without ever being able to decrypt either the inputs or the results. Long thought impossible, methods for such "fully homomorphic encryption" have now been devised.[24] Though practical applications are coming online, many algorithms remain still too slow to operate in real world settings. Proposed protocols would perform regressions and other analyses on encrypted data supplied by survey participants, for example, but only allow the statistics to be decrypted if those participants verify that the calculations have been done properly and to their satisfaction.[25]

Once practical, fully homomorphic encryption could have profound implications for the privacy of numerous applications, including cloud computing, search engines, tax preparation and "personal data lockers." But even effectively hiding inputs to research will not necessarily preserve the privacy of participants, especially if the statistical findings are subject to linkage or "differencing" attacks. As an example of the latter, consider that asking two simple questions—how many employees of a company make more than $1 million in salary, and how many employees of a company who are not the CEO make more than $1 million—could reveal whether or not the CEO makes over a $1 million.[26]

8. Restricted Computation and Output: **Differential Privacy**

Consider, for example, a dataset *D* that contains an individual's personal information in one "row" and another dataset *D'* that is missing that row but is otherwise identical. Two datasets are said to be adjacent if they differ by only in a single row like this. A research protocol would be considered privacy preserving if it could not distinguish between D and an adjacent D'. Alas, it also would not be very useful because too many different datasets would all look the same. But what if the protocol could barely make such a distinction? Specifically, consider the probabilities that it generates a given answer to a given question when applied to D as compared to D'. The ratio of those two probabilities should be as close as possible to one. In fact, the log of that ratio measures the loss of privacy incurred when the protocol answers the given question.[27] If the log of that probability ratio is always less than ε for any adjacent datasets, the protocol is said to provide ε-differential privacy.

Cynthia Dwork and her colleagues not only formulated this definition and showed it captures basic intuitions about privacy loss, but also devised explicit research protocols that provide ε-

[24] Craig Gentry, *Fully Homomorphic Encryption Using Ideal Lattices*, STOC 169-178 (2009).

[25] Adriana Lopez-Alt, Eran Tromer & Vinod Vaikuntanathan, *On-the-Fly Multiparty Computation on the Cloud via Multikey Fully Homomorphic Encryption*, STOC (2012).

[26] It may seem straightforward for a system to simply rule out lines of questioning like this. Provably, however, no algorithm can reliably determine whether a given set of questions that seem to inquire only about statistical aggregates would nevertheless yield answers that, taken together, reveal private information (See C. Dwork, op. cit.).

[27] For a protocol *M* that yields research result *γ = M(D)* when applied to database *D,* Dwork defines the loss of privacy as $L = \ln \frac{\Pr[M(D) = \gamma]}{\Pr[M(D') = \gamma]}$ where *D'* is adjacent to *D.*

differential privacy.[28] In this system, data is held by a trusted curator who only accepts certain questions from an investigator. Calculations are performed behind a firewall, and answers are returned only after injecting a small amount of carefully calibrated noise. It suffices, for example, to draw that noise from a Laplace distribution with parameter 1/ε when responding to a counting query. Other aggregate statistics, including regression coefficients and contingency tables, can be handled similarly. There is a limit on the number of questions that can be allowed, however, since each question depletes the given "privacy budget" by an amount that depends on ε.

Conceptually, choosing a parameter ε for a differentially private protocol determines how a research project will trade accuracy against privacy. The smaller the ε, the less leakage of information – but this comes at a cost of more noise and less accuracy. As a practical matter, how to provide differential privacy by designing, implementing, or combining various algorithms is the subject of intense research.[29] Increasingly, scientists are devising real world systems implementing differential privacy to enable data research with minimal accuracy-privacy losses.

8. Summary

In assessing the privacy protecting research protocols presented above, it is useful to distinguish between those that affect the *collection* of data and those that deal with *computations* and *output*. The former protocols, if built into a data project from the start, can help ease privacy concerns later. Researchers should therefore be aware of techniques like secure multi-party computation or fully homomorphic encryption before setting out to collect data, and are encouraged to use them when appropriate.

In many cases, however, researchers are eager to work with datasets that have already been compiled for other purposes. For such administrative datasets, privacy protecting protocols can only be applied at the computational and output stages. Because so much of the evidence-base for policymaking derives from administrative data, it is worth focusing in more detail on comparing and contrasting post-collection protocols, such as data enclaves, de-identification, and differential privacy.

For these post-collection scenarios, the main privacy concern involves risk of re-identification, since information is not obscured or protected when initially recorded. Most controversial among scientists and policymakers is the practice of de-identification. If performed carefully, de-identification is often considered acceptable in many practical applications. But as discussed above, there have been numerous actual and potential examples of re-identification. In contrast, research protocols that enforce differential privacy effectively rule out the identification of

[28] Cynthia Dwork , Frank McSherry , Kobbi Nissim & Adam Smith, *Calibrating noise to sensitivity in private data analysis*, in THEORY OF CRYPTOGRAPHY CONFERENCE (TCC), Springer, 2006.

[29] One promising example is the Census Bureau's OnTheMap Project, which provides probabilistic differential privacy. A "synthetic database" has been constructed by carefully perturbing and aggregating actual payroll tax records in each state. By querying this dataset, members of the public can receive approximate but quite accurate answers to a large class of counting and geographic questions. *See* http://onthemap.ces.census.gov.

individuals altogether, no matter what post-processing or linkages might ever be attempted. Data enclaves have similarly proven quite safe. Re-identification is unheard of, but the theoretical possibility remains that the release of precise statistics—even if aggregated, averaged, or otherwise sanitized—would allow linkage, differencing, or other attacks to compromise privacy.

At the end of the day, society must decide how to trade-off privacy concerns against research potential.[30] How well society understands, facilitates, and regulates privacy-preserving research will, in turn, determine whether the public will benefit from advances in empirical behavioral and social science or whether that value will flow strictly to those private interests that hold enormous and ever growing stores of sensitive information.

V. Building Institutions

According to economic theory, roadblocks such as high transactions costs can be lowered by forming *institutions*. Nobel Laureate Oliver Williamson in particular emphasized the study of institutions and their governance.[31] One type of institution that already addresses data privacy in research settings is the Institutional Review Board (IRB). However, IRBs, which were initially conceived to address research ethics generally, remain anchored in a paradigm involving direct engagement with individual study participants, which is untenable in the context of administrative data research. New institutions, including recent kinds of research facilities and networks, offer a promising path toward ethical, privacy-aware data sharing between organizations and researchers.

IRBs

In federally funded human subject research, IRBs are the institution responsible for evaluating whether a research project comports with ethical frameworks. Yet, in today's data economy, research of administrative data is increasingly taking place outside of universities and traditional academic settings. With information becoming a raw material for production, organizations are regularly exposed to and closely monitoring vast quantities of administrative data about citizens, consumers, patients and employees. This includes not only companies in industries ranging from technology and education to financial services and healthcare, but also non-profit entities, which seek to advance societal causes, and even political campaigns.

Whether the proposed revisions to the Common Rule address some of these new concerns or exacerbate them is hotly debated. But whatever the final scope of the Common Rule, it is clear

[30] This choice differs from any individual's choice between, for example, taking an expensive vacation or buying a new car. Whereas that decision affects only the specific individual concerned, privacy or validity leakages can affect other researchers and their subjects. Indeed, the reliability of privacy protection and of scientific research are *public goods* which, like national security or lighthouses, are neither excludable nor rival.

[31] OLIVER E. WILLIAMSON, THE MECHANISMS OF GOVERNANCE (Oxford University Press 1996).

that while raising challenging ethical questions, a broad swath of academic research will remain neither covered by the rules nor subject to IRB review. Currently, gatekeepers for ethical decisions range from private IRBs to journal publication standards, association guidelines and peer reviews. A key question for further debate is whether there is a need for new principles as well as new structures for review of academic research that is not covered by the current or expanded version of the Common Rule.[32]

To be sure, privacy and data protection laws provide a backstop in cases involving commercial uses of data, setting boundaries like consent and avoidance of harms. But in many cases where informed consent is not feasible and where data uses create both benefits and risks, legal boundaries are more ambiguous and rest on vague concepts such as "unfairness"[33] (in the U.S.) or the "legitimate interests of the controller" (in Europe).[34] An uncertain regulatory terrain could jeopardize the value of important research, which could be perceived as ethically tainted or become hidden from the public domain to prevent scrutiny.[35] Concerns over data ethics could diminish collaboration between researchers and private sector entities, restrict funding opportunities, and lock research projects in corporate coffers contributing to the development of new products without furthering generalizable knowledge.[36]

To address ethical questions about corporate data research, Ryan Calo foresaw the establishment of "Consumer Subject Review Boards."[37] Calo suggested that organizations should "take a page from biomedical and behavioral science" and create small committees with diverse expertise that could operate according to predetermined principles for ethical use of data.[38] The idea resonated in the Obama White House's legislative initiative, the Consumer Privacy Bill of Rights Act of 2015, which required the establishment of "Privacy Review Boards" to vet non-contextual data uses.[39]

[32] National Research Council, Proposed Revisions to the Common Rule for the Protection of Human Subjects in the Behavioral and Social Sciences (Committee on Revisions to the Common Rule for the Protection of Human Subjects in Research in the Behavioral and Social Sciences. Board on Behavioral, Cognitive, and Sensory Sciences, Committee on National Statistics, Division of Behavioral and Social Sciences and Education, Washington, DC, The National Academies Press, 2014).

[33] FTC Policy Statement on Unfairness, Appended to International Harvester Co., 104 F.T.C. 949, 1070 (1984). See 15 U.S.C. § 45(n).

[34] Article 29 Working Party, WP 217, Op. 06/2014 on the Notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC, Apr. 9, 2014, http://ec.europa.eu/justice/dataprotection/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf.

[35] The Common Rule's definition of "research" is "a systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to *generalizable* knowledge." (Emphasis added).

[36] Jules Polonetsky, Omer Tene, & Joseph Jerome, *Beyond the Common Rule: Ethical Structures for Data Research in Non-Academic Settings*, 13 Colo. Tech. L. J. 333 (2015).

[37] Ryan Calo, *Consumer Subject Review Boards: A Thought Experiment*, 66 Stan. L. Rev. Online 97 (2013), http://www.stanfordlawreview.org/online/privacy-and-big-data/consumersubject-review-boards.

[38] Stan. L. Rev. Online Symposium Issue*, Privacy and Big Data: Making Ends Meet, September, 2013, http://www.stanfordlawreview.org/online/privacy-and-big-data; also see stage setting piece*, Jules Polonetsky & Omer Tene, *Privacy and Big Data: Making Ends Meet*, 66 Stan. L. Rev. Online 25 (2013).

[39] CONSUMER PRIVACY BILL OF RIGHTS §103(c) (Administration Discussion Draft 2015), *https://www.whitehouse.gov/sites/default/files/omb/legislative/letters/cpbr-act-of-2015-discussion-draft.pdf*.

In Europe, the European Data Protection Supervisor announced the creation of an Advisory Group to explore the relationships between human rights, technology, markets and business models from an ethical perspective, with particular attention to the implications for the rights to privacy and data protection in the digital environment.[40] A number of organizations are working to advance new types of review boards, but these efforts are still nascent. Supporting the development of these entities would be a valuable to ensure ethical oversight over research data that is not subject to IRBs.

In *Beyond IRBs: Ethical Guidelines for Data Research*, Omer Tene and Jules Polonetsky discussed possible guidelines for industry-wide or corporate IRBs.[41] They offered suggestions as to what projects would be subject to review; how a review board would be structured; when review would be conducted; and which principles would apply. They offered that even research initiatives that are not governed by the existing ethical framework should be subject to clear principles and guidelines. They noted, "As the field of data ethics develops and grows, policymakers should seek to harmonize the principles and procedures governing academic research, corporate research, and corporate product development using personal data, as well as research projects affecting individuals in real ways."[42]

Administrative Data Research Facilities

Because IRBs primarily examine research ethics before a project starts, they can recommend a broad range of pre-collection and post-collection protocols for data privacy. Yet as noted above, more and more information is compiled by agencies, offices, or companies for purposes other than research. Such administrative data, which comprises a by-product of standard enterprise activity, stands in contrast to the results of surveys, lab experiments and field trials that are designed by academics and approved by IRBs to test specific hypotheses.

If handled properly, administrative datasets can nevertheless be hugely valuable for research. Yet using data for secondary purposes, such as research or program evaluation, can entail high transaction costs. When faced with a collective action problem with high transaction costs such as this, economists seek institutional solutions—especially ones that engender trust. We therefore suggest setting up a network of trustworthy data intermediaries called Administrative Data Research Facilities (ADRFs) to allay transaction costs that impede data driven research. Each ADRF would develop expertise solving data sharing privacy and ethics problems for a given industry sector. In addition to setting forth procedures to enhance accuracy, privacy and efficacy

[40] European Data Protection Supervisor*,* Ethics Advisory Group, Dec. 3, 2015, *https://secure.edps.europa.eu/EDPSWEB/edps/site/mySite/Ethics*. See also
Curtis Naser, *The IRB Sledge-Hammer, Freedom and Big-Data*, https://bigdata.fpf.org/papers/the-irb-sledge-hammer-freedom-and-big-data/.

[41] Omer Tene & Jules Polonetsky, *Beyond IRBs: Ethical Guidelines for Data Research*, 72 WASH. & LEE L. REV. ONLINE 458 (2016).

[42] Id.

at the data input, computation and output stages, an ADRF would institute accountability measures for auditing and monitoring compliance with data sharing rules.

While an ADRF could go a long way to lowering transaction costs for both data suppliers and users, setting up multiple ADRFs is only the first phase of a more comprehensive plan. The next phase would involve organizing an Administrative Data Research Network (ADRN) whose members comprise ADRFs committed to sharing best practices and high standards. Such an association of data intermediaries would, for example, create working groups on topics like: compliance and legal matters; researcher credentialing; ADRF accreditation; data security; systems and operations; private and proprietary data protections; government and public relations; research and reproducibility standards; corporate relations and contracting; data interfaces and linking; and more. Any researcher or facility found to be jeopardizing the ADRN's reputation for trustworthiness would lose their status, privileges, and data access, presenting a forceful deterrent to misbehavior.

A similar model has already been tried and tested in the UK. According to a 2012 report titled *The UK Administrative Data Research Network: Improving Access for Research and Policy,* an interagency Administrative Data Taskforce headed by Sir Alan Langlands recommended to "set up an independent organization that would help social and economic researchers gain access to administrative data in a safe and lawful way." The goal was to create a single point of access for researchers who want to use administrative data; an institution that would screen researchers, make sure they are properly trained to handle potentially sensitive information, provide safe rooms for the researchers to access data in, and take on the task of negotiating for data access as well as find safe ways to link different datasets together without compromising the privacy of any individual.

Consequently, the UK Economic and Social Research Council decided to fund an ADRN to facilitate access to government sourced administrative data. The ADRN not only provides de-identification services – using a trusted third party to link administrative datasets without leaking identifying information about the individuals involved – but also helps researchers prepare their proposals before they go to an IRB ("Approvals Panel"), which determines whether their research project is lawful, ethical, feasible, of high scientific merit and of potential benefit to society. The ADRN then provides a secure environment where researchers can access the linked de-identified data. Before researchers can remove their final results from the secure environment, ADRN staff check that their findings are relevant to the approved research project and do not contain directly identifying personal information that would allow any individual to be identified.

The ADRN-UK comprises an Administrative Data Service, which coordinates the network, as well as four Administrative Data Research Centers, one in each country in the UK: England, led by the University of Southampton; Northern Ireland, led by Queen's University Belfast; Scotland, led by the University of Edinburgh; and Wales, led by Swansea University. Other parties to the network include national statistics authorities, government departments and agencies (the data custodians), the ESRC (the funder) and the UK Statistics Authority, which leads the ADRN Board that reports directly to Parliament.

The ADRN-UK focuses especially on government datasets, and is especially notable for its facilities and procedures that can link such datasets in straightforward and timely but secure ways. Laws and traditions vary from country to country, of course, but there are also examples and potential partners set up by governments in Germany, Denmark, and elsewhere.

In the U.S., the Bureau of the Census runs a network of Federal Statistical Research Data Centers (FSRDC) where data from a dozen different agencies can be made available to qualified researchers. Some states and localities participate in similar schemes. Access to data is carefully regulated before, during, and after any calculations are performed. At the federal level, researchers must obtain Census Bureau Special Sworn Status that makes them subject to prosecution for breaking confidentiality restrictions.[43] On the one hand, this limits privacy risks; so much so that there has never been a reported breach of confidentiality. On the other hand, this also limits the replicability and reliability of research results since it is nearly impossible for anyone other than the original investigator to verify the accuracy of the data or the calculations.

The U.S. also has several data centers and archives that are not run directly by the government but serve the needs of researchers interested in evidence-based policy. For example, the National Opinion Research Center (NORC) at the University of Chicago runs a "data enclave" as one of its signature data governance solutions. This is a system that allows the sharing, among a closed community of researchers, of datasets that are too sensitive to be shared broadly.[44] In addition to archiving, curating, and indexing the data, NORC provides extensive privacy protection by restricting access. The Inter-University Consortium for Political and Social Research (ICPSR) is another example of a data intermediary.[45] "Virtual data enclaves" can also be used to enable remote access by qualified researchers.[46]

In contrast to these institutions, which have a broad purview, there are examples of data intermediaries that focus on a given industry sector. In the U.K., for example, the Consumer Data Research Centre based at the University College London works specifically with high street retailers; the Urban Big Data Centre based at the University of Glasgow works with smart cities. Because researchers and managers associated with each such facility know the data and the data providers, they build up a reputation for trust and expertise as data intermediaries in their space. The Administrative Data Service provides an overall framework of standards, procedures, training, and other forms of support for their operations.

Sector-specific data specialists already exist in the U.S. as well. For example, researchers interested in exploring supermarket scanner data can consult the Kilts Center at Chicago Booth,

[43] *See* discussion *supra* Part IV(2).

[44] *See* NORC at the University of Chicago Data Enclave, http://www.norc.org/Research/Capabilities/Pages/data-enclave.aspx.

[45] *See* ICPSR, Data enclaves, http://www.icpsr.umich.edu/icpsrweb/content/ICPSR/access/restricted/enclave.html.

[46] John Abowd & Julia Lane, *New approaches to confidentiality protection: Synthetic data, remote access and research data centers*, in INT'L WORKSHOP ON PRIVACY IN STATISTICAL DATABASES 282-289 (Springer, Berlin Heidelberg, June 2004).

which acts as an ADRF for this category of administrative data. Those interested in university data can turn to the Institute for Research on Innovation and Science (IRIS) at the University of Michigan. To study cities' administrative data, experts, datasets, and systems meet at the Center for Urban Science Progress (CUSP) at NYU. With these institutions in place, a researcher who needs data held an ADRF does not have to renegotiate a new data access agreement with original data producers every time.

Yet the U.S. landscape still lacks facilities in numerous industry sectors. And American ADRFs could benefit from an ADRN-US, modeled after the UK network and charged with accrediting ADRF-members and providing a staff of experts in law, ethics, technology, research methodology, and government relations to support them. Such coordinated support for data intermediaries would facilitate evidence-based policy making.

Ultimately, at least one ADRF would cover each major data-producing sector in the economy, including, for example: online retailers; traditional retail chains; credit card companies; financial services; automotive companies; payroll processors; as well as all levels of state, local, and federal agencies. An ADRF with a given specialization would negotiate a standard Data Use Agreement with each industry member on behalf of researchers. The ADRF would then be responsible for preparing each dataset for study by creating the necessary documentation, metadata, basic data hygiene and versioning, approximate summary statistics, citation information and archiving services. This would include scrutinizing data sets to flag and minimize selection bias such as excluding vulnerable populations or reflecting preexisting societal biases.[47] For credentialed researchers with sound research plans, access to the data would be granted using admissible privacy protecting protocols as appropriate.

While it is natural to think of an ADRF as an institution hosted at a university, other organizations could play that role, including NGOs, corporations, or government agencies—especially ones like the Census Bureau that are not only actively engaged in generating data but also in facilitating its use by independent researchers. Of course, this will require establishing robust and transparent accountability mechanisms to foment trust in entities that lack the institutional tradition of academic establishments. Having received funding to establish a clearinghouse for government and other administrative data, for example, the Census Bureau would be an important partner. This would be especially significant since all network members would abide by explicit and streamlined procedures for sharing, linking, and protecting datasets.

Whereas FSRDC system and the ADRN-UK system were created by government, in the U.S., sector-specific ADRFs and the ADRN-US could coordinate closely with federal agencies without necessarily depending on them for initiation, governance, or even financing. A s is often the case in the U.S., there is an important role for private philanthropy. Indeed, some of the existing ADRFs have been established by grants, and some ADRFs are beginning to develop long-term sustainability plans to make them self-sufficient. Both IRIS in Michigan and the CDRC in the U.K.,

[47] *See* Madelyn Rose Sanfilippo, *An Unequal Information Society: How Information Access Initiatives Contribute to the Construction of Inequality*, Ph.D. Dissertation, Indiana University, 2016, https://eric.ed.gov/?id=ED571243.

for example, deliver services that data donors are willing to pay for, such as data cleansing, linking, archiving, and analysis. In this way, the private benefits that ADRFs supply can help pay for the public benefits they provide by facilitating evidence-based policymaking.

Conclusion

To promote policymaking based on evidence derived from private or public administrative data, at least four stakeholder groups must cooperate: government agencies, private sector corporations, independent researchers and philanthropic funders. Academics such as Gary King, Director of the Institute for Quantitative Social Science at Harvard University, have long discussed the need for a “grand bargain” among these groups with respect to administrative data. Lowering the transactions costs associated with studying data would benefit all. Led by the Census Bureau, government agencies at all levels are beginning to work more seamlessly with academics. At the same time, private sector corporations continue to encounter repeated and disparate data requests from a steady stream of academics or government agencies. In many cases, businesses and government agencies already realize that outside researchers can add enormous value by cleaning, compiling, archiving, and analyzing administrative data in ways that could not be performed internally.

Researchers, in turn, have been slow to organize facilities or networks to deal with access to administrative data. Certain researchers who have gained such access through personal connections or otherwise may not be eager to share the wealth and are less concerned about the difficulties their colleagues face when trying to access administrative data. More generally, academics, who are laser-focused on what is needed to publish their next paper, spend little time on finding ways to cooperate with colleagues to create more streamlined processes for granting access to data or generating reliable evidence. Journal editors, while troubled by reproducibility requirements, continue to offer waivers liberally.

At their best, philanthropies can provide incentives to help solve such collective action problems. Some already do so through support of ADRFs, which in the long run can be structured to generate self-sustaining revenue. Besides launching more ADRFs, foundations should work on establishing governance mechanisms, offices, working groups, and membership criteria for an ADRN.

In the end, there is no single solution to the problems limiting researchers’ access to private data. But the technical and institutional solutions discussed above would greatly facilitate access and reduce transaction costs. Achieving the potential for data science to improve policymaking while also protecting privacy is well within reach.