Data Protection's Composition Problem

*Aaron Fluitt, Aloni Cohen, Micah Altman, Kobbi Nissim, Salome Viljoen and Alexandra Wood**

I. A Time of Reckoning for the Information Ecosystem

Is it possible to piece together the confidential data of almost everyone in the US from statistics published by the Census Bureau—without breaching Census security or policy? Could someone—a doctor, a nosy neighbour, or a foreign state actor—determine whether a particular person participated in a genetic study of hundreds of individuals, when each individual contributed only tiny trace amounts of DNA to a highly complex and aggregated genetic mixture? Could police detectives re-trace a suspect's every movement over the course of many months and thereby learn intimate details about the suspect's political, religious, and sexual associations—without having to deploy any sort of surveillance or tracking devices? Could someone reliably deduce the sexual preferences of a Facebook user without looking at any content that user has shared?

Until recently, most people probably never imagined that their highly sensitive personal data could be so vulnerable to discovery from seemingly innocuous sources. Many continue to believe that the privacy risks from purely public, statistical, and anonymised data are merely theoretical, and that the practical risks are negligibly small.[1] Yet all of the privacy violations described above are not only theoretically possible—they have already been successfully executed.[2]

DOI: 10.21552/edpl/2019/3/4

* Aaron Fluitt is a Privacy Law Fellow at the Institute for Technology Law & Policy at Georgetown University. For correspondence: <jaf294@georgetown.edu>. Aloni Cohen is a Postdoctoral Associate at the Hariri Institute for Computing and Computational Science and Engineering at Boston University and the Boston University School of Law: <aloni@bu.edu>. Micah Altman is Director of Research at the Center for Research in Equitable and Open Scholarship at the Massachusetts Institute of Technology. For correspondence: <escience@mit.edu>. Kobbi Nissim is a McDevitt Chair in Computer Science at Georgetown University and affiliated with Georgetown University Law Center and Ethics Lab. For correspondence: <kobbi.nissim@georgetown.edu>. Salome Viljoen is a Joint Research Fellow at the Digital Life Initiative at Cornell Tech and the Information Law Institute at NYU Law. For correspondence: <sviljoen@cyber.harvard.edu>. Alexandra Wood is a Fellow at the Berkman Klein Center for Internet & Society at Harvard University. For correspondence: <awood@cyber.harvard.edu>.
The authors describe contributions to this Essay using a taxonomy from Liz Allen et al, *Publishing: Credit Where Credit Is Due*, 508 Nature 312 (2014). MA, KN, and AW provided the core formulation of the Essay's goals and aims; and AF and AC led the writing. All authors contributed through commentary, review, editing, and revision. The authors wish to thank Salil Vadhan, Elisabeth Perlman, Aaron Bembenek, and the members of the Privacy Tools Project and the Bridging Privacy Definitions Working Group for their thoughtful comments. This material is based upon research supported by the Alfred P Sloan Foundation and the US Census Bureau under cooperative agreement no CB16ADR0160001. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the Alfred P Sloan Foundation or the US Census Bureau.

1 See, eg, Ann Cavoukian and Daniel Castro, *Big Data & Innovation, Setting the Record Straight: DeIdentification Does Work* (16 June 2014) <http://www2.itif.org/2014-big-data-deidentification.pdf> accessed 25 August 2019.

2 The first case describes a recently-revealed test-attack on the publicly-released statistical tables from the 2010 US Census, discussed in more detail below. The second case recounts a 2008 attack on a genetic database that prompted the National Institutes of Health to remove public access to the database while it assessed 'the broader scientific, ethical, and policy implications' of the development. See Nils Homer et al, 'Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays' (29 August 2008) 8 PLOS Genetics 4; Elias A Zerhouni and Elizabeth G Nabel, 'Protecting Aggregate Genomic Data' (3 October 2008) 322 Science 44. The third case describes the criminal investigation, prosecution, and conviction of Timothy Ivory Carpenter in 2011. *Carpenter v United States*, 138 SCt 2206, 2217 (2018). The fourth case of 'Facebook Gaydar' has been repeatedly referenced in privacy and technology-related news since its announcement in 2009. See, eg, Ki Mae Heussner, '"Gaydar" on Facebook: Can Your Friends Reveal Sexual Orientation?' *ABC News* (22 September 2009) <https://abcnews.go.com/Technology/gaydar-facebook-friends/story?id=8633224> accessed 25 August 2019.

To elaborate on the first example, researchers in 2018 revealed that the underlying confidential data from the 2010 US Decennial Census could be reconstructed using only the statistical tables published by the Census Bureau.[3] They demonstrated a type of attack, called a *database reconstruction attack*, that leveraged the large volumes of data from the published statistical tables in order to narrow down the possible values of individual-level records.[4] The researchers were able to reconstruct the sex, age, race, ethnicity, and fine-grained geographic location (to the block-level) reported by Census respondents *exactly* for 46% of the US population.[5] They also showed that, if they relaxed their conditions and allowed age to vary by up to only one year, these five pieces of information could be reconstructed for 71% of the population.[6] Further, the researchers showed that the reconstructed records could be completely *re-identified*—meaning they were able to assign personally identifiable information to individual records—using commercial databases available at the time.[7] They concluded that, with this attack, they could putatively re-identify 138 million people, and they confirmed that these re-identifications were accurate for 52 million people, or 17% of the US population.[8]

These findings are startling. The last time the Census Bureau performed such a simulated re-identification attack on census datasets, the re-identification rate was only 0.0038%. The 2018 test attack demonstrates that previous risk assessments underestimated the re-identification risk by a factor of at least 4,500![9]

The foregoing examples of real-world privacy attacks all leverage one particular vulnerability that we refer to as *composition effects*.[10] This vulnerability stems from the cumulative erosions of privacy that inhere in every piece of data about people. These erosions occur no matter how aggregated, insignificant, or anonymised the data may seem, and even small erosions can combine in unanticipated ways to create big risks.[11]

Privacy and data protection failures from unanticipated composition effects reflect a type of data myopia—a short-sighted approach toward addressing increasingly-ubiquitous surveillance and privacy risks from Big Data analytics, characterised by a near-total focus on individual data processors and processes and by pervasive underestima-

3 See Simson L Garfinkel, John M Abowd and Christian Martindale, 'Understanding Database Reconstruction Attacks on Public Data' (October 2018) 16(5) ACMQUEUE 28–53 <https://queue.acm.org/detail.cfm?id=3295691> accessed 25 August 2019. Reconstruction attacks were presented as a tool for analysing privacy in Irit Dinur and Kobbi Nissim, 'Revealing Information while Preserving Privacy' (2003) Proc Of the ACM Symposium on Principles of Database Systems (PODS) 202.

4 ibid.

5 See John Abowd, 'Stepping-up: The Census Bureau Tries to Be a Good Data Steward in the 21st Century' (Presentation at the Simons Institute for the Theory of Computing, Berkeley, 4 March 2019) <https://bit.ly/2lDaXwk> accessed 25 August 2019.

6 ibid.

7 ibid.

8 ibid.

9 See John Abowd, 'Tweetorial: Reconstruction-abetted re-identification attacks and other traditional vulnerabilities' <http://blogs.cornell.edu/abowd/special-materials/245-2/> accessed 25 August 2019.

10 See Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan and Adam Smith, 'Composition Attacks and Auxiliary Information in Data Privacy' (14 Proc ACM SIGKDD Int'l Conf. on Knowledge, Discovery & Data Mining, 2008) 265, 265–66.

11 For a review of known characteristics of data that are contributing to accelerated privacy risk in practice, see Micah Altman et al, 'Practical Approaches to Big Data Privacy over Time' (2018) 8 Int'l Data Privacy L 29.

tion of systemic risks accumulating from independent data products. The failure to recognize accumulation of risk in the information ecosystem reflects a more general societal blind spot to cumulative systemic risks, with parallels in collective failures to foresee or forestall global financial crises, and to adequately address mounting risks to the natural environment.

As the volume and complexity of data uses and publications grow rapidly across a broad range of contexts, the need to develop frameworks for addressing cumulative privacy risks is likely to become an increasingly urgent and widespread problem. Threats to privacy are growing due to the accelerating abundance, and richness, of data about individuals being generated and made publicly available. Furthermore, substantial increases in computing power and algorithmic improvements are making the execution of such attacks more technically feasible. These threats will be impossible to overcome unless regulations are designed to explicitly regulate cumulative risk in a manner that is consistent with the science of composition effects.

II. Understanding Composition

Broadly speaking, composition in the data privacy context relates to the accumulation of privacy risk or harm across a sequence of decisions related to the use of data. We distinguish between two complementary concepts: *composition effects* and *composability*.

Composition effects are the cumulative results of multiple uses of data vis-a-vis data privacy. In the data privacy context, composability is a property of certain processes used to preserve privacy, typically referred to as formal privacy concepts, that enables one to reason about—and thereby manage and control—composition effects on privacy in a modular way. If an approach to preserving privacy is composable, then the composition effects of multiple data uses employing the approach can be understood by analysing each data use by itself. We refer to protections as *composable data protections* when they can be shown to limit privacy loss in a manner that degrades predictably and gradually across repeated applications. This can empower an organisation like the Census Bureau to perform multiple data releases while controlling privacy risks.

1. Composition Effects

Over the last two decades, privacy researchers have revealed an inconvenient truth that applies universally to every data release. Under what has come to be called the *fundamental law of information recovery*, releasing "overly accurate" estimates of 'too many" statistics completely destroys privacy.'[12] The reason is that 'the more information that is released, the more determined the underlying data is.'[13] With the accumulation of

12 ibid at fn 84 (citing Cynthia Dwork and Guy N Rothblum, 'Concentrated Differential Privacy' (Working paper, 2016) <https://arxiv.org/pdf/1603.01887.pdf> accessed 25 August 2019 (citing Dinur and Nissim (n 3) 202 et sqq)).

13 Hector Page, Charlie Cabot and Kobbi Nissim, 'Differential Privacy: An Introduction for Statistical Agencies' (2018) National Statistician's Quality Review into Privacy and Data Confidentiality Methods 5.

enough data releases, whatever their form, 'an adversary may be able to reconstruct, either exactly or with very high accuracy, the entire dataset from these summaries.'[14]

A classic logic puzzle illustrates this phenomenon in a stylised way:[15]

> A man opens his door to a census taker, who asks how many people reside at the address and their ages. The man explains that it is just him and his three daughters. Instead of providing his daughters' ages, the man tells the census taker, 'The product of my daughters' ages is 36, and the sum is 13.' He then dismisses the census taker, noting 'I have to get my oldest daughter to her piano lesson.' The census taker thanks the man and accurately records the daughters' ages in his notes.

How was the census taker able to deduce the daughters' ages from the information provided? Each piece of information—the product of the ages, the sum, and the existence of an oldest daughter—narrows down the possible age combinations and ultimately reveals the exact ages. Figure 1 illustrates with a dotted circle the possible combinations of the three daughters' ages with a product of 36. Of those, the dashed circle contains the possible age combinations with a sum of 13. The solid circle contains the only combination that also has an oldest child: 2, 2, and 9. Although the possible age combinations that satisfy each clue independently may seem overwhelmingly vast, together the three clues eliminate all but one set of possible age combinations.

While significantly more complex, the reconstruction of individual-level data from the 2010 Decennial Census is based on the same intuition. Each piece of statistical data published by the Census Bureau acts as a constraint on the possible values of an individual's personal information. If enough data are published, it can enable one to narrow down the range of possible values significantly, and, in some cases, even to a point where an individual's personal information can be determined exactly.

Composition effects can take many forms. For instance, releasing either the first half of a credit card number or the second would not allow somebody to charge the card. But releasing both would—a composition effect that does not fit into the earlier mould. While this example is extremely simplistic, popular approaches to preserving privacy are vulnerable to similar composition effects both in theory and in practice.[16]

In many well-known re-identification attacks—including on the Netflix and AOL datasets and Massachusetts medical records—purportedly anonymised data was sufficiently rich to link to outside public sources of information.[17] While these can be

14 ibid.

15 This puzzle is adapted from Richard Rider, 'Census Puzzle' (Submission to Mathsisfun.com, 2017) <https://www. mathsisfun.com/census.html> accessed 25 August 2019.

16 Ganta et al (n 10); see also Aloni Cohen, 'New Guarantees for Cryptographic Circuits and Data Anonymization' (DPhil thesis, Massachusetts Institute of Technology 2019).

17 See Paul Ohm, 'Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization' (2010) 57 UCLA LR 1701.

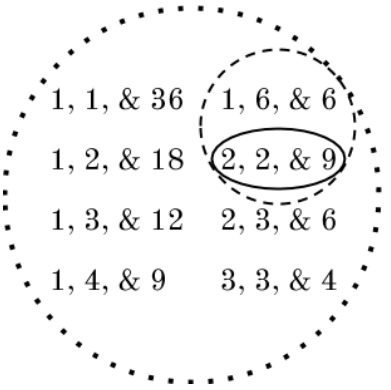

***Figure 1**. Census taker puzzle as an example of composition effects*

thought of within the framework of composition effects and accumulated privacy risk, they should not be taken as representative. Composition effects describe a more universal phenomenon: the degradation of privacy protection resulting from multiple uses of data, even if each use is intended to respect privacy.

Privacy risks stemming from the combination or linkage of datasets that already contain individual-level identifying (or quasi-identifying) attributes have been thoroughly covered in academia and regulations for years. By contrast, law and policy has paid relatively scant attention to an attacker's ability to reconstruct individual records from aggregated statistics by leveraging what can be very subtle and unintuitive composition effects; yet the risks from such attacks are particularly significant precisely because no individual-level records are required. Once the underlying data is reconstructed from the statistics and individual data subjects have been singled out, the effort needed to link the reconstructed data to outside sources of identifiable information may often be trivial by comparison.

2. Composability

Composition effects can make it quite difficult to reason about the combined effects of a collection of actions, especially when individual actions are made by different people, at different times, and with minimal coordination. At its most abstract, composability allows one to reason about a whole collection of actions or decisions by reasoning about each of the individual actions or decisions separately. Composability is a feature of a measurable property relevant to decision-making (eg, price, risk, revenue). When a property is composable, there is some calculus or method for measuring the whole by measuring the constituent parts.

For an example outside the domain of privacy, price is typically composable: the total price of four $50 chairs and a $200 table is $400. In this case, the calculus for combining the parts is very simple (ie, adding up the prices), but the calculus for gauging composition effects may be much more complex in different contexts. In contrast, as

anyone who has lived in a tiny apartment can attest, it is not always possible to fit a queen-sized bed (33 square feet), a desk (12 square feet), and a dresser (9 square feet) into a small bedroom (70 square feet). Whether the area occupied by furniture fits in a room is not a composable property.[18]

How data privacy risks compose is worth particular attention; if the aim of data protection is to limit potential harms resulting from information releases, then it is important to understand how protection degrades with multiple releases.

Returning to the census taker puzzle, because the composition effects of the three clues could be leveraged to deduce the age of the three daughters, whatever data protection the father's evasiveness provided was not composable. That is, while each of the father's statements did not *independently* reveal information the father intended to keep confidential, the combined effect of the statements was the unintended disclosure of the daughters' ages.

Composability of effective data protection rarely occurs by accident. K-anonymity[19] for example, is not a composable data protection: even if each of two releases are each 100-anonymous, it is possible that, in combination, they fail to preserve even 2-anonymity.[20] Other privacy properties may be composable in an abstract sense but do not provide composable data privacy. For example, we can correctly infer that the combination of two (properly) HIPAA-redacted databases composes in a way that preserves the redactions—so redaction composes in this abstract sense. However, the combination of redacted databases may create risks that are far greater than the risks of each individually—through mechanisms akin to those in the census-taker puzzle. Thus redaction is *not* a composable data protection. More broadly, redaction, sampling, swapping, and aggregation may compose in the general sense, yet these common approaches to preserving privacy are *not* composable data protections. The 2010 Census example above underlines the distinction. Although the procedures used by Census to protect data were neither technically reversed nor invalidated by merely combining multiple releases, the protection against risk offered by their methods degraded rapidly and unpredictably.

In contrast, differential privacy[21] *is* a composable data protection. The maximum privacy risk to an individual from a single differentially private data release is bounded

18 One could reason about whether the furniture would fit in the room using their *dimensions*—length and width—rather than area measured in square feet. So there are composable quantities relevant for the problem of fitting furniture in a room, just not area.

19 *k*-anonymity is a privacy concept sometimes used to de-identify sensitive data. The parameter *k* in *k*-anonymity is a number that intends to capture the strength of the privacy guarantee, with a higher number considered more private. Typical values of *k* are 5–10, and *k* =1 offers no privacy at all. See Latanya Sweeney, 'k-anonymity: A model for protecting privacy' (2002) 10(05) International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 557–570.

20 Ganta et al (n 10).

21 Differential privacy is a mathematical privacy concept, where the parameter ε limits the privacy risks to individuals. Composability is one of the distinguishing features of differential privacy, and a collection of composition theorems give provable bounds on the combined compositional effects of multiple differentially private analyses. See Alexandra Wood et al, 'Differential Privacy: A Primer for a Non-Technical Audience' (2018) 21 Vand J Ent & Tech L 209; Page, Cabot and Nissim (n 13). In light of the 2010 Census reconstruction, differential privacy will be used to generate data products from the 2020 Decennial Census responses.

by the parameter epsilon (ε), and the maximum risk from multiple releases is a slowly-growing function of the ε values of each release. This is no accident: composability was designed into that privacy concept from the outset.

III. Implications of Composition Effects for Data Protection

In a world where data sources are becoming increasingly detailed and made available to wider audiences with greater frequency, it has become exceedingly difficult (and in many cases impossible) to predict how fast privacy degrades with each new data use—and traditional disclosure avoidance methods fail to provide composable data protections against this degradation. Fortunately, some formal models for preserving privacy, like differential privacy, *do* provide some composable protection, and thus facilitate the holistic understanding, tracking, and managing of cumulative privacy loss for data subjects across many formally private data releases.

Current formal methods provide composable data protection against individual privacy harm from computations over personal data. However, these formal methods do not necessarily provide protection against other privacy risks. In the aggregate, individual decisions to click on a link can lead to virtual group red-lining;[22] and cameras that individually are useful only for issuing speeding tickets can combine to form a national surveillance network.[23] If we do not design legal systems that provide composable protections at multiple levels, many small decisions will likely result in a cascade of societal data protection failures.

Legal treatment of privacy risk should strive to draw experience from other areas of the law—such as environmental protection and financial systems regulation—that incorporate formal quantitative assessment of cumulative and systemic risks. Experiences in environmental and financial regulation demonstrate that standardisation of technologies, processes, and assessment metrics facilitates risk measurement and management.[24] An important policy question that arises in these regulatory contexts involves the degree to which regulations should dictate the specific protections to be applied, since an overly rigid approach can stifle innovation and hinder growth.

Because of the rapid pace of technological change relating to data privacy, it will be challenging to strike the right balance between the need for some degree of technological neutrality to ensure the regulatory framework remains flexible and durable, and the powerful benefits of having common metrics and centralised processes by which to assess and manage composition effects. Because of these substantial technical chal-

22 See Latanya Sweeney, 'Discrimination in online ad delivery' (2013) arXiv preprint arXiv:1301.6822.

23 See eg, Linda M Merola and Cynthia Lum, 'Emerging surveillance technologies: privacy and the case of License Plate Recognition (LPR) technology' (2012) Judicature 96, 119; and Altman et al (n 11).

24 See eg, Dennis D Hirsch, 'Protecting the Inner Environment: What Privacy Regulation Can Learn From Environmental Law (2006) 41 Georgia L Rev 1, 39–42 (Allowing polluting facilities 'to choose their own control method [makes] it harder for government officials to track emission levels and to enforce [emissions] fee requirements.'); Iman Anabtawi and Steven L Schwarcz, 'Regulating Systemic Risk: Towards an Analytical Framework' (2011) 86 Notre Dame LR 1349, 1412.

lenges, data protection law needs to develop an evolving toolkit for assessing and addressing composition effects that is flexible enough to deal with current knowledge gaps and adaptable enough to increase in sophistication as our technical and empirical capacities mature. Two critical components of an adaptive strategy for addressing uncertainties are ongoing monitoring of consequences of particular approaches to protecting privacy, and a continuously iterative process for repeatedly assessing consequences of previous approaches in light of new evidence obtained through the monitoring. The GDPR's explicit concern about identifiability of ostensibly-anonymised information suggests an opportunity for regulation to move away from compartmentalised assessments of privacy risk, toward a more holistic and evolving approach that could provide the flexibility needed to recognize and address composition effects on privacy. It may be premature to say how effective the EU data protection regime will be at guarding against composition effects, but there is urgent need for action on this front, before the cumulative composition effects of Big Data carry us beyond the crisis point.

Because privacy losses inevitably accumulate, protecting privacy will require a more systemic approach to privacy regulation. The prevailing approach to protecting personal data at the level of individual data processors, rather than treating risks to personal data at the macro or systemic level, typifies the *tyranny of small decisions*:[25] although each step seems small, together they bring society over a cliff. If data protection regulation is to be successful, it must recognize that small privacy risks can multiply unexpectedly, and potentially catastrophically—unless protections are explicitly implemented to limit cumulative risk. Moreover, it is critical to acknowledge that privacy losses *always* accumulate—even when information protections are well-designed.

A fundamental role of the legal system is to constrain individual actors so that the societal impacts of their actions are predictable and proportionate. The consequences of a global privacy catastrophe stemming from the uncoordinated decisions of individuals and organisations are too severe to address after it has occurred. Society therefore needs to provide ex-ante protections on information risks *before* massive and irreversible damage has been done. Theoretical advances in the science of data protection provide the tools necessary to develop effective ex-ante solutions, and data protection laws must leverage and apply this scientific progress.

25 See Laurence Tribe, 'Constitutional Calculus' (1985) 98 Harv LR 592, 611–12 (and fn 119).