Does the GDPR restrict the dissemination of de-identified data?

Connor Colson

Laws regulating data protection are popping up internationally, federally, and in states. These laws, typified by the European Union's adoption of the GDPR, impose restrictions on the collection, distribution, and processing of an individual's personal data. Yet these laws almost uniformly apply only to data which can be tied back to an individual, thus allowing de-identified or anonymous data to be shared without restriction.[1] This de-identification has been the "main paradigm used in research and elsewhere to share data while preserving people's privacy."[2]

This regulatory environment has enabled the free flow of information, revolutionizing the way businesses and governments function. This trends show every sign of continuing into the future, as the "big data" revolution is expected to grow exponentially[3] in coming years due to the developing world coming online, the internet of things, AI and cloud computing.[4] However with the rise in big data the capacity for large scale de-anonymization has also improved.[5] In recent years, individuals have been re-identified by browser histories, medical records, hospital discharge data, taxi trajectories in NYC, bike sharing trips in London, subway data in Riga, and mobile phone and credit card datasets.[6]As the technology for large scale collection and use of

[1] Victor Richardson, Sallie Milam, and Denise Chrysler, *Is Sharing De-identified Data Legal? The State of Public Health Confidentiality Laws and Their Interplay with Statistical Disclosure Limitation Techniques*, 43 Journal of Law, Medicine, & Ethics, 83, 83 (2015); C. Christine Porter, *De-Identified Data and Third Party Data Mining: The Risk of Re-Identification of Personal Information,* 5 Shidler J.L. Com. & Tech. 3, 4 (Sep. 23, 2008). http://www.lctjournal.washington.edu/Vol5/a03Porter.html.

[2] Luc Rocher, Julien M. Hendrickx & Yves-Alexandre de Montjoye, *Estimating the success of re-identifications in incomplete datasets using generative models*, 10 NATURE COMMUNICATIONS 1, 2 (2019), https://doi.org/10.1038/s41467-019-10933-3.

[3] https://www.sciencedirect.com/science/article/abs/pii/S0306437914001288

[4] Rocher et al., *supra* note 2, at 2.

[5] Groden, Samantha, Summer Martin, and Rebecca Merrill. *Proposed Changes to the Common Rule: A Standoff Between Patient Rights and Scientific Advances?* Journal of Health & Life Sciences Law. Retrieved March 26, 2017.

[6] Rocher et al., *supra* note 2, at 2.

individuals data continues to expand, it is worth assessing whether existing de-identification practices are sufficient to forestall legal liability.

Specifically, the new spat of privacy regulation, including the European General Protection Regulation (GDPR) and California Consumer Protection Act (CCPA), raise the standard for anonymization dramatically, requiring "each and every person in a dataset has to be protected for the dataset to be considered anonymous."[7] Additionally the GDPR directly addresses the risk of de-anonymization, specifically stating that once information is deemed "vulnerable" or only "pseudonymized" the protections afforded to anonymous data fall away.[8] Due to its wide breadth, recent adoption, and unique recognition of re-identification risk, we are looking at the effect of this paper and other re-identification advances on compliance with GDPR.

One of the most significant defenses for the existing anonymization procedures is the theory of "plausible deniability".[9] Essentially because released datasets are virtually always incomplete, then "journalists and researchers can never be sure they have re-identified the right person even if they found a match."[10] The practice of releasing only a subset or sampled portion of the total data attempts to strike a balance between providing enough information to be useful, while covering a small enough portion of individuals that virtually anyone specifically identified could reasonably claim there is a "data doppelgänger" outside of the released sample.

This plausible deniability defense has been a cornerstone of modern de-identification, providing cover for organizations which release or process personal data.[11] Previous statistics

[7] *Id.*
[8] *Id.*
[9] Luc Rocher, Julien M. Hendrickx & Yves-Alexandre de Montjoye, *Estimating the success of re-identifications in incomplete datasets using generative models*, 10 NATURE COMMUNICATIONS 1, 2 (2019), https://doi.org/10.1038/s41467-019-10933-3.
[10] Rocher et al., *supra* note 2, at 2.
[11] Rocher et al., supra note 2, at 2.

based approaches to confirm a positive identification were population level estimators.[12] These require significant researcher manpower, complete datasets, and accurate population level statistics, posing a practical barrier to wide scale re-identification.[13]

The recent paper by the lab of Yves-Alexandre de Montjoye titled “Estimating the success of re-identification in incomplete datasets using generative models” lays out a method to get around the limitations on population level estimators. For purposes of this paper it is sufficient to state that these researchers created a mathematical formula for the likelihood that an individuals record is unique in the complete population, “fed” the algorithm with publicly available sources of data to train it, and modeled the resulting distribution to extrapolate this data to the remainder of the population not included in the data disclosure.[14] The advantage of this approach is that its conclusions are mathematically verifiable, effective on heavily sampled or incomplete datasets,[15] reproducible across a wide variety of datasets with minimal alteration, and has been demonstrated to score 39% better than the best theoretically achievable prediction using only population uniqueness.[16]

The approach laid out within this research paper significantly degrades the defense of plausible deniability in response to the threat of de-anonymization. As the amount of data collected continues to increase and algorithms improve, it is expected that re-identification will keep getting easier.[17] While prior studies have managed to effectively re-anonymize some individuals despite measures to anonymize datasets,[18] these advancements have not yet resulted

[12] *Id.*
[13] *Id.*
[14] *Id.* at 2-6.
[15] Rocher et al., *supra* note 2, at 3.
[16] *Id.*
[17] Groden, Samantha, Summer Martin, and Rebecca Merrill. *Proposed Changes to the Common Rule: A Standoff Between Patient Rights and Scientific Advances?* Journal of Health & Life Sciences Law. Retrieved March 26, 2017.
[18] C. Christine Porter, *De-Identified Data and Third Party Data Mining: The Risk of Re-Identification of Personal Information,* 5 Shidler J.L. Com. & Tech. 3, 4 (Sep. 23, 2008).

in large scale de-anonymization of public datasets. This is partly because sufficient demographic information tied to name are not openly available in useable formats, and even with the Yves-Alexandre de Montjoye's method for confirming a positive identification, collecting the requisite information manually is time intensive and thus limited to a subset of the larger dataset.

However, with this tools capacity to eliminate plausible deniability even with incomplete or irregular data, there are two ways that GDPR could expose companies to legal risk. (1) While named data is harder to find publicly than anonymous data, companies and data brokers around the world have access to exactly the named type of datasets necessary for larger scale de-identification. Repositories like customer lists, medical data, tax information, government agencies, or social media companies collect the type of information necessary to link individuals across databases on a larger scale. The companies themselves could open themselves up to GDPR liability by misusing this information to re-identify, inadvertently transferring vulnerable information to a bad actor, or through a data breach. (2) In the second scenario, the technology remains fundamentally limited to de-identifying individuals in small quantities, but because these identifications are now verifiable, companies are open to GDPR liability even though only a small subset of the dataset was affected.

Having established that the recent advancement in re-identification by Yves-Alexandre de Montjoye's lab could expose data holding companies to liability, we explore the legal mechanics by which this could occur in relation to the GDPR.

To discuss why GDPR is unique in its recognition of re-identification risk, it is necessary to compare its approach to other information privacy laws. The United States tends to regulate narrow classes of special data which are restricted, such as Health Data (HIPAA), Children (COPPA), or Credit Information (FACTA). These privacy laws, typified by HIPPA, have avoided the problems posed by re-identification through methods such as the Safe Harbor Provision.[19] These provisions lay out a set of best practices, which if followed and certified by an expert, provide an effective defense to subsequent litigation.[20] This defense is sufficient even if the "covered entity" knew of studies or methods by which the data could be re-identified, because the receiver of the information is not presumed to have the capacity to do with less than "actual knowledge".[21] This is all in service of the principle that entities should have a simple metric to determine if de-identification procedures were sufficient.[22]

Put into practice, this concept is illustrated by *Southern Illinoisan v. Illinois Department of Public Health*. In this case, the Illinois Department of Public Health refused to provide patient information based on state confidentiality law "precluding disclosure of . . . the identity, or any group of facts which tends to lead to the identity, of any person. . ." alongside expert testimony that patients could be identified by matching the data requested with publicly available data.[23] The court held that while experts were capable of de-anonymizing this information, this expertise was rare, and the relevant question was whether "a member of the general public could perform the multi-step procedure to match identities."[24] This approach standardizes data protection best practices, encourages compliance by providing "safe harbor" protection, enables efficient

[19] Victor Richardson, Sallie Milam, and Denise Chrysler, *Is Sharing De-identified Data Legal? The State of Public Health Confidentiality Laws and Their Interplay with Statistical Disclosure Limitation Techniques*, 43 Journal of Law, Medicine, & Ethics, 83, 84-85 (2015).

[20] HEALTH AND HUMAN SERVICES, GUIDANCE REGARDING METHODS FOR DE-IDENTIFICATION OF PROTECTED HEALTH INFORMATION IN ACCORDANCE WITH THE HEALTH INSURANCE PORTABILITY AND ACCOUNTABILITY ACT (HIPAA) PRIVACY RULE. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#idrisk

[21] *Id*. at https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#specificstudies

[22] *Id*. at https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#specificstudies

[23] Victor Richardson, Sallie Milam, and Denise Chrysler, *Is Sharing De-identified Data Legal? The State of Public Health Confidentiality Laws and Their Interplay with Statistical Disclosure Limitation Techniques*, 43 Journal of Law, Medicine, & Ethics, 83, 83 (2015);

[24] *Id*. at 84.

transfer of data, and provides effective safeguards to de-anonymization.[25] Because this system provides litigation protection by statute to adherents,[26] companies are much less vulnerable to sudden technology improvements like the one discussed in Yves-Alexandre's article.

In comparison, the GDPR fundamentally approaches data privacy protection differently. The GDPR is a regulation in EU law on data privacy for all data subjects within the territory of the European Union and the European Economic Area.[27] This regulation was implemented on May 25, 2018 and had far reaching implications for international companies. Similarly to previous laws like HIPAA, it exempted truly anonymous data from regulation,[28] but uniquely it directly addresses the problem of de-anonymized data.[29] Specifically, it introduces the concept of "pseudonymous data", which is data that does not contain obvious identifiers but could be re-identified through the use of additional external information.[30] Rather than providing protection to those who unsuccessfully de-identify data, the GDPR explicitly states that data which can be re-identified is equivalent at law to unprotected "personal data."[31]

Secondly, when assessing whether an individual can be re-identified the GDPR looks to "all the means reasonably likely to be used . . . either by the controller or by another person to identify the natural person directly or indirectly."[32] In asking what means are "reasonably likely" to be used, the test is objective: What are the costs and time required for identification?[33] This

[25] *Id.*
[26] *Id.*
[27] GDPR Art. 3 §2. https://gdpr-info.eu/art-3-gdpr/
[28] GDPR Recital 26. https://gdpr-info.eu/recitals/no-26/
[29] GDPR Recital 26. https://gdpr-info.eu/recitals/no-26/
[30] Luc Rocher, Julien M. Hendrickx & Yves-Alexandre de Montjoye, *Estimating the success of re-identifications in incomplete datasets using generative models*, 10 NATURE COMMUNICATIONS 1, 4 (2019), https://doi.org/10.1038/s41467-019-10933-3.
[31] GDPR Recital 26. https://gdpr-info.eu/recitals/no-26/
[32] *Id.*
[33] *Id.*

analysis requires that all available technologies at the time of processing are considered, and that the potential for technological developments is a fundamental part of the analysis.[34]

When contrasted with the blanket immunity provided by HIPAA the results are striking. The GDPR does not concern itself with best practices, instead setting out an objective test for compliance which requires companies to stay abreast of technological changes.[35] Because of this difference between these two styles of law, companies are at uniquely increased litigation and compliance risk from de-identification for GDPR than comparable statutes such as HIPAA.

> The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. 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. The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.[36]

Beyond recognizing the potential for de-identification, this section states that once "pseudonymized" data can be attributed to an individual then the protection afforded to anonymized data is negated.[37] Coupled with the GDPR's assertion that "that each and every person in a dataset has to be protected for the dataset to be considered anonymous,"[38] it appears

[34] *Id.*
[35] *Id.*
[36] *Id.*
[37] *Id.*
[38] Luc Rocher, Julien M. Hendrickx & Yves-Alexandre de Montjoye, *Estimating the success of re-identifications in incomplete datasets using generative models*, 10 NATURE COMMUNICATIONS 1, 2 (2019), https://doi.org/10.1038/s41467-019-10933-3.

that the GDPR has acknowledged the potential for de-anonymization, while stating that the de-anonymization of even a single individual within the dataset removes the protection anonymization provides.[39] Referencing back to the underlying advancement in de-anonymization tech discussed in Yves-Alexandre's article, the continual advancements in identification technology do not need to fundamentally reshape the status of anonymous data to put companies at significant compliance risk. The GDPR is known to have significant fines to enforce the privacy of EU "data subjects", and even a conservative reading of the technological advancement in the last thirty years suggests that private and public datasets are vulnerable enough to allow a small number of positive identifications through. Due to the drafting of the GDPR, these small errors are all that is required to open firms up to liability.

Reading further the statute proposes a method to determine whether data should be considered de-anonymized.[40] The determination asks companies to ascertain what means are "reasonably likely to be used."[41] This test does not ask companies to simply take "reasonable precautions" and instead proposes a broad objective analysis based on the time, cost, and technological developments available at the time of processing.[42] This analysis specifically looks to the state of technology, rather than industry best practices or statutory requirements.[43] Secondly the statute makes the data provider liable for the actions reasonably likely to be undertaken by the controller or "by another person."[44] Finally, the statute requires companies to consider the available technology "at the time of processing."[45] The definition of processing within the GDPR is a term of art, but generally encompasses all manner of touching personal

[39] GDPR Recital 26: https://gdpr-info.eu/recitals/no-26/.
[40] GDPR Recital 26: https://gdpr-info.eu/recitals/no-26/.
[41] *Id.*
[42] *Id.*
[43] *Id.*
[44] *Id.*
[45] *Id.*

data, and is not limited to first contact. When taken together, these sections suggest that organizations providing anonymized datasets needs to (1) assess the current state of de-anonymization technology objectively, (2) forecast reasonable technological advancements, (3) reasonably predict the behavior of both the controllers it provides data too and all other unrelated parties, (4) at the moment of processing rather than the moment of data production.[46]

By the language of the GDPR assessed above data providers are not insulated from the potential for de-anonymization by adherence to industry best practices, or statute, or the defense that the data was anonymized when it was provided.[47] Instead the GDPR explicitly acknowledges the potential for re-identifying data, states that data which is pseudorandomized but possible to trace back to individuals with external information has the same legal requirements as personal data, and charges companies with conducting an objective analysis of current and upcoming technology before sharing anonymous data.[48] In addition, it requires them to forecast not only the potential abuse by controllers but also action by "another person".[49] Finally, it states that if a single individual can be positively identified then the entire dataset loses its anonymity protection, and individuals which provide this data remain liable during the period which the information is available for "processing."[50] Taken together this is a dramatic expansion on the traditional United States conception of privacy laws, with a myriad of new and complicated obligations required to stay in compliance in the face of constantly improving re-identification technology.

[46] *Id*.
[47] *Id*.
[48] *Id*.
[49] *Id*.
[50] *Id*.

Companies face a significant and complicated issue in complying with GDPR for its datasets, and the relative newness of the regulations mean that there is vanishingly little caselaw on the subject. As such we will engage in a hypothetical to illustrate the potential for liability and misuse which these technological advancements pose. A fairly average scenario would occur when a midsize company with extensive contacts in Europe collects a significant volume of sales and customer data from the region. This information contains a lot of demographic data of the customers, their spending habits, and their zip codes if they had ever placed an order on the site.

Presumably the company recognizes the value of this dataset, and presuming that the absence of personally identifiable information such as names, phone numbers, social security numbers, or addresses were included, decides to share the information with business partners to inform future advertising spending. This use is allowed under the GDPR, but the organization is charged with the responsibility to conduct an objective analysis of the likelihood that this information will be personally attributable during the course of the partnership. Having read about advancements in re-identification and discussed the topic with his business partners, the owner is aware of the potential for experts in the field to recombine information based on a variety of data points, but he believes that his business is protected from liability because at the time the data was shared his business partners lacked the capacity or know how to re-identify the data.

In this scenario, the owner of this business could still be charged under the GDPR for distribution of personal data if the secondary company intentionally or inadvertently de-identified individuals from the dataset. (1) the owners lack of “actual knowledge” that his associates were capable of re-identification is irrelevant, as the test is objective.[51] (2) The

[51] GDPR Recital 26: https://gdpr-info.eu/recitals/no-26/.

moment the judge needs to look at in assessing the adequacy of data anonymization is not the moment the data was shared, but instead the moment at which the data was "processed."[52] This change means that companies have to follow up with places they have loaned data to, as they are accountable for prompt compliance at the moment of "processing" rather than the moment of "transfer."[53] (3) Finally the Company cannot seek protection by claiming only a small subset of the total dataset was compromised, because the GDPR explicitly states that any dataset which includes pseudorandomized information is equivalent to unprotected personal data at law.[54]

In short this legislation, in comparison to earlier privacy laws, leaves a large gap within which significant technological advancements in re-identification have the potential to destabilize existing "anonymous" dataset designations. The recent paper discussed above provides exactly the type of advancement in re-identification that the GDPR provides for, explaining why the authors themselves stated that "even heavily sampled anonymized datasets are unlikely to satisfy the modern standards for anonymization set forth by GDPR and seriously challenge the technical and legal adequacy of the de-identification release-and-forget model."[55] At minimum, this paper exposes the risk that companies subject to GDPR incur by providing inadequately anonymized datasets. Looking to the future as these technologies continue to improve, regulators will have to either find a way to maintain anonymity, restrict access to potentially useful data, or decide that the value provided by open data policies outweighs individuals interest in privacy.

[52] *Id*.
[53] *Id*.
[54] *Id*.
[55] Luc Rocher, Julien M. Hendrickx & Yves-Alexandre de Montjoye, *Estimating the success of re-identifications in incomplete datasets using generative models*, 10 NATURE COMMUNICATIONS 1, 1 (2019). https://doi.org/10.1038/s41467-019-10933-3.