BRUSSELS PRIVACY SYMPOSIUM

on

IDENTIFIABILITY: POLICY AND PRACTICAL SOLUTIONS FOR ANONYMISATION AND PSEUDONYMISATION

Framing the Discussion

by

Ira Rubinstein
Adjunct Professor and Senior Fellow, NYU School of Law
Senior Fellow, Future of Privacy Forum

Introduction

Deidentification—the process of modifying personal data to ensure that data subjects are no longer identifiable—is one of the primary measures that organizations use to protect privacy. Proper deidentification enables organizations to safely share data sets for a broad range of valuable purposes without endangering the privacy interests of data subjects. This matters. Government agencies routinely collect, process, and share huge troths of citizens' data for a wide range of administrative purposes and to ensure accountability regarding their own activities. Commercial firms providing financial, healthcare, retail or marketing services match or exceed government collection and use of data, and they often rely on deidentified data to develop or improve products and services. And, of course, academic researcher rely on many sorts of data for a wide range of public health and social science research. More recently, and under the rubric of open data, governments and other large organizations have started to publicly release large data sets to promote the public good and lend support both to commercial endeavors and funded research. In short, deidentified data is a vital aspect of the digital economy. We all benefit from it in many ways ranging from education programs, to improved traffic flows and urban planning, to anti-theft and fraud programs, to genetic research.[1] Not surprisingly, both European data protection and U.S. sectoral privacy laws regulate deidentification, seeking to achieve an optimal balance privacy concerns and data utility.

Over the past decade, however, computer scientists and mathematicians have demonstrated that deidentification is not foolproof and regulators have struggled with how best to respond to these new developments. Most privacy laws worldwide define their scope of application based on whether information is identifiable or not. Indeed, many privacy laws associate privacy harm with "personal data" (or, to use the American term, "personally identifiable information" (PII)), while treating anonymous data as unregulated. But "identifiability" and "anonymity" are ambiguous terms, and the tools and techniques for transforming one into the other are highly contested. This leaves us with many questions in need of answers.

Should we define these terms in binary fashion or are they better understood as the end-points of a wide spectrum? Are certain tools and techniques best suited for specific research fields and lines of inquiry, or certain types of data, and can experts in statistical disclosure control and computer science reach agreement on where the affinities reside? Given the inevitable trade-offs between privacy and data utility, are there optimal ways to balance these competing interests? Should regulators adjust data protection requirements in light of anticipated risk levels and to ensure that data custodians have strong incentives for investing in and using state of the art methods of protection? Are the tools and techniques that support privacy-protective uses of datasets best understood in terms of appropriate safeguards that minimize risk under specific circumstance or should we insist on provable privacy guarantees that eliminate risk entirely? Are scientific experts themselves any closer to resolving these disputed issues?

[1] *See* Jules Polonetsky, Omer Tene & Kelsey Finch, *Shades of Gray: Seeing the Full Spectrum of Practical Data Deidentification* 56 SANTA CLARA L. REV. 593 (2016), Appendix A.

In Europe, these questions are more timely and complicated than ever given the recent approval of the General Data Protection Regulation (GDPR). While the GDPR changes prior data protection law in many ways, it continues to allow data controllers to meet their obligations on a relaxed basis, or even remove certain data from the scope of the Regulation, when the data in question are no longer identifiable. The GDPR also introduces the related concept of "pseudonymisation," defined as the processing of personal data in such a way as to prevent attribution to an identified or identifiable person without additional information that must be held separately. Although such data remains subject to the remit of the Regulation, the GDPR recognizes that pseudonymisation potentially reduces the risks for data subjects and therefore relaxes certain requirements when controllers use this technique. It also allows pseudonymisation to be a factor in meeting certain obligations (such as data security and data protection by design) and when considering the compatibility of different uses of data with the conditions of initial collection.

While the introduction of this new concept may be viewed as a positive development, signaling a welcome shift from a binary to a tripartite approach to identifiability, the GDPR treatment of pseudonymisation raises more questions than it answers. For example, what technical and organizational measures are required to ensure that pseudonymized data has met regulatory standards (which currently varies per Member State's law and policy)? When organizations utilize such measures, which legal requirements are relaxed under the GDPR and by how much? Does the GDPR provide sufficient incentives for organizations to use this technique as part of an overall compliance strategy?

This symposium seeks to address these and other related questions about identifiability, anonymisation and pseudonymisation, and to surface and discuss practical solutions that rely on these techniques. We have invited a group of academic experts and prominent policy makers to examine the technical, policy and ethical aspects of deidentification reidentification, and expect to learn a great deal from them about these topics.

The Deidentification Debate: Is it Safe to Go in the Water?

Beginning in the 1990s and accelerating in the past few years, several well-publicized incidents have shown that data sets that apparently were deidentified remain vulnerable to reidentification attacks. Indeed, many commentators believe that a well-known trio of reidentification cases call into question the underlying validity of deidentification.[2] These incidents raised serious doubts for many about the extent to which deidentification remains a

[2] The three cases famously involve the public release of (1) deidentified hospitalization records of state employees including then-Massachusetts Governor Weld; (2) twenty million search queries of 650,000 AOL users, and (3) more than 100 million ratings from over 480,000 Netflix customers on nearly 18,000 movie titles. All three incidents involved linkage attacks, in which an individual or entity trying to reidentify a data subject takes advantage of auxiliary or background information to link an individual to a record in the deidentified data set. *See* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1717-23. (2010).

credible method for using and deriving value from large data sets while protecting privacy. Both legal and technical experts are sharply divided on the efficacy of deidentification and related solutions. Some critics argue that it is impossible to eliminate privacy harms from publicly released data using deidentification due to the growing availability of background data, which allow attackers to identify data subjects by mounting linkage attacks. Defenders of deidentification counter that despite the theoretical and demonstrated ability to mount such attacks, the likelihood of reidentification for most data sets remains minimal.

Which side is right? One would like to think that the relevant scientific experts would have sorted out their differences by now and resolved any lingering doubts about deidentification techniques. Unfortunately, this is not the case. To the contrary, the community of computer scientists, statisticians, and epidemiologists who write about deidentification and reidentification seem deeply divided, not only in how they view the implications of linkage attacks, but in their goals, methods, interests, and measures of success. Indeed, some commentators argue that the experts fall into distinct camps of "pragmatists" and formalists."[3] In general, pragmatists share an expertise in deidentification methods and value practical solutions for sharing useful data to advance the public good. Accordingly, they devote a great deal of effort to devising methods for measuring and managing the risk of reidentification for clinical and other specific disclosure scenarios.[4] In sharp contrasts, formalists are less concerned with finding practical solutions than with achieving mathematical rigor in defining privacy, modeling adversaries, and quantifying the probability of reidentification. They seek provable privacy guarantees using methods first developed in cryptography and more recently applied in theoretical research associated with differential privacy.[5]

Pragmatists consider it difficult to gain access to auxiliary information and consequently give little weight to well-known reidentification attacks, in which the data subjects may be distinguishable and unique but no one is ever identified on an individual basis. And they point to empirical studies and meta-analyses showing that the risk of reidentification in properly deidentified data sets is, in fact, very low. Formalists object to such studies on the grounds that these efforts to quantify the efficacy of deidentification "are unscientific and promote a false sense of security by assuming unrealistic, artificially constrained models of what an adversary might do."[6] Unlike the pragmatists, they take very seriously proof-of-concept demonstrations

[3] *See* Ira Rubinstein & Woodrow Hartzog, *Anonymization and Risk,* 91 WASH. L. REV. 703, 714-17 (2016).

[4] *See, e.g.,* KHALED EL EMAM, GUIDE TO THE DE-IDENTIFICATION OF PERSONAL HEALTH INFORMATION (2013).

[5] Differential privacy has been described as "a set of techniques based on a mathematical definition of privacy and information leakage from operations on a data set by the introduction of non-deterministic noise. Differential privacy holds that the results of a data analysis should be roughly the same before and after the addition or removal of a single data record (which is usually taken to be the data from a single individual). In its basic form differential privacy is applied to online query systems, but differential privacy can also be used to produce machine-learning statistical classifiers and synthetic data sets." SIMSON L. GARFINKEL, NAT'L INST. OF STANDARDS & TECH., DEIDENTIFICATION OF PERSONAL INFORMATION (NISTIR 8053) 4 (2015), http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf.

[6] ARVIND NARAYANAN & EDWARD W. FELTEN, NO SILVER BULLET: DEIDENTIFICATION STILL DOESN'T WORK (2014), http://randomwalker.info/ publications/no-silver-bullet-deidentification.pdf.

of reidentification, while minimizing the importance of empirical studies showing low rates of reidentification in practice.

This split among the experts is concerning for several reasons. Pragmatists and formalists represent distinctive disciplines with very different histories, questions, methods, and objectives. Accordingly, they have not—until very recently—shown much inclination to engage in fruitful dialogue or join in finding ways to resolve their differences by placing deidentification on firmer foundations. Of course, this makes it very difficult for policy makers to judge whether current deidentification requirements should be maintained, reformed, or abandoned. And this uncertainty, in turn, has very broad consequences.

1. It affects the privacy of data subjects across a broad range of contexts.
2. It affects privacy-driven organizations, many of which devise a compliance strategy premised on the identifiable/non-identifiable distinction and take steps to transform one into the other with the goal of limiting or eliminating their obligations under applicable privacy statutes and regulations. Clearly, the lack of certainty around deidentification undermines this strategy.
3. It endangers valuable research whether by creating doubts in data subjects about how safe it is to participate in studies using deidentified data or making it impractical for researchers to engage in longitudinal research or to reuse existing data for secondary purposes that are inconsistent with the original terms of collection.
4. It has serious implications for "open data." A key argument in favor of open data within the scientific community is that openness promotes transparency, reproducibility, and more rapid advancement of new knowledge and discovery. Indeed, many scientific journals and funding agencies now require that researchers make experimental data publicly available; however, they remain divided over what steps researchers must take to protect individuals' privacy before releasing data sets in the open, and regulatory uncertainty only exacerbates these problems.

Emerging Trends

So far, we have given a fairly conventional account of the deidentification debate. Rather than take sides or offer a more detailed analysis of disputed issues, this next section briefly considers three emerging trends that might suggest a way to advance the discussion: the idea of identifiability as a continuum; the broad support for risk-based approaches to deidentification; and the signs of convergence between pragmatists and formalists. All three trends provide additional context for the papers presented at today's symposium.

Identifiability as a Continuum

Our first trend is the growing recognition among privacy scholars regarding the inadequacy of any strictly binary distinction between identifiability and non-identifiability. Critics like Ohm (and others) insist that regulators abandon this distinction completely, arguing that the list of potential identifiers is inexhaustible and "will never stop growing until it includes everything."[7] However, most commentators now agree that identifiability and anonymity are better understood as end-points on a wide spectrum with many interim states that pose a variety of graduated privacy risks.[8] Thus, they call for revising and refining the concept of identifiability rather than abandoning it. Five years ago, Schwartz and Solove proposed a tripartite classification that distinguishes information depending on whether it refers to (1) an identified person, (2) an identifiable person, or (3) a non-identifiable person.[9] Moreover, they argue that the applicability of the Fair Information Privacy Principles (FIPPs) should turn on these categories. While all FIPPs generally should apply to information that refers to an identified person, only data quality, transparency, and security should apply to identifiable data.[10] Their approach requires an *ex ante,* probabilistic and contextual assessment of which of the three categories a given data set falls into (and hence how it should be treated in terms of the FIPPs).[11]

More recently, Polonetsky, Tene and Finch suggested a new conceptualization that recognizes multiple gradations of identifiability across a much broader spectrum of personal data.[12] Their scheme—which is also available as a visual guide to practical data deidentification[13]—distinguishes ten gradations or categories of data depending upon the treatment of *direct identifiers*, *indirect identifiers*, and *safeguards or controls* (which include both internal and external controls).[14] The first two three components may be either intact, partially masked, eliminated or transformed, while the third may be either not in place, or limited or in place. Using these distinctions, they arrange the ten categories into four main groupings: First,

[7] Ohm, *supra* note 2, at 1742.

[8] This has long been the preferred way for computer scientists to understand these terms; *see, e.g.,* Andreas Pfitzmann and Marit Hansen, A terminology for talking about privacy by data minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management (Version v 0.34)(2010), http://dud.inf.tu-dresden.de/Anon_Terminology.shtml.

[9] Paul M. Schwartz & Daniel J. Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, 86 NEW YORK UNIV. L. REV. 1814, 1870-72, 1877-79 (2011)(characterizing a person as "identified" when her identity is "ascertained" or he or she can be "distinguished" from a group; as "identifiable" when specific identification is "not a significantly probable event" (i.e., the risk is low to moderate); and as "non-identifiable" when the risk of identification is no better than "remote").

[10] *Id.* at 1879-83 (notice, access, and correction rights would not apply, while the authors are silent on the remaining FIPPs).

[11] *Id.* at 1878 (noting that this assessment depends on "the means likely to be used by parties with current or probable access to the information, as well as the additional data upon which they can draw" as well as additional contextual factors such as "the lifetime for which information is to be stored, the likelihood of future development of relevant technology, and parties' incentives to link identifiable data to a specific person"). *See* Rubinstein & Hartzog, *supra* note 3, at 53 (suggesting changes in this method of assessment due to its failure to treat the public availability of released data as an overriding factor in assigning data sets to categories 1, 2, or 3).

[12] Polonetsky, Tene & Finch, *supra* note 1.

[13] *See* https://fpf.org/wp-content/uploads/2016/04/FPF_Visual-Guide-to-Practical-Data-DeID.pdf.

[14] Internal controls "encompass security policies, access limits, employee training, data segregation guidelines, and data deletion practices" while external controls "involve contractual terms that restrict how partners use and share information, and the corresponding remedies and auditing rights to ensure compliance that aim to stop confidential information from being exploited or leaked to the public." Polonetsky, Tene & Finch, *supra* note 1, at 606.

personal data with different degrees of identifiability (i.e., information containing direct and indirect identifiers); second, pseudonymous data (i.e., information from which direct identifiers have been eliminated or transformed but indirect identifiers remain intact); third, deidentified data (i.e., data sets from which direct and indirect identifiers have been removed or manipulated to break the linkage to real world identities); and, finally, anonymous data (i.e., data sets from which direct and indirect identifiers have been removed or manipulated and mathematical and technical guarantees to prevent re-identification have been applied).

Although somewhat complex as compared with binary or tripartite schemes, their more granular approach to the identifiability spectrum provides a sound basis for addressing several legal conundrums. For example, regulators and firms (especially advertisers) have long disputed whether unique identifiers (such as IP addresses) are personally identifiable. A binary response might lead to the wrong analysis by ignoring the gradations in the identifiability spectrum and thereby failing to distinguish among subtly different cases. As Polonetsky, Tene and Finch rightly observe, "if an identifier can be cleared by a user, its dissemination and retention controlled, and strong technical and legal constraints prevent it from being linked to personal information," then it should warrant more flexible legal treatment in which some obligations apply but not others.[15] The co-authors also stress the importance of considering all relevant factors in classifying "key-coded" data (i.e., personal data that have been stripped of direct identifiers and replaced by a key to avoid unwanted or unintended reidentification).[16] In the hands of the curator who holds the key, or researchers with access to the key, the data are clearly personal data. As to third parties, if there are strong controls limiting key access to approved researches only and the method of securing the key is sufficiently strong to thwart an attack by a determined adversary, then key-coded data should not be treated as personal data but rather as non-identifiable data, at least in the hands of those who cannot unlock it. As the co-authors also point out, however, a strict reading of the "any other person" language in Recital 26 undermines their analysis by imputing reidentification to third parties "who do not hold a key, based on the capabilities of the party who first coded the data."[17]

A very recent decision by the European Court of Justice sheds a little more light on how to read Recital 26 but also leave some questions unanswered. On October 16, 2016 the Court ruled that dynamic IP addresses in the hands of a website may constitute "personal data" even where only a third party (i.e., an Internet Service Provider) has the additional data (such as its customer's name and address) necessary to identify the individual.[18] In reaching this conclusion, the Court compared two ways of interpreting the following italicized language in Recital 26, which states "to determine whether a person is identifiable, account should be

[15] *Id.* at 611-13.

[16] *Id.* at 613-14. As the co-authors note, *id.* at 614, "Key-coded data are used extensively in a range of circumstances where limited re-identification is necessary or desirable, including pharmaceutical research, …. For example, in clinical trials, health institutions typically must maintain an ability to link research data back to specific patients, in order to alert them of a treatable condition they discover or contain the spread of an infectious disease").

[17] *Id.* at 614.

[18] *See* judgment in Case C-582/14, *Patrick Breyer v Bundesrepublik Deutschland*, ECLI:EU:C:2016:77.

taken of all the *means likely reasonably* to be used either by the controller or by *any other perso*n." Academics refer to these two approaches as the "absolute/objective approach" and the "subjective/relative approach."[19] The former treats IP addresses as "personal data" if *any* third party (including an ISP over whom a website lacks legal authority) is able to determine the identity of the individual, while the latter would reach this conclusion only if a website has the legal and practical means (and not merely an abstract possibility) of obtaining the additional identifying information from the third party.[20]

While a literal reading of Recital 26 suggests the absolute approach, the Court seemingly embraced the relative approach. First, the Court observed that combining a dynamic IP address with the additional identifying information held by the ISP would not constitute a means likely to be used to identify the data subject "if the identification of the data subject was prohibited by law or practically impossible on account of the fact that it requires a disproportionate effort in terms of time, cost and man-power, so that the risk of identification appears in reality to be insignificant."[21] Next, it stated that under German law, the website in question has the legal means, particularly in the event of a cyber-attack, to contact the competent authority and through it obtain additional identifying information from the ISP in order to bring criminal proceedings.[22] It therefore concluded that, in this case, the website "has the means which may likely reasonably be used in order to identify the data subject, with the assistance of other persons, namely the competent authority and the internet service provider, on the basis of the IP addresses stored."[23] The Court did not, however, explicitly reject the absolute approach and, so we will have to await further developments in law and policy before we can know the full range of circumstances under which the Court would apply (or not apply) the relative approach.

A Risk-Based Approach to Deidentification

Our second trend is the broad agreement—at least among academics—that instead of focusing on anonymisation as a perfect end-state that prevents all privacy harm, the law and policy of identifiability should be "designed around the processes necessary to lower the risk of reidentification and sensitive attribute disclosure."[24] This risk-based perspective predominates among both critics[25] and defenders[26] of deidentification and also receives support from leading

[19] *See* F.J. Zuiderveen Borgesius, *Singling Out People without Knowing Their Names - Behavioural targeting, pseudonymous data, and the new Data Protection Regulation*, 32 COMPUTER L. & SECURITY REV. 256, 263-65 (2016).

[20] *See* Case C-582/14, *Patrick Breyer v Bundesrepublik Deutschland,* par. 25.

[21] *Id.,* par. 46.

[22] *Id.,* par. 47.

[23] *Id.,* par. 48.

[24] Rubinstein & Hartzog, *supra* note 3, at 729.

[25] *See, e.g.,* Ohm, *supra* note 2, at 1761 (recommending that regulators focus on factors "that help reveal the risk of reidentification and threat of harm") and identifying five factors for reducing such risks, *id*. at 1764-68.

[26] *See, e.g.,* Khaled El Emam & Bradley Malin, Appendix B: Concepts and Methods for De-identifying Clinical Trial Data, in SHARING CLINICAL TRIAL DATA: MAXIMIZING BENEFITS, MINIMIZING RISK 240-43 (Inst. of Med. ed., 2015) (describing an eleven-step, risk-based process for deidentifying data).

data scientists and experts in statistical disclosure control.[27] Some computer scientists—especially those who seek provable privacy guarantees using formalistic methods such as differential privacy— remain skeptical of the risk-based approach.[28] But as discussed below, estimating privacy risk and exploring privacy-utility trade-offs is becoming more central to this community as well. Regulators (mostly) follow a risk-based approach too although, in some cases, they long for more certainty (the Article 29 Working Party) or settle for less (the HIPAA safe harbor standard). These differences are worth exploring in a bit more detail by contrasting the current European vs. U.S. regulatory approaches.

In Europe, the Data Protection Directive defines an identifiable person as "one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity",[29] and the General Data Protection Regulation (GDPR), which takes effect in May 2018, largely retains this definition. As to data not meeting this broad definition, Recital 26 of the Directive clarifies that its provisions "shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable." As previously highlighted, this recital further states that "to determine whether a person is identifiable, account should be taken of all the *means likely reasonably to be used*" to identify them,[30] while the corresponding recital in the GDPR offers even more explicit language indicating that we should view anonymisation in terms of a risk-based reasonableness test.[31]

And yet there is a tension in European data protection law between this reasonableness test and what appears to be more of an "impossibility" standard, according to which data may be rendered anonymous only when it is "retained in a form in which identification of the data subject is *no longer possible*."[32] Recent guidance from the Article 29 Data Protection Working Party regarding anonymisation techniques fails to resolve this tension. On the one hand, the Working Party assesses anonymisation primarily in terms of the strengths and weaknesses of various technical measures, quite explicitly framing this exercise in terms of "the residual risk"

[27] *See* Mark Elliot, Elaine Mackey, Kieron O'Hara & Caroline Tudor, THE ANONYMISATION DECISION MAKING FRAMEWORK (2016), http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf, (describing a new holistic approach to anonymisation that provides an end to end methodology for assessment of risk and control of reidentification).

[28] *See, e.g.,* Cynthia Dwork & Rebecca Pottenger, *Towards Practicing Privacy*, 20 J. AM. MED. INFORMATICS ASS'N 102, 102 (2013),(dismissing deidentification as a "sanitization pipe dream"); NARAYANAN & FELTEN, *supra* note 6.

[29] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, OFFICIAL JOURNAL L 281, 23/11/1995 P. 0031 – 0050 ("Data Protection Directive"), Art. 2(a).

[30] Data Protection Directive, Recital 26 (emphasis added).

[31] *See* Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), [2016] OJ L119/1, Recital 26 (stating that "To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.").

[32] Data Protection Directive, Recital 26 (emphasis added).

of identification inherent in each of them.[33] On the other hand, it sometimes conceptualizes anonymisation as requiring a zero (or near-zero) probability of reidentification; it must be "irreversible."[34] As others have argued, any such standard is not only impractical but also conflicts with a risk-based approach, which can never eliminate risk entirely.[35]

In the U.S., regulators are also mostly adhering to a risk-based approach. This is very clear from the Federal Trade Commission's three-part test,[36] and the HIPAA expert determination standard,[37] but less so in the safe harbor standard, which offers a very straightforward method for achieving legally-recognized deidentification (by removing eighteen enumerated data elements) rather than requiring data controllers to engage in an individualized risk assessment based on the specific facts and circumstances of a given data release.[38] Recent NIST guidance also recommends that government agencies contemplating a data release evaluate the risks arising from releasing deidentified data and offers detailed guidance on how to conduct a risk assessment.[39]

Convergence Between Pragmatists and Formalists

Our third trend consists in some preliminary but welcome signs of convergence between pragmatists and formalists over the need for a more flexible and holistic approach to data

[33] Article 29 Data Protection Working Party, *Opinion 5/2014 on Anonymisation Techniques*, 0829/14/EN WP 216 (April 10, 2014), http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommenda-tion/files/2014/wp216_en.pdf, p. 7.

[34] *Id.* at 5 & 7; *see also id.* at 8 (referring to an earlier opinion that "clarified that the 'means . . . reasonably to be used' test" helps assess whether any given anonymisation process is sufficiently robust, i.e., "whether identification has become 'reasonably' impossible'"). If not an oxymoron, this phrase ("reasonably impossible") betrays some logical confusion.

[35] *See, e.g.,* Khaled El Emam & Cecilia Álvarez, *A Critical Appraisal of the Article 29 Working Party Opinion 05/2014 on Data Anonymization Techniques,* 5 INTERNATIONAL DATA PRIVACY LAW 73 (2015) (suggesting that a zero-risk approach has practical disadvantages and otherwise rejecting the notion that achieving zero risk of reidentification in anonymized data is a legal requirement under European law).

[36] *See* FEDERAL TRADE COMMISSION, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE 21 (2012), https://www.ftc.gov/sites/de-fault/files/documents/reports/federal-trade-commission-report-protecting-con-sumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf (stating that data is not "reasonably linkable" to the extent that a company: (1) takes reasonable measures to ensure that the data is deidentified; (2) publicly commits not to try to reidentify the data; and (3) contractually prohibits downstream recipients from trying to re-identify the data).

[37] *See* HIPAA Privacy Rule, 45 C.F.R. § 164.514(b)(1) (requiring an expert determination using "generally accepted statistical and scientific principles and methods" of deidentification to establish that there is a "very small" risk that the deidentified information "could be used, alone or in combination with other reasonably available information, . . . to identify an individual who is a subject of the information").

[38] *See* 45 C.F.R. § 164.514(b)(2).

[39] GARFINKEL, DEIDENTIFICATION OF PERSONAL INFORMATION, *supra* note 5, at 12-16. It is encouraging to note that the National Committee on Vital and Health Statistics, Subcommittee on Privacy, Confidentiality & Security recently held hearings at which Garfinkel and several symposium participants testified on the need for policy changes in the HIPAA deidentification rule. *See* Hearings on De-Identification and the Health Insurance Portability and Accountability Act (HIPAA) (May 24, 2016), http://www.ncvhs.hhs.gov/meeting-calendar/agenda-of-the-may-24-25-2016-ncvhs-subcommittee-on-privacy-confidentiality-security-hearing/.

releases. Just six years ago, Paul Ohm published his justly famous law review article alerting the legal profession to the failure of anonymisation in the face of ever more powerful methods for mounting linkage attacks. But Ohm's analysis focused almost exclusively on the "release and forget" model in which data custodians release deidentified data to the public usually after singling out and modifying identifying information by means of suppression, generalization, or aggregation.[40] Notably, Ohm's work—and the contretemps it inspired—largely ignored alternative data release models, such as the data use agreement model, the simulated dataset model, and the enclave model.[41] But different models in fact take different approaches, depending on the goals of the research and intended uses of the data sets, the relevant risks involved (including the risk of reidentification), and the associated harms.[42] More recent work from the UK Information Commissioner's Office[43] and a group at the University of Manchester,[44] demonstrates the necessity of deploying different anonymisation techniques based on assessing reidentification risk in context and identifying the right tool for the job at hand. By way of illustration, the University of Manchester group's "Anonymisation Decision-Making Framework" embodies a "total system approach" consisting in ten components:

1. Describe your data situation
2. Understand your legal responsibilities
3. Know your data
4. Understand the use case
5. Meet your ethical obligations
6. Identify the processes you will need to assess disclosure risk
7. Identify the disclosure control processes that are relevant to your data situation
8. Identify who your stakeholders are and plan how you will communicate
9. Plan what happens next once you have shared or released the data
10. Plan what you will do if things go wrong.[45]

The framework supports data custodians who need to understand the correct level of anonymisation to apply in a specific situation. Component 7 draws heavily on statistical disclosure control methods including both non-perturbative methods (such as sampling, choice of variables, and level of detail) and perturbative methods (such as data swapping, overimputation, rounding, cell and value suppression, and *k*-anonymity).[46] Additionally, it considers environmental controls governing who has access to the data, what analyses may or

[40] Ohm, *supra* note 2, at 1711-16.

[41] *See* SIMSON L. GARFINKEL, NAT'L INST. OF STANDARDS & TECH., DEIDENTIFICATION OF PERSONAL INFORMATION (NISTIR 8053) DE-IDENTIFYING GOVERNMENT DATASETS (DRAFT NIST SPECIAL PUBLICATION 800-188) 18-19 (2016).

[42] Rubinstein & Hartzog, *supra* note 3.

[43] United Kingdom, Information Commissioner's Office, *Anonymisation: Managing Data Protection Risk Code of Practice* (2012), http://tinyurl.com/ICO-ANON.

[44] Elliot, Mackey, O'Hara & Tudor, *supra* note 27.

[45] *Id.* at 3-4, 67-118.

[46] *Id.* at 43-52.

may not be conducted, where the data access/analysis may be carried out, and how access is obtained.[47]

This emerging holistic approach is by no means wedded to the simple deidentification methods associated with the release and forget model. Furthermore, and along somewhat similar lines, a group at Harvard University comprised of experts in computer science, social science, statistics, and law, has set itself the task of refining and developing definitions and measures of privacy and data utility, and at the same time designing an array of technological, legal, and policy tools for social scientists to use when dealing with sensitive data.[48] This project is notable not only for its multidisciplinary approach but for combining (1) techniques for estimating privacy risk and measuring and defining utility with (2) an intensive effort to design and test a variety of algorithms for "privacy-preserving" analysis and sharing of data. These include algorithms for statistical estimation, managing the privacy budget, synthetic data generation, and data summaries, all of which draw upon recent advances in differential privacy. [49] The Harvard project is holistic insofar as it seeks to facilitate and complement these computational privacy tools with a variety of legal instruments, including "custom policies, licenses, contracts, and other legal agreements carefully tailored to the specific needs of researchers (and their subjects) working with specific types of data under different technical approaches."[50]

The Symposium Panels

Today's symposium brings together experts from multiple disciplines including law, computer science, statistics, engineering, social science, ethics and business to address a range of topics from technology, open data, and pseudonymisation to regulation, policy, and ethics. Their work reflects the emerging trends discussed above.

We have organized the papers into four topical panels.

The first panel covers "Deidentification Frameworks" and has papers by Dr. Mark Elliot, Dr. Elaine Mackey, and Dr. Kieron O'Hara, by Orit Levin and Javier Salido, and by Dr. Micah Altman, David R. O'Brien, Urs Gasser, and Alexandra Wood.

- In a paper entitled "The Anonymisation Decision Making Framework," Elliot et al. describe a new, holistic approach to anonymisation that provides an end to end methodology for assessment of risk and control of reidentification. This framework incorporates legal, ethical, policy and statistical insights. While it has been developed in the context of the current UK regulatory environment, it also provides valuable insights for interpreting the GDPR.

[47] *Id.* at 52-60.

[48] *See* Salil Vadhan, Gary King, Latanya Sweeney, Edoardo Airoldi & Urs Gasser, Project Description: *Privacy for Social Science Research*, National Science Foundation (NSF) Award No. 1237235 (September 19, 2012), http://privacytools.seas.harvard.edu/files/privacytools/files/projectdescription_1.pdf?m=1363618082.

[49] *Id.* at 10.

[50] *Id.* at 13.

- In their paper entitled “The Two Dimensions of Data Privacy Measures,” Levin and Salido describe a practical framework for use with big data. It begins by considering two factors in the design of data privacy measures: the desired data utility (with its corresponding deidentification techniques), and the anticipated data sharing scenarios (with the corresponding feasible data security measures). It then examines the different levels of potential risk to data subjects for possible combinations of deidentification techniques and data sharing scenarios, with the goal of guiding practitioners in their design of deidentification measures that comply with the GDPR.
- Finally, in their paper entitled “Practical Approaches to Big Data Privacy Over Time,” Altman et al. examine a range of long-term data collections in social science research and identify the characteristics of these programs that drive their unique sets of risks and benefits. They argue that many uses of big data, across academic, government, and industry settings, have characteristics like those of traditional long-term research studies. They discuss the lessons that can be learned from longstanding data management practices in such research and potentially applied in the context of newly emerging data sources and uses.

The second panel covers “Risk-Based Approaches” and has papers by Khaled El Emam, Eloise Gratton, Jules Polonetsky and Luk Arbuckle, by Monica Dias, Frank Petavy and Alessandro Spina, and by Gergely Acs, Claude Castelluccia, and Daniel Le Metayer.

- In a paper entitled “The Seven States of Data,” El Emam et al., map the spectrum of identifiability to a risk-based approach to deidentification based on practices in the statistical disclosure control community. They seek to define precise criteria for evaluating the different levels of identifiability and propose a new point on this spectrum that would allow broader uses of pseudonymous data under certain conditions.
- In their paper entitled “Notes on the Anonymization of Clinical Study Reports for the Purpose of Ensuring Regulatory Transparency,” Dias et al. discuss the scientific methodology and the technical and legal challenges for the anonymization of clinical data. Their work reflects the recently developed guidance of the European Medicines Agency (EMA) on on the anonymisation of clinical reports in the pharmaceutical industry.
- Finally, in their paper entitled “Testing the Robustness of Anonymisation Techniques: Acceptable versus Unacceptable Inferences,” Acs et al. take issue with the risk-based criteria put forward by the Article 29 Working Party in its “Opinion on Anonymization Techniques,” namely, “singling out,” “linkability,” and “inference.” The co-authors argue that these risk-based criteria are neither necessary nor effective in deciding on the robustness of an anonymization algorithm. They propose an alternative approach relying on the notions of acceptable versus unacceptable inferences, which is based on a newly developed technique they call “differential testing.”

The third panel covers “New Perspectives” and has a diverse set of papers by Dr. Daniel C. Barth-Jones, Gemma G. Clavell and Iris Huis in ’t Veld, and by Dr. Nicola Jentzsch.

- In a paper entitled, “Why a Systems-Science Perspective is Needed to Better Inform Data Privacy Deidentification Public Policy, Regulation and Law,” Barth-Jones argues in favor of a systems perspective to better understand how multidimensional technical and regulatory interventions can effectively combine to create practical controls for countering wide-spread reidentification threats. He rejects the “dystopic” vision of Ohm and other critics of deidentification because their work ignores important underlying mathematical realities regarding information entropy and signal detection theory. Finally, he suggests that systems modeling and quantitative policy analyses, including uncertainty analyses, provide the necessary scientific tools to critically evaluate the potential impacts of pseudonymisation and anonymisation in various regulatory schemes.
- In their paper entitled “Tailoring Responsible Data Management Solutions to Specific Data-Intensive Technologies: A Societal Impact Assessment Framework,” Clavell and Huis in ’t Veld develop a societal impact assessment (SIA) framework tailored to data-intensive technologies. Unlike other similar assessment tools, an SIA framework is designed to evaluate the risks, externalities and consequences of technologies, policies, programs, and systems, taking account of a wide range of concerns and stakeholders. The four main pillars of the SIA framework are desirability, acceptability, ethics, and data management, and the paper explores the lessons learned from implementing this approach in several real-life technologies and projects involving anonymisation.
- Finally, in a paper entitled “Competition and Data Protection Policies in the Era of Big Data: Privacy Guarantees as Policy Tool,” Jentzsch considers how different concepts of identifiability help expand the tools available to data protection <u>and</u> competition authorities in supervising firms that rely on Big Data and personalization. Her paper raises novel questions regarding mergers between data-rich firms. In particular, it addresses whether dominance in data might undermine competition, and whether competition authorities might condition mergers of data-rich firms on certain privacy guarantees to ensure that pre-merger promises and post-merger actions are properly aligned.

The final panel covers “Law and Policy” and has papers by Dr. Sophie Stalla-Bourdillon and Alison Knight, Dr. Waltraut Kotschy, and Michael Hintze. All three authors analyze and critique the provisions on anonymisation and pseudonymisation and related policies in European data protection law, especially the GDPR.

- In a paper entitled “Anonymous data v. Personal data—A False Debate: An EU Perspective on Anonymisation, Pseudonymisation and Personal Data,” Stalla-Bourdillon and Knight call attention to terminological and doctrinal ambiguities in how the Data Protection Directive and the GDPR have defined anonymisation and related terms. Their analysis identifies a static approach to anonymisation as the main reason for shortcomings in several recent regulatory positions and instead develops a more dynamic understanding of whether anonymized data is likely to be reidentified based on the purpose of any further processing and future data linkages, as well as any obligations assumed by third parties with whom the data has been shared.

- In a paper entitled “Identifiability: Policy and Practical Solutions for Anonymization and Pseudonymisation,” Kotschy also focuses on weaknesses in the GDPR definitions of anonymisation and pseudonymisation. Relying on Austrian law as a helpful model, she considers what it means for any given technique of anonymisation and pseudonymisation to be “sufficiently safe” to achieve the policy goals of the relevant GDPR provisions and perhaps confer greater legal advantages on parties who satisfy emerging standards of safe treatment.
- Finally, in a paper entitled “Viewing the GDPR Through a De-Identification Lens: A Tool for Clarification and Compliance,” Hintze takes issue with the binary approach to deidentification and instead distinguishes four levels of identifiability. He refers to these as: (1) identified, (2) identifiable, (3) Article 11 deidentified, and (4) anonymous / aggregated. Hintze argues that EU regulatory guidance should be more attuned to these different levels and illustrates his point by analyzing various obligations under the GDPR (including notice, consent, access, data retention limitations, and data security) through the lens of this simplified deidentification spectrum.