Association for Computing Machinery
acm Technology Policy Council
TechBriefs



# Data Privacy Protection

by Micah Altman, Aloni Cohen, and Kobbi Nissim

## PROBLEM

***Proliferating data collection, advanced algorithms, and powerful computers have made it easy to piece together information about individuals' private lives from public information as controls over information privacy become increasingly ineffective.***

## POLICY IMPLICATIONS

- Proliferating data collection, use, and publication present rapidly accumulating risks of private information disclosure that require regulation to mitigate.
- Traditional approaches to anonymization, deidentification, and disclosure control fail to protect information at its current scale and are entirely unable to deal with new ways of utilizing information, such as generative AI.
- Inherently imperfect legal and technical solutions must balance individuals' and stakeholders' needs for data privacy and accuracy.

## DATA PRIVACY PROTECTION: BY THE NUMBERS

| | |
|---|---|
| **4.8** | Billions of individuals worldwide whose personal information is for commercial sale.[1] |
| **462** | Projected 2031 total revenue of worldwide "data broker" market in billions of U.S. dollars.[2] |
| **156** | Comparable revenues for that global market in 2012 in billions of U.S. dollars.[2] |
| **81** | Percentage of Americans concerned about how companies use their data.[3] |
| **38.4** | Percentage of 2010 U.S. Census responses that can be reidentified with high confidence.[4] |
| **77** | Percentage of Americans concerned that they don't understand what the government does with their data.[3] |
| **95** | Percentage of individuals in a study of 1.5 million Europeans who could be uniquely characterized from just four random location records without using other personal data.[5] |
| **10,000** | Minimum number of data types available for purchase from a major data broker.[1,6] |
| **6.14** | Millions of person-years estimated to be required annually to read all of the privacy policies for web services used and sites visited by U.S. internet users.[7] |
| **350** | Number of consumer privacy bills considered by U.S. state legislatures in 2023.[8] |
| **3,519** | Number of computer science papers on privacy posted to arXiv in 2023.[9] |

ILLUSTRATION: © GETTY IMAGES/TRAITOV



ACM's Technology Policy Committees provide cutting-edge, apolitical, non-lobbying scientific information about all aspects of computing to policy makers in the United States and Europe. To tap the deep expertise of ACM's 100,000 members worldwide, contact ACM's Global Policy Office at acmpo@acm.org or +1 202.580.6555.

acm Technology Policy Council

In an era of ubiquitous data collection and massive computing capacity, society faces unprecedented challenges to protecting information privacy. Technical advances in the fields of computer science, statistics, and information science both starkly illuminate privacy risks and have the potential to provide a foundation for privacy protection that is more systematic and effective.

## Privacy Risks Proliferate with Increased Data Collection and Analysis

Foundational research conducted in computer science and statistics over the past two decades has revealed a truth that challenges every data analysis. Under what has come to be called the "fundamental law of information recovery," every useful data analysis invariably leaks some information, and these leaks accumulate across computations. Too many independent analyses, even if highly aggregated, will inevitably reveal the underlying data itself.[10]

***We face unprecedented challenges to protecting information privacy.***

Further, research across a range of fields underscores that the predictability of many human behaviors in general can be used to distinguish specific individuals. For example, in a dataset of location information for 1.5 million representative people in a small European country, just four random location records uniquely characterized 95% of individuals without using direct personal identifiers—even when location was measured only once every hour and with relatively low location precision.[10] More generally, individuals may be uniquely distinguished through the measurement of a small number of behaviors observable across a variety of scales, including typing rhythm, walking patterns, shopping habits, writing style, and movie preferences.[11]

Additionally, increases in the frequency, precision, and other dimensions of data measurement have a high potential to create new and substantial privacy risks. For example, a niche mobile app that occasionally uses coarse location information to provide localized weather forecasts poses lower risks but could still use collected data to behaviorally distinguish the unique owner of the phone. A similar but widely adopted weather app collecting more frequent and granular location information would represent a much greater threat to personal privacy, potentially enabling private information to be inferred about an individual user's employment, exercise habits, health, and associations, or even potentially facilitating systematic surveillance.[6] Generally, "small privacy risks can multiply unexpectedly, and potentially catastrophically unless protections are explicitly implemented to limit cumulative risk."[12]

## Anonymization Alone Cannot Adequately Protect Personal Privacy

As a general concept, individual information privacy encompasses both what others may learn about a person as a direct or indirect result of observations of or interactions with them and what others can do with that information.

The term "anonymization" is used colloquially to refer to a collection of overlapping but fundamentally different concepts, which are often only weakly protective. For example, in some common legal frameworks, anonymization is defined in terms of the removal of names or other specific data elements. It also can refer to the application of specific data transformations or to an expert's determination that the data cannot readily be linked to an individual. Modern privacy theory, however, calls many of these techniques into question. Computer science has converged on an approach to anonymization that is more coherent and personally protective: Analyses (or other computations) are anonymized to the extent that they guarantee that (almost) no information specific to any individual is revealed as a result of the inclusion of their data in that analysis.

Such an approach to anonymization mitigates harm to individuals that may result from the use of their own specific information. However, anonymization in general cannot, even in theory, be used to ensure that algorithms based on personal data will be secure, protect groups of people from discrimination, protect information property rights, be explainable, be reasonable, or be used for legitimate purposes.

***Too many independent analyses of a dataset will inevitably reveal the underlying data itself.***

In practice, current legal and technical anonymity safeguards are concerned primarily with the protection of individuals. Current formal anonymization methods alone do not provide substantial protection against inferences about groups, such as families, tribes, targeted organizations, or marginalized communities.

## Advanced Privacy-Protecting Technologies Must Be Baked-In and Widely Implemented

Twenty years ago, the discovery of the fundamental law of information recovery[10] heralded a profound new challenge to privacy protection. Recognition of the limits of traditional approaches to protection—such as deidentification and simple aggregation, followed by release and benign neglect of the resulting data—subsequently gained significant and widespread currency in both legal and technical scholarship.[13]

As the traditional approaches to protecting individual privacy have proved increasingly vulnerable to attack and prone to failure, modern privacy-enhancing technologies (PETs) have evolved to provide potentially more reliable and adaptable approaches to data protection. PETs offer new approaches to controlling information risks to individuals from data use, including inference. While these new technologies sometimes require substantially more computing power than older protection technologies, PETs have the capability of providing protection that is more flexible, precise, and reliable.

Differential privacy, a formal mathematical framework, is generally considered to be the state of the art for strong anonymization. It provides provable quantifiable control over *inferential risk* (i.e., how much others can learn about any individual as a result of the inclusion of their data in an analysis). Other PETs, such as homomorphic encryption and secure multiparty computation, limit what analyses may be performed directly on the data but do not directly protect anonymity. Technical controls on use, such as personal data stores, access limitations, and logging, can limit who accesses data directly and increase accountability for data use. They do so by facilitating the enforcement of usage policies within a computer system and by supporting usage audits.[14]

***Strong anonymization mitigates harm.***

Strong privacy does not occur by accident, nor should it be implemented as an afterthought. Best practices include designing controls spanning the entire data life cycle from collection to disposal, planning for multiple tiers of access to support different users and their needs, and tailoring information controls to specific intended data uses and potential privacy harms.[15] Data processing should control inferential risk directly, preferably by using formal methods or, alternatively, by analysis based on conservative assumptions about the threat environment.

Controls should be targeted to provide a meaningful level of protection to individuals and implemented with transparency.[15,16] These controls on inferential risk should be combined with the aforementioned technical controls limiting direct access and use. Also required are procedural, economic, educational, legal, and policy controls on data processing that recognize the interactions between these domains. Those, in turn, must support monitoring and mitigation of cumulative risk. Such controls may protect data sufficiently to allow it to be used with necessary accuracy and detail.[15,17]

## Privacy Regulation Must Keep Pace with Privacy Protection Technologies

Privacy risks have grown substantially with the explosion of online services, social media, and personal devices. There is now broad and substantial concern among stakeholders—including consumers, the media, and government—about these risks. Further, the fundamental economics of information prevent effective governance of privacy through purely market-based solutions and commercial self-regulation. In general, market-based solutions to privacy are plagued by the presence of externalities, information asymmetries, and human cognitive limits, as well as discouraged by economies of scope and scale.[18]

Managing privacy risks requires upgrading the current state of technical practice. Traditional disclosure limitation approaches have often failed to adequately protect privacy and will grow weaker over time. Managing privacy risks also requires new, broad and systematic regulation. As it stands, the underpinnings of data protection law fail to account for modern developments in the scientific understanding of information privacy.[13,19]

***Privacy does not occur by accident and must not be an afterthought.***

The law needs to fully recognize the illusory nature of perfect privacy and accuracy and the inadequacy of current practices based on deidentification and aggregation. It can, however, rely on state-of-the-art privacy-enhancing technologies that provide strong protection guarantees. Modernized privacy protection thus should include rigorous protection measures as a matter of both private sector initiative and governmental mandate and require transparent and accountable data processing to address cumulative privacy risks.

## KEY CONCLUSIONS

- To effectively protect privacy, controls must systematically address every stage of the data life cycle from collection to publication to disposal.
- Effective data protection requires combining conservative threat assumptions, rigorous technical methods that limit inferences, and complementary non-technical controls on data use.
- Wherever reliable anonymization is needed, data policies should prefer the use of new privacy-enhancing technologies.
- Regulation of data processing should reflect the need for multiple data access strategies to support a range of uses, the need to explicitly manage cumulative privacy loss for individuals, and transparency about protective methods, privacy guarantees, and the accuracy of analytical results.

## NOTES AND SOURCES

1. Henrik Twetman and Gundars Bergmanis-Korats, “Data Brokers and Security,” NATO Strategic Communications Centre of Excellence, January 14, 2021.
2. See: *Data Brokers Market Report*, Transparency Market Research, July 2022 https://www.transparencymarketresearch.com/data-brokers-market.html; Marc van Lieshout, “The Value of Personal Data,” in *Privacy and Identity Management for the Future Internet in the Age of Globalisation*, ed. Jan Camenisch, Simone Fischer-Hübner, and Marit Hansen, in IFIP Advances in Information and Communication Technology, vol. 457 (Cham, Switzerland: Springer, 2015), https://doi.org/10.1007/978-3-319-18621-4_3. Note that estimates and predictions are highly uncertain because of the opacity of the market, and because of the variability in rates of market growth.
3. Colleen McClain, Michelle Faverio, Monica Anderson, and Eugenie Park, “How Americans View Data Privacy,” Pew Research Center, October 18, 2023.
4. See: John M. Abowd and Michael B. Hawes, “Confidentiality Protection in the 2020 US Census of Population and Housing,” *Annual Review of Statistics and Its Application* 10 (2023): 119–44; Table 2. The percentage in the table represents the proportion of people who can be uniquely linked using high-quality external information.
5. Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel, “Unique in the Crowd: The Privacy Bounds of Human Mobility,” *Scientific Reports* 3, no. 1 (2013): 1–5.
6. Acxiom has testified to providing more than 10,000 measures, including scores for sensitive health attributes such as “vaginal itch scores” and “erectile dysfunction scores.” See: Micah Altman, Alexandra Wood, David R. O’Brien, and Urs Gasser, “Practical Approaches to Big Data Privacy Over Time,” *International Data Privacy Law* 8, no. 1 (February 2018): 29–51.
7. Aleecia M. McDonald, and Lorrie Faith Cranor, “The Cost of Reading Privacy Policies,” *I/S: A Journal of Law and Policy for the Information Society* 4, no. 2 (Summer 2008); Table 7. The table reflects the broadest and most reliable data analysis on this topic. Studies of smaller samples of privacy policies confirm that the policies remain frequent, lengthy, and time-consuming to read. See: Chiarra Castro, “You Need a Whole Workweek Every Month to Read Privacy Policies—and That’s Bad News,” TechRadar, October 25, 2023, https://www.techradar.com/computing/cyber-security/you-need-a-whole-workweek-every-month-to-read-privacy-policiesand-thats-bad-news; Kevin Littman-Navaro, “We Read 150 Privacy Policies. They Were an Incomprehensible Disaster,” *New York Times*, https://www.nytimes.com/interactive/2019/06/12/opinion/facebook-google-privacy-policies.html.
8. Heather Morton, “2023 Consumer Data Privacy Legislation,” September 28, 2023, National Conference of State Legislatures, https://www.ncsl.org/technology-and-communication/2023-consumer-data-privacy-legislation.
9. Based on arXiv metadata search of titles and abstracts using the term “privacy,” published in 2023, with classification “computer science.”
10. See: Cynthia Dwork and Aaron Roth, “The Algorithmic Foundations of Differential Privacy,” *Foundations and Trends® in Theoretical Computer Science* 9, nos. 3–4 (August 2014): 211–407; Irit Dinur and Kobbi Nissim, “Revealing Information While Preserving Privacy,” in PODS 2003: *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* (2003): 202–10.
11. See: Abdulaziz Alzubaidi and Jugal Kalita, “Authentication of Smartphone Users Using Behavioral Biometrics,” *IEEE Communications Surveys & Tutorials* 18, no. 3 (2016): 1998–2026; Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard, “Surveying Stylometry Techniques and Applications,” ACM Computing Surveys 50, no. 6 (November 2018): 1–36; Arvind Narayanan and Vitaly Shmatikov, “Robust De-Anonymization of Large Sparse Datasets,” in *SP ’08: Proceedings of the 2008 IEEE Symposium on Security and Privacy* (2008): 111–25.
12. Aaron Fluitt, Aloni Cohen, Micah Altman, Kobbi Nissim, Salome Viljoen, and Alexandra Wood, “Data Protection’s Composition Problem,” *European Data Protection Law Review* 5, no. 3 (2019): 285.
13. Paul Ohm, “Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization,” *UCLA Law Review* 57 (2010): 1701; Ira S. Rubinstein and Woodrow Hartzog, “Anonymization and Risk,” *Washington Law Review* 91, no. 2 (June 2016): 703; National Academies of Sciences, Engineering, and Medicine, *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps* (Washington, DC: The National Academies Press, 2017), https://doi.org/10.17226/24893.
14. Differential privacy is a formal mathematical framework for quantifying anonymization risks. Homomorphic encryption enables computations to be performed on encrypted content without decrypting it, thereby protecting the confidentiality of the data from the owner of the computation system. Secure multiparty computation enables multiple data controllers to jointly compute prespecified functions over their collective data without revealing that data. See: Alexandra Wood, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, James Honaker, Kobbi Nissim, David R. O’Brien, Thomas Steinke, and Salil Vadhan, “Differential Privacy: A Primer for a Non-Technical Audience,” *Vanderbilt Journal of Entertainment and Technology Law* 21, no. 1 (Fall 2018); “Emerging Privacy-Enhancing Technologies: Current Regulatory and Policy Approaches,” OECD Digital Economy Papers, no. 351 (March 8, 2023).
15. See: Micah Altman, Alexandra Wood, David R. O’Brien, Salil Vadhan, and Urs Gasser, “Towards a Modern Approach to Privacy-Aware Government Data Releases,” *Berkeley Technology Law Journal* 30, no. 3 (May 2016): 1967–2072; *NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management, Version 1.0*, National Institute of Standards and Technology, January 16, 2020, https://doi.org/10.6028/NIST.CSWP.10; NASEM, *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps* (207).
16. Joseph P. Near and David Darais, *Guidelines for Evaluating Differential Privacy Guarantees*, National Institute of Standards and Technology, December 2023, https://doi.org/10.6028/NIST.SP.800-226.ipd; Micah Altman and Aloni Cohen, “Natural Differential Privacy—A Perspective on Protection Guarantees,” *PeerJ Computer Science* 9 (2023); Ron S. Jarmin, John M. Abowd, Robert Ashmead, Ryan Cumings-Menon, Nathan Goldschlag, Michael B. Hawes, Sallie Ann Keller, et al., “An In-Depth Examination of Requirements for Disclosure Risk Assessment,” *Proceedings of the National Academy of Sciences* 120, no. 43 (October 24, 2023).
17. Shawn Cole, Iqbal Dhaliwal, Anja Sautmann, and Lars Vilhuber, eds., *Handbook on Using Administrative Data for Research and Evidence-Based Policy* (Cambridge, MA: Abdul Latif Jameel Poverty Action Lab, 2020).
18. For example, when someone posts a group photo on social media, they reveal information about others (an externality); pragmatically skim the lengthy and complex terms of service (an example of bounded cognition); and are unaware of the potential use of their photo in training the next generation of commercial AI models (an information asymmetry). See more generally: Alessandro Acquisti, Curtis Taylor, and Liad Wagman, “The Economics of Privacy,” *Journal of Economic Literature* 54, no. 2 (June 2016): 442–92; Bertin Martens, “An Economic Perspective on Data and Platform Market Power,” March 22, 2021; Avi Goldfarb and Verina F. Que, “The Economics of Digital Privacy,” *Annual Review of Economics* 15 (2023): 267–86.
19. Micah Altman, Aloni Cohen, Kobbi Nissim, and Alexandra Wood, “What a Hybrid Legal-Technical Analysis Teaches Us About Privacy Regulation: The Case of Singling Out,” *Boston University Journal of Science & Technology Law*, 27.1 (Winter 2021): 1; Micah Altman, Aloni Cohen, Francesa Falzon, Evangelia Anna Markatou, Kobbi Nissim, Michel Jose Reymond, Sidhant Saraogi, and Alexandra Wood, “A Principled Approach to Defining Anonymization as Applied to EU Data Protection Law” (May 9, 2022); Daniel J. Solove, “Data Is What Data Does: Regulating Based on Harm and Risk Instead of Sensitive Data,” *Northwestern University Law Review* 118, no. 4 (2024): 1081.

## ADDITIONAL INFORMATION

With 100,000 members in 190 countries, the nonprofit **Association for Computing Machinery** is the world’s largest and longest-established organization of professionals involved in all aspects of computing. Under the auspices of the global ACM Technology Policy Council, policy committees in the U.S. and Europe provide cutting-edge, apolitical, non-lobbying information about computing and its social impacts to policy makers at all levels of government in many forms, including briefings, testimony, consultation, and rulemaking comments, reports, and analyses.

**To tap the deep expertise of ACM’s global membership, please contact ACM’s Global Policy Office at *acmpo@acm.org* or +1 202.580.6555. To receive ACM TechBriefs quarterly, in the body of a one-line email send *subscribe ACM-tpc-tech-briefs* followed by your first and last names to *listserv@listserv.acm.org***

## AUTHORSHIP & ACKNOWLEDGMENTS

**Micah Altman** is a research scientist at the Center for Research on Equitable and Open Scholarship, and the Massachusetts Institute of Technology. Aloni Cohen is an assistant professor of computer science and data science at the University of Chicago. Kobbi Nissim is a professor in the Department of Computer Science at Georgetown University and an affiliate professor at Georgetown Law. Nissim’s work was supported by NSF grant no. CCF2217678 and a gift to Georgetown University. The authors wish to thank their collaborators—Aaron Bembenek, Mark Bun, Aaron Fluitt, Marco Gaboardi, James Honaker, David R. O’Brien, Thomas Steinke, Salil Vadhan, Salome Viljoen, and Alex Wood—with whom they have undertaken the research on which this TechBrief draws, as well as Professor of Computer Science Stephen Chong of Harvard University for his expert review. This brief may be cited as “*ACM TechBrief: Data Privacy Protection*, ACM Technology Policy Council (Issue 11, Spring 2024).”