INFORMATION ACCESS

Balancing privacy versus accuracy in research protocols

Restricting data at collection, processing, or release

By **Daniel L. Goroff***

Designing protocols for research using personal data entails trade-offs between accuracy and privacy. Any suggestion that would make empirical work less precise, open, representative, or replicable seems contrary to the needs and values of science. A careful reexamination has begun of what "accuracy" or "privacy" should mean and how research plans can balance these objectives.

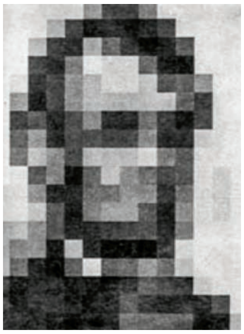
Attitudes toward research that analyzes personal data should depend both on how well the protocol generates valuable statistics and on how well it protects confidential details. There is always some risk of a leak, so it hardly makes sense to support a study incapable of producing valid and robust results. It would also be reassuring to know that the same or better scientific reliability could not be obtained via some other protocol that provides more privacy protection.

POLICY

PARSING PROTOCOLS. A given research plan can be assessed by comparing it along accuracy and privacy dimensions with other potential protocols. Many purport to deliver more than they do on either score. Research on even a simple population statistic—say, average salary—involves collecting, processing, and releasing data. Various protocols can introduce obfuscation, or not, at any combination of these three stages. Eight examples follow, starting with traditional methods whose strengths and shortcomings motivate more recent approaches.

Open data. Suppose a researcher wishes to study faculty wages. Some U.S. states publish names, salaries, and other information about public university employees. There are no restrictions on data collecting and sampling, linking and analysis, or release and reuse. This is the ideal supported by "open data" advocates. It facilitates accuracy but not confidentiality. People who care about keeping their pay private need to be aware of such policies before they decide to take a position.

Data enclaves for federal data. Suppose a researcher wishes to study U.S. wage and employment trends more broadly. Academics can apply for access to Research Data Centers run by the U.S. Census Bureau (*1*). Approved researchers are subject to prosecution for misuse of private information under the same terms as government officials. Computations typically take place in a data enclave disconnected from the rest of the world. Papers must be reviewed by the Census Bureau before they can be released, mainly to ensure that information is aggregated or obfuscated enough to protect individuals' privacy. This is akin to how pixelating the photo of an unfamiliar face renders it unidentifiable (see image). Federal enclaves have produced no known security breaches and are becoming less cumbersome to use, but replication is problematic.



Aggregation and averaging, as in the pixelated image, can hide identities. Linkage with other information, like the familiar portrait of President Lincoln on U.S. currency, can undermine obfuscation. Image from (*16*).

Nondisclosure agreements for online business data. Suppose a researcher wishes to study the relation between salary and other behaviors. Online companies often ask or draw inferences about users' income, usually for unstated purposes. Researchers who seek such data rarely gain access without signing a nondisclosure agreement (NDA) that gives the company control over what details may be released. Arrangements like this usually protect proprietary interests of businesses rather than privacy interests of customers. NDAs can also preclude replication of results or reuse of data (*2*).

Anonymization of administrative data. Suppose a researcher wishes to study earnings of cab drivers. New York City recently released "anonymized" data about every taxi trip taken in 2013. These data were reidentified by exploiting weak encoding and by linking with other publicly available data sets. Not only is it possible to track earnings of each cabbie by name, one can also map GPS coordinates on either end of each ride and even deduce the trip times, fares, and tips of certain celebrities (*3*).

This joins many other examples of data sets that were released with assurances that they had been scrubbed of any personally identifiable information but were easily linked with other public information to yield private confidences, including health records of Governor Weld (*4*) and movie rental histories of Netflix users (*5*). Sweeney even suggests that a vast majority of Americans can be uniquely identified using only zip code, sex, and birthday data (*6*). So anonymization can reduce accuracy while failing to protect private information against "linkage attacks." In other words, "sanitizing data doesn't" and "deidentified data isn't" (*7*).

Randomized response in survey data. A researcher may want to estimate what percentage of a group lives in poverty and so gives each person a coin to flip, together with these instructions: "If it lands heads, truthfully answer yes or no to the question 'Is your income below the poverty line?' If it lands tails, flip again. If the second toss is a head, answer truthfully, but if the second toss is a tail, then lie by giving the answer opposite to what is true." Twice the fraction of yes responses minus one-half provides a good estimate of the actual fraction sought (*8*).

Even if you know who answered what, that does not tell you who is impoverished. The usefulness of this technique depends on having lots of participants, all of whom follow instructions. There are privacy-preserving variants that provide more efficient estimators, but some accuracy is sacrificed in any case.

Multiparty computation for reporting sensitive data. Suppose a researcher would like to calculate the average salary of a group, but without anyone ever communicating her own. Say there are three people. Each generates two random numbers and gives one to each of the other two participants. Everyone then adds the two random numbers she generated to her own salary, subtracts the two numbers she was given, and reports the result. All the random numbers cancel when these three results are added, so their sum equals the sum of the salaries. Dividing by three gives the average. Special and more convoluted computations can secretly carry out operations beyond just taking averages (*9*).

Although no individual's salary was communicated, this protocol does not necessarily keep participants from finding out one another's personal information. If, for example, all but one collude by using the same

*Vice President and Program Director, Alfred P. Sloan Foundation, New York, NY 10111, USA. E-mail: goroff@sloan.org. *Opinions or errors are the author's own rather than those of the foundation or its grantees.*

method to compute their average salary, that group could deduce what the salary was of their original colleague. The protocol delivers completely accurate results but can also endanger privacy.

Fully homomorphic encryption of cloud data. Suppose a researcher wishes to study salaries using bank data. People routinely and confidently send such information encrypted over the Internet. Financial institutions decrypt the messages and perform calculations. But what if the bank or other data receiver not only could perform calculations without ever decrypting the private information but also could return encrypted answers that only the sender could decode? Long thought to be impossible, "fully homomorphic encryption" methods have recently been devised (*10*) to do just that. Most algorithms are still too slow for practical applications. Proposed protocols could, however, analyze a population's encrypted data but only allow statistics to be decrypted if participants verify that calculations have been done to their satisfaction (*11*).

By giving control over their data to potential subjects rather than to researchers, such techniques jeopardize plans for replicability and reuse, as well as for representative or even adequate sampling. Supposing there are results to release, it may still be possible for a researcher to violate the privacy of individuals who participate in the study. Any protocol that allows exact counts of subpopulations is vulnerable to a "differencing attack," for example. To find out whether the CEO of a company earns more than $1 million, just make two simple inquiries: how many employees earn over $1 million in salary, and how many who are not the CEO earn over $1 million. It may seem straightforward to rule out lines of questioning like this. Provably, however, no algorithm can reliably determine whether a given set of questions that seem to ask only about statistical aggregates would nevertheless have answers that, taken together, reveal private information (*12*).

Differential privacy for curated data. Consider a data set D that contains my personal information and another data set D' that is missing my data but otherwise the same. A research protocol would be privacy-preserving if it could not distinguish between D and an adjacent D'. It also would not be very useful. But what if the protocol could barely and rarely make such a distinction? Consider the probabilities that a certain methodology generates a given answer to a given question when applied to D as compared with D'. The ratio of those two probabilities should be as close to one as possible. The log of that ratio measures the loss of privacy incurred when the protocol answers the given question. If the log is always less than ε for any adjacent data sets, the protocol provides ε-differential privacy.

Dwork, McSherry, Nissim, and Smith formulated this definition, showed it captures basic intuitions about privacy, and devised research protocols that provide ε-differential privacy (*13*). Data are held by a trusted curator who only accepts certain questions from the investigator. The curator performs calculations behind a firewall but only returns answers after adding a small amount of carefully chosen noise. It suffices, for example, to draw noise from a Laplace distribution with parameter $1/\varepsilon$ when responding to a counting query. There are limits on the type and number of questions allowed, as each could deplete a privacy budget by as much as ε.

Privacy is breached when "secure" data can be linked with publicly available data.

Choosing ε for a differentially private protocol determines how the research will trade accuracy against privacy. The smaller ε is, the less leakage of information but at the cost of more noise. One promising application is the Census Bureau's OnTheMap Project (*14*). Payroll records in each state have been carefully perturbed and aggregated to create a "synthetic database." The public can query that database to receive approximate, but quite accurate, answers to a large class of counting and geographic questions (*15*).

PICKING PROTOCOLS. Setting aside administrative, financial, legal, or institutional factors that do not bear directly on accuracy and privacy, some basic suggestions for comparing protocols are clear. Potential subjects considering participation in a study should ask if there is another protocol that would yield at least as reliable scientific results while offering better privacy protection. Researchers designing studies should ask if the protocols will actually deliver the levels of accuracy and privacy anticipated.

Funders or others deciding on whether a research plan moves forward should also ask about the broader incentive effects of using a particular methodology. Accuracy and privacy achieved by a protocol are public goods and, hence, subject to free-rider problems. To increase chances of curing a disease, say, every patient wants accurate research but preferably using other people's data rather than their own. To decrease chances of linkage or other attacks, every researcher wants all other projects held to high thresholds of privacy protection but preferably not their own.

Policy-makers reviewing U.S. legislation should also ask about laws like FERPA, HIPAA, or the Privacy Act of 1974 that govern data collection and use by educators, health care providers, or federal officials, respectively (*17*). Do these actually promote accuracy and privacy, or are they based on outmoded ideas about anonymization and identifiability, for example? Unlike other countries, the United States has no legislation specifically regulating or facilitating the use of personal information by academic researchers.

Critically, society as a whole must also ask about promising and threatening aspects of new information technologies. How well society balances the accuracy and privacy of research protocols will determine the extent to which "big data" either allows everyone to benefit from advances in empirical science or only those private interests who hold enormous and growing stores of sensitive information about us all. ■

REFERENCES AND NOTES

1. RDC Research Opportunities, Center for Economic Studies (CES), https://www.census.gov/ces/rdcresearch/.
2. L. Einav, J. Levin, *Science* **346**, 1243089 (2014).
3. Riding with the Stars, Passenger Privacy in the NYC Taxicab Dataset; http://research.neustar.biz/2014/09/15/.
4. L. Sweeney, *J. Law Med. Econ.* **25**, 98 (1997).
5. A. Narayanan, V. Shmatikov, *Proc. of IEEE Symp. on Security and Privacy* (IEEE, 2008), pp. 111–125.
6. How Unique are You? http://aboutmyinfo.org/.
7. C. Dwork, in *Privacy, Big Data, and the Public Good*, J. Lane *et al*., Eds. (Cambridge Univ. Press, Cambridge, 2014), pp. 296–322.
8. $E(r) = p/2 + p/4 + (1 - p)/4$, where r is the fraction reporting "yes" and p is the true proportion.
9. M. Prabhakaran, A. Sahai, *Secure Multi-party Computation* (IOS Press, Amsterdam, 2013).
10. C. Gentry, *STOC '09: Proc. of the 41st ACM Symp. on Theory of Computing* (ACM, New York, 2009), pp. 169–178.
11. A. López-Alt, E. Tromer, V. Vaikuntanathan, *STOC '12: Proc. of the 44th ACM Symp. on Theory of Computing* (ACM, New York, 2012), pp. 1219–1234.
12. C. Dwork, in *Automata, Languages and Programming* (Springer, New York, 2006), pp. 1–2.
13. C. Dwork, F. McSherry, K. Nissim, A. Smith, *Proc. 3rd Theory of Cryptography Conference* (*TCC*) (Springer, New York, 2006), pp. 265–284.
14. A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, L. Vilhuber, *Proc. 24th IEEE International Conf. on Data Engineering* (*ICDE*) (IEEE, 2008), pp. 277–286.
15. On the Map, http://onthemap.ces.census.gov.
16. L. D. Harmon, B. Julesz, *Science* **180**, 1194 (1973).
17. The Family Educational Rights and Privacy Act (FERPA) and the Health Insurance Portability and Accountability Act (HIPAA).

10.1126/science.aaa3483

PHOTO: DIBROVA/THINKSTOCK