Chasing the Golden Goose: What is the path to effective anonymisation?

by Omer Tene[1], Gabriela Zanfir-Fortuna[2]

PinG (Forthcoming 2017)

Abstract

Searching for effective methods and frameworks of de-identification often looks like chasing the Golden Goose of privacy law. For each answer that claims to unlock the question of anonymisation, there seems to be a counter-answer that declares anonymisation dead. In an attempt to de-mystify this race and un-tangle de-identification in practical ways, the Future of Privacy Forum and the Brussels Privacy Hub joined forces to organize the Brussels Symposium on De-identification - "Identifiability: Policy and Practical Solutions for Anonymisation and Pseudonymisation". The event brought together researchers from the US and the EU, having academic, regulatory and industry background, discussing their latest solutions for such an important problem. This contribution looks at their work in detail, puts it in context and aggregates its results for the essential debate on anonymisation of personal data. The overview shows that there is a tendency to stop looking at anonymisation/identifiability in binary language, with the risk-based approach gaining the spotlight and the idea of a spectrum of identifiability already generating practical solutions, even under the General Data Protection Regulation.

Key-words: anonymisation, identifiability, privacy, personal data, pseudonymisation

I. Introduction

De-identifying personal data can very well represent a Golden Goose for protecting privacy and other rights of those whose data make up immense databases, while allowing the use of that data for unlimited purposes. The benefits of anonymisation are significant. For instance, framing this discussion under EU data protection law is clear: if a controller is processing data that has been de-identified so as to become anonymous, then the data protection regulatory framework does not apply to that processing operation because the data is not personal and, hence, does not fall in the material scope of data protection law. This principle, recognized under Directive 95/46[3], is also spelled out in the General Data Protection Regulation[4] (GDPR), under Recital 26:

> "The principles of data protection should therefore not apply to anonymous information, namely information that does not relate to an identified or

[1] Senior Fellow, Future of Privacy Forum.

[2] PhD; Fellow, Future of Privacy Forum.

[3] 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, OJ L 281, 23/11/1995 P. 0031 – 0050; see Recital 26.

[4] 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), OJ L 119/1, which will become applicable on 25 May 2018.

identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable."

The same stands true for most privacy laws worldwide, because their scope of application is defined based on whether information is identifiable or not[5]. However, in practice things are not at all as clear as they may seem in legal wording. Numerous studies have shown that re-identifying de-identified data, as well as identifying an individual using different categories of data points is usually possible with the appropriate tools[6]. Should, then, anonymisation be considered unachievable?

Recent guidance from the Information Commissioner's Office (ICO) suggests that the answer to this question may not be relevant after all: "It may not be possible to establish with absolute certainty that an individual cannot be identified from a particular dataset, taken together with other data that may exist elsewhere. The issue is not about eliminating the risk of re-identification altogether, but whether it can be mitigated so it is no longer significant. Organisations should focus on mitigating the risks to the point where the chance of reidentification is extremely remote"[7]. Furthermore, the regulator sees the value of anonymisation techniques beyond taking processing operations outside the scope of data protection laws: "it is also a means of mitigating the risk of inadvertent disclosure or loss of personal data"[8]. In other words, even if data protection or privacy laws would apply to data that has been "reversibly anonymised", it would still pay off for organisations to anonymise the data they are processing. It then becomes essential to understand to what extent and how could compliance mechanisms be adjusted to accommodate processing of data that undergo "reversible anonymisation".

The French Supreme Administrative Court (*Conseil d'Etat*) recently dealt with the question of whether processing personal data that is subject to two specific de-identification techniques, "hashing" and "salting", would still allow individuals to be entitled to exercise their rights as data subjects[9]. The case concerned monitoring of MAC addresses of mobile phones by JCDecaux, through their panels showing ads in a Parisian public market. The French DPA (*CNIL*) did not authorize this processing operation because the controller did not provide mechanisms for the exercise of the data subjects, claiming that it anonymises the data to the extent that the French data protection law is not applicable[10]. The Court upheld the decision of the CNIL. The main argument of the French judges was that even if the "*hashing* and *salting* techniques have the purpose to obstruct access of third parties to that data, they allow the data controller the possibility to identify the data subjects and they do not prohibit correlation of records related to the same individual, or inferring information about

[5] *I. Rubinstein* in his Framing the Discussion paper of the Brussels Privacy Symposium on Identifiability: Policy and Practical Solutions for Anonymisation and Pseudonymisation.

[6] See, for instance, *Y.-A. de Montjoye, C. A Hidalgo, M. Verleysen, V. D Blondel*, Unique in the Crowd: the Privacy Bounds of Human Mobility, Nature Scientific Reports, Volume 3, 2013; *Paul Ohm*, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, 57 UCLA Law Review 1701, 1717-23, 2010; *Alessandro Acquisti, Ralph Gross*, Predicting Social Security Numbers from Public Data, Proceedings of the National Academy of Science, July 7, 2009; *Pierangela Samarati, Latanya Sweeney*, Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression, Technical Report SRI-CSL-98-04, 1998 and its second version *Latanya Sweeney*, K-Anonymity: A Model For Protecting Privacy, 10 (5) International Journal of Uncertainty, Fuzziness & Knowledge-based Systems 557, 2002.

[7] ICO, "Big data, artificial intelligence, machine learning and data protection" Report, 1 March 2017, Paragraph 134.

[8] Idem, Paragraph 139.

[9] Conseil d'État, 10ème – 9ème ch. réunies, Decision of 08.02.2017, "JCDecaux France".

[10] Conseil d'État, "JCDecaux France", paragraph 3.

him or her"[11]. The Court considered that the purpose of the processing operation (monitoring the behavior of passersby, including measuring the repetitiveness of their walking-byes and the pattern of their movements between ad panels) is incompatible with processing anonymised data[12], and therefore the claim of the controller that it processes anonymous data is not substantiated.

The area between what is personal and what is anonymous convincingly looks like quicksand, and the legal implications of understanding where in that area the processed data stands are momentous. Contributions presented and discussed within the Brussels Symposium on De-identification, organized[13] by the Future of Privacy Forum and the Brussels Privacy Hub substantially inform this debate.

Rubinstein provided a comprehensive framework to initiate the discussions, summarizing two decades of scholarship and policymaking on anonymisation/de-identification, exploring the visions of formalists, pragmatists and those who plead for convergence[14]. Rubinstein asks poignant questions – "Should we define these terms in binary fashion or are they better understood as the end-points of a wide spectrum?"; "Given the inevitable tradeoffs between privacy and data utility, are there optimal ways to balance these competing interests?"; "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?"

The contributions selected for the Summit tackled these questions, organized in four panels, which also delineate the structure of this paper, starting with analyzing practical (II) de-identification frameworks, followed by a closer look to (III) risk-based approaches, a discussion on (IV) new perspectives and (V) law and policy developments, with a focus on the GDPR. The conclusions (VI) will show that there is a tendency to stop looking at anonymisation/identifiability in binary language, with the risk-based approach gaining the spotlight and the idea of a spectrum of identifiability[15] already generating practical solutions.

II. De-identification frameworks

1. A ten-steps framework to anonymisation understood as a "risk management process"

Mackey, Elliot and O'Hara introduced their "Anonymisation Decision-making Framework" (ADF), which "attempts to unify the technical, legal, social and ethical aspects of anonymisation to provide a comprehensive guide to doing anonymisation in practice".

Their framework is built around five underpinning principles. The first one informs that one "cannot decide whether data are safe to share or not by examining the data alone". This means that practitioners will need to assess whether a set of data is anonymised in relation with the environment of that data. The second principle

[11] Conseil d'État, "JCDecaux France", paragraph 8 (unofficial translation).

[12] Conseil d'État, "JCDecaux France", paragraph 8.

[13] 8 November 2016 in Brussels.

[14] Available here https://fpf.org/wp-content/uploads/2016/11/Rubinstein_framing-paper.pdf (last time visited on 9 March 2017).

[15] For an analysis of the spectrum of identifiability, see *J. Polonetsky, O. Tene and K. Finch*, Shades of Gray: Seeing the full spectrum of practical data de-identification, in Santa Clara Law Review, vol. 56, 2016.

asserts that, notwithstanding the first one, the data still needs to be examined, together with the context. According to the third principle, “anonymisation is a process to produce safe data but it only makes sense if what you are producing is safe useful data”. According to the fourth principle, “zero risk is not a realistic possibility if you are to produce useful data”, therefore anonymisation “is best understood as a risk management process”. The last principle shows that the measure one puts in place to manage re-identification risk “should be proportional to the risk and its likely impact”.

The ADF enshrines ten components, clustered in three core anonymisation activities: (1) a data situation audit – understanding the processing operation, its context, the legal obligations and the ethical dimension, (2) risk analysis and control – assessing disclosure risks and identifying the disclosure control processes that are relevant to your data situation and (3) impact management – identifying who the stakeholders are, planning further steps after anonymised data was shared and having a back-up plan if anything goes wrong after sharing data.

2. Choosing the appropriate de-identification technique based on the data-sharing scenario used

Levin and Salido propose in their paper “The Two Dimensions of Data Privacy Measures” a framework that would help data controllers choose the most effective de-identification technique for their datasets without factoring in the nature or content of data. The authors claim the grid they propose leads to identifying the appropriate de-identification technique across different industries.

Their model comprises eleven de-identification techniques applied to nine sharing scenarios. The authors classify the effectiveness of each technique for each scenario as “conservative”, “optimal”, “risky”, “inappropriate”, “for future study” or “not applicable”. For instance, masking of identifiers is considered risky if access to data is provided under a Service Level Agreement or contract. But, if it were applied in the case where access to data is provided within a legal entity, masking of data would be an optimal de-identification technique.

The authors encourage regulators to use their model and “define a sufficiently protected area within the two axes such that the level of data protection inside this area would be considered sufficient in the eyes of data subjects and regulators and applicable to a wide range of industries and use cases”. The terminology, classification and understanding of the characteristics of known techniques for de-identification of tabular data used for the paper were sourced from ISO/IEC JTC1 CD 20889 “Privacy enhancing data de-identification techniques” (which is currently under debate).

3. Borrowing best de-identification practices from researchers and their datasets

In their paper “Practical Approaches to Big Data Privacy Over Time”, Altman, Wood, O’Brien and Gasser look at de-identification techniques and other privacy protections deployed by researchers to their datasets, aiming to inform commercial and government actors on best practices that have been tested by the research community. The authors argue that “many uses of big data, across academic, government, and industry settings, have characteristics similar to those of traditional long-term research studies”. Starting from this hypothesis, they look in depth to how researchers have been deploying different combinations of privacy controls to their datasets.

Even if they found that the characteristics of using big data for research purposes and for commercial or governmental purposes are similar, the authors show that "the review processes and safeguards employed for long-term data collection and linkage activities in commercial and government settings differ from those used in the research context in a number of key respects". For instance, "commercial and government actors often rely heavily on certain approaches, such as notice and consent or de-identification, rather than drawing from the wider range of privacy interventions that are available and applying combinations of tailored privacy controls at each stage of the information lifecycle, from collection, to retention, analysis, release, and post-release".

The impact of time on privacy should play a more prominent role when deciding which are the most effective de-identification techniques. As highlighted in the paper, "key risk drivers for big data that are related to the time dimension include the age of the data, the period of collection, and the frequency of collection".

In their concluding remarks, the authors recommend "using a combination of controls to manage the overall risk resulting from identifiability, threats and vulnerabilities", pointing out that "several clusters of controls for addressing identifiability and sensitivity can be implemented, such as notice, consent, and terms of service mechanisms in combination with robust technical disclosure limitation techniques, formal application and review in combination with data use agreements and disclosure limitation techniques, and secure data enclaves with auditing procedures".

III. Risk-Based Approaches

1. Introducing "Flexible pseudonymous data" in the spectrum of identifiability

In their paper "The Seven States of Data: When is Pseudonymous Data not Personal Information?", El Emam, Gratton, Polonetsky and Arbuckle define the spectrum of identifiability and specific criteria for the placement of different types of data along this spectrum. They use a risk-based approach for evaluating identifiability which is consistent with practices in the disclosure control community. Using precise criteria for evaluating the different levels of identifiability, the authors proposed a new point on this spectrum that would allow broader uses of pseudonymous data under certain conditions.

The initial six states of data identified reflect the type of data sharing that is happening today, based on the authors' observations: public release of anonymized data, quasi-public release of anonymized data and non-public release of anonymized data qualify as "not-PII" (not-personally identifiable information), while protected pseudonymized data, "vanilla" pseudonymized data and raw personal data qualify as "PII". The first three states of data refer mainly to types of open data, as well as data that requires qualified access. *Protected pseudonymous data* refers to data where "only masking of direct identifiers has been applied and no de-identification methods are used", but which have "additional contractual, security, and privacy controls in place". *"Vanilla" pseudonymous data* is "pseudonymous data without any of the additional contractual, security or privacy controls in place", while *raw personal data* refers to "data that has not been modified in any way or that has been modified so little that the probability of re-identification is still very high".

The authors define in their paper three specific criteria that would further reduce the risk of re-identification for protected pseudonymous data: "(1) No processing by humans; (2) No PII leakage from analytics results; and (3) No sensitive data." Data that comply with these criteria would be "*flexible pseudonymized data*", an intermediary category between not-PII and PII, which would not require consent for processing.

In conclusion, by adding more conditions and safeguards to the existing state of protected pseudonymous data, the authors propose that "more flexibility can be granted for the use and disclosure of the data while still being consistent with contemporary risk management frameworks".

2. Testing the robustness of anonymization techniques with a machine learning process

In their paper "Testing the Robustness of Anonymisation Techniques: Acceptable versus Unacceptable Inferences", Acs, Castelluccia and Le Metayer dismantle the guidance issued by European Data Protection Authorities on anonymisation techniques[16], by deeming the criteria laid out there as neither necessary, nor effective to decide upon the robustness of an anonymisation algorithm. The criteria put forward by the Article 29 Working Party in their 2014 Opinion referred to the following risks a data controller should consider: singling out, linkability and inference.

The authors consider that the criteria are not necessary "because they do not take into account the type of information that can be derived. In some cases, this information may actually be insignificant, noisy or even useless". As for their effectiveness, they consider that "it depends very much on the precise meaning of inference". According to their assessment, "the only way to make this criterion meaningful would be to qualify it and consider inferences of attributes about specific individuals with sufficient accuracy", which would lead to a threshold issue – "where should the red line be put to decide upon 'specific' and 'sufficient'".

The ability to perform inferences is "the key issue with respect to both privacy and utility". The authors believe that "there are acceptable and unacceptable disclosures: 'learning statistics about a large population of individuals is acceptable, but learning how an individual differs from the population is a privacy breach'". However, they acknowledge that certain group inferences "can still be harmful, which means that the release of the resulting anonymized dataset should still be reviewed and controlled by a privacy ethics committee".

The main challenge identified is "to provide criteria to distinguish between acceptable and unacceptable inferences". The solution found by the authors is to use a machine learning process, called "differential testing", to predict "the sensitive attribute of users (attributes that are usually not quasi-identifiers but rather represent some information not to be revealed about the user such as medical diagnosis, salary, locations, etc.)".

[16] Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymization Techniques, adopted 10 April 2014.

3. Anonymisation – key to publishing Clinical Study Reports by the European Medicines Agency

Spina, Dias and Petavy presented their ongoing work for the paper "Notes on the anonymisation of Clinical Study Reports for the purpose of ensuring regulatory transparency". The European Medicines Agency (EMA) adopted a Policy on the publication of clinical data for medicinal products for human use in 2014[17]. EMA started to publish clinical data submitted by pharmaceutical companies to support their regulatory applications for human medicines under the EU centralised procedure, on the basis of the Policy, in October 2016[18].

The Policy generally refers to "ways and means to anonymise data and protect patients from retroactive identification". In order to implement this, EMA developed a guidance document addressed to pharmaceutical companies on the anonymisation of clinical reports. The paper aims to discuss the scientific methodology and the technical and legal challenges for the anonymisation of clinical study reports.

IV. New Perspectives

1. Pleading for a Systems-Science perspective to better inform the de-identification public policy

In his paper "Why a Systems-Science perspective is needed to better inform data protection de-identification public policy, regulation and law", Barth-Jones argues that "data privacy policy for de-identification must take a systems perspective in order to better understand how combined multi-dimensional (i.e., involving both technical de-identification and administrative/regulatory responses) interventions can effectively combine to create practical controls for countering widespread re-identification threats".

The author makes the case that "rumors of de-identification's death have been greatly exaggerated". He identifies the main reasons for this formalist approach – "the vast majority of re-identification demonstrations have been conducted against data without any proper statistical disclosure limitation methods applied, or have blatantly ignored the impact of disclosure controls where they have been applied"; and the fact that "it assumes as a default that the actors and forces of re-identification are omnipresent, omniscient, omnipotent and relentless". Barth-Jones seconds the conclusion of Rubinstein and Harzog, who argued that the first law of privacy policy is that "there are no silver-bullet solutions" and that the best way to move policy past the purported failures of anonymisation is to instead focus on the process of minimizing the risk of re-identification[19].

Therefore, Barth-Jones pleads for de-identification to be given a fair chance, using "improved re-identification research steps, combined with the use of systems modeling and quantitative policy analyses including uncertainty analyses". These

[17] "European Medicines Agency policy on publication of clinical data for medicinal products for human use", EMA/240810/2013, 2 October 2014.

[18] According to information available on the website of the institution. See http://www.ema.europa.eu/ema/?curl=pages/special_topics/general/general_content_000555.jsp (last time visited on 9 March 2017).

[19] *Ira Rubinstein* and *Woodrow Hartzog*, Anonymization and Risk, Washington Law Review, June 2016 91(2):703-760.

methods can provide “the necessary scientific tools to critically evaluate the potential impacts of pseudo/anonymisation in various regulatory schemas and should be pursued routinely when conducting data privacy policy evaluations”.

2. Bringing the human dimension to anonymisation

Galdon Clavell and in’t Veld build a framework to assess the societal impact of data intensive technologies, which they deem to be “sensitive both to the technological and economic concerns of engineers and decision-makers and to societal values and legislation”. The purpose of their paper, “Tailoring Responsible Data Management Solutions to Specific Data-Intensive Technologies: A Societal Impact Assessment Framework”, is to provide policy-makers and engineers with the tools to think about ethics and technology and lead them “towards value-sensitive and privacy-enhancing solutions like anonymisation”.

The authors recall that “data relates to human beings with rights and values”. Therefore, “aspects of legality, ethics, desirability, acceptability and data management policy have to be critically considered in order to make sure that rights and values are respected”. The proposed framework is called “Eticas” and it has four pillars: Law and Ethics, Desirability, Acceptability and Data Management.

The Law and Ethics dimension “relates to the legal and moral standards guiding a project and results in the preconditions for a project in a specific field”. It focuses on the relevant legislation and the social values that are involved in a specific context. The Desirability dimension “refers to the justification of the need for a technology or its specific functionalities” and it involves a clear “problem definition”. The purpose is to avoid “technological solutionism”. The Acceptability dimension “involves the inclusion of public opinion and values in a technological innovation or research project”. The outcome of stakeholder consultations could be implemented in the design process. Finally, the Data Management dimension refers to the legal framework of privacy and data protection, ethical principles, but also to broader considerations relating to individual control and consent, methods of anonymisation, and how privacy issues can be designed into technologies and projects.

The authors conclude that the Eticas framework is malleable, because “it can be adapted to different systems and contexts, as well as to the resources of the organizations performing the assessment”. However, they acknowledge that its success “depends on a genuine commitment from all stakeholders”, particularly from technology designers, “which should adopt a mind-shift from technology inventors to solution providers”, while considering the values, needs and expectations of the communities beyond their user base.

3. De-identification as policy tool for Data Protection Authorities and Competition Authorities

Jentzsch explores the complicated environment at the interaction of competition law and data protection law in the era of Big Data, looking specifically at how “privacy guarantees” can enable “a more effective monitoring of industry players”, both from the perspective of Data Protection Authorities (DPAs) and of Competition Authorities (CAs).

In his paper, “Competition and data protection policies in the era of Big Data: Privacy Guarantees as Policy Tools”, Jentzsch starts from the assumption that “information asymmetries are a key ingredient for competition”, because they protect trade secrets and they induce uncertainty about the competitors' innovations and future movements. The author observes that the increasing complexity of analytical methods used by companies creates transparency challenges, in the sense that firms are now able to monitor consumers and rivals in an unprecedented manner. This is why he argues that "we need to discuss how some of the recently developed privacy guarantees can be utilized as tools for upholding information asymmetries needed to ensure competition."

Jentzsch looks at how anonymisation of databases can play a part in evaluating mergers and preventing the abuse of a dominant position by CAs. "Authorities in charge for enforcing legislation relating to unfair commercial practices can use the 'degree of differentiation' spectrum to prosecute any misleading promises of firms regarding anonymisation of data. (…) For example, in merger cases, authorities need to define the relevant market (product-wise, geographic and temporal), before assessing dominance and its anticompetitive effects. If a merger creates or strengthens a dominant position stifling competition, it might be prohibited. Databases play a critical role in the merger of data-intensive firms or in evaluating the abuse of a dominant position." The author develops specific recommendations for both DPAs and CAs to use different privacy guarantees as policy tools. For instance, he proposes that CAs "should condition a merger of data-rich firms on provable privacy guarantees", such as "randomization and/or generalization or preventing linkability of the data".

One of the conclusions of the study is that using privacy guarantees for supervision provides an incentive for companies "to use de-personalized information to a greater extent in order to avoid scrutiny by supervisors". Moreover, "such deployment could spur investments in the development of more *efficient privacy guarantees and mechanisms*."

V. Law and policy

1. Looking at the incentives under the GDPR to anonymise and pseudonymise personal data

Kotschy analyses in his paper - "The new General Data Protection Regulation: Is there sufficient pay-off for taking the trouble to anonymize or pseudonymise data?", whether there are sufficient incentives for data controllers to anonymize and pseudonymise data in the framework of the new General Data Protection Regulation. He assesses all provisions and recitals of the GDPR relevant to the two processes and concludes that while using anonymised data results in clear, significant, consequences – "the GDPR is not applicable", the rewards for using pseudonymised data are not that clear. There are "no precise legal consequences", the author observes, pointing out that "the 'pay-off' for pseudonymisation in data protection has not (yet) been fully exploited".

The paper provides insight into how the Austrian data protection law differentiates between personal data and "indirectly personal data" – a concept introduced in 2000. These are still personal data, but they identify the data subject only indirectly, "in the sense that additional information would be needed to reveal the full identity of the data subject". According to the author, "all identifiers which together directly identify this person (such as the name, date of birth, residence etc.) are encrypted and the user of such data has no access to the encryption algorithm".

Kotschy explains that, under the Austrian law, using “indirectly personal data” triggers “several privileges for the controllers involved”, such as having “no obligation to notify the processing of indirectly personal data to the DPA, no restriction for disclosing such data to third parties, no obligation to obtain permission from the DPA for transfers to third countries, no obligation to inform the data subjects about transfers to third parties”. In addition, “access rights of data subjects are suspended”. This is not the case under the GDPR, as Kotschy points out.

2. Proposing a fluid line between personal data and anonymised data, with a dynamic approach to anonymisation

Framing the debate under the GDPR, Stalla-Bourdillon and Knight argue in their paper “Anonymous data v. Personal data—A false debate: An EU perspective on anonymisation, pseudonymisation and personal data”, that the state of anonymised data should be comprehended dynamically: “anonymised data can become personal data again, depending upon the purpose of the further processing and future data linkages, implying that recipients of anonymised data have to behave responsibly”. They claim that the “attempts” of EU data protection regulators to clarify the terms of the dichotomy personal data/anonymised data “have partly failed”.

The authors analyze the guidance issued by the ICO and the Article 29 Working Party on anonymisation techniques, as well as the legal requirements within Directive 95/46 and the GDPR with regard to anonymisation and the definition of personal data. They argue that, even if the Article 29 WP is “sympathetic to a risk-based approach”, its position is problematic because it “suggests that an acceptable re-identification risk requires near-zero probability, an idealistic and impractical standard that cannot be guaranteed in a big data era”. Looking at the provisions of the GDPR, the authors point out that, at least in its Preamble, the regulation adopts a risk-based approach to anonymisation, relying on the test of “means reasonably likely to be used” by the data controller and third parties to identify a data subject. They consider it is necessary to “revisit the very concept of personal data as defined under EU law” in order to fully understand the implications of a dynamic approach to anonymisation.

Their argument is that identifiability is not the only key component of the concept of personal data, another component equally important being the context in which the personal data are processed, or the “relate to” component of the definition. To support their claim, the authors refer to the *Breyer*[20] case, where the Advocate General Campos Sanchez-Bordona considered that, indeed, “context is crucial for identifying personal data, and in particular characterizing IP addresses as personal data”[21]. The Court followed the same approach, as it excludes identifiability “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.”[22]

The authors conclude that “a dynamic approach to anonymisation therefore means assessing the data environment in context and over time and implies duties and obligations for both data controllers releasing datasets and dataset recipients”. They

[20] CJEU, Case C-582/14, Breyer v Bundesrepublik Deutschland, 19.10.2016, ECLI:EU:C:2016:779.

[21] Opinion of the Advocate General Campos Sánchez-Bordona, CJEU C-582/14, Breyer v Bundesrepublik Deutschland, 12.05.2016, ECLI:EU:C:2016:339, at [68].

[22] CJEU, Case C-582/14, Breyer v Bundesrepublik Deutschland, 19.10.2016, ECLI:EU:C:2016:779, at [46].

also acknowledge that more research is necessary in the field to fully comprehend the variety of categories of processing and the interplay between the different components of data environments.

3. Making the case for de-identification as key for GDPR compliance

In his paper “Viewing the GDPR Through a De-Identification Lens: A Tool for Clarification and Compliance”, Hintze makes a compelling analysis of the implications of de-identifying data for compliance with the GDPR, arguing that de-identification brings significant incentives for data controllers to comply with key requirements under the EU data protection law framework: lawful grounds for processing (in particular consent and legitimate interests), notice, data retention, data security, as well as data subject rights of access, deletion and other controls.

He identifies four levels of identifiability, looking at the provisions of the GDPR: identified data, identifiable data, Article 11 De-identified data and anonymous/aggregate data.

Identified data “identifies or is directly linked to data that identifies a specific natural person (such as a name, e-mail address, or government-issued ID number).” Identifiable data “relates to a specific person whose identity is not apparent from the data; the data is not directly linked with data that identifies the person; but there is a known, systematic way to reliably create or re-create a link with identifying data. Pseudonymous data as defined in the GDPR is a subset of Identifiable data.” Article 11 De-identified data “may relate to a specific person whose identity is not apparent from the data; and the data is not directly linked with data that identifies the person”, while anonymous/aggregate data “is (1) stored without any identifiers or other data that could identify the individual or device to whom the data relates; and (2) aggregated with data about enough individuals such that it does not contain individual-level entries or events linkable to a specific person.”

The author argues that, for instance, “Article 6(4) of the GDPR supports the idea that de-identification can be used to help justify a basis for lawful processing other than consent”. As for the notice obligation – he suggests that “the more strongly de-identified the data is, the more likely discoverable notice will be appropriate”, which means that an individualized Notice for each kind of processing operation will not be required by the supervisory authorities.

Hintze also draws attention to the fact that “Article 12(2) of the GDPR specifies that if the controller can demonstrate that it is not in a position to identify the data subject (i.e., Article 11 De-Identified data), it need not comply with Articles 15 to 22. Those articles include the right of access (Article 15), rectification (Article 16), erasure (Article 17), data portability (Article 20), and the right to object to the processing of personal data or obtain a restriction of such processing under certain circumstances (Articles 18 and 21)”.

A substantial conclusion of the article is that “the GDPR requirements in each area should be interpreted and enforced in a way that will encourage the highest practical level of de-identification and that doing so will advance the purposes of the regulation”.

VI. Conclusion

The difficult questions surrounding anonymisation and identifiability are not going anywhere soon. As showed in the introductory part of this paper, the questions started to appear in Courts and regulators are paying more and more attention to them. With the entering into force of the GDPR and its vast (extra)territorial application, finding good and practical answers is more important than ever.

The "De-identification frameworks" proposed by the papers debated at the Brussels Privacy Symposium do just that. They describe possible practical solutions, organized in frameworks that understand anonymisation as a risk management process. One fundamental idea they have in common is that the assessment for identifying the most effective anonymisation technique should give more weight to the environment or context where that data is processed than to the content of the data itself (Subsections I.1 and I.2). On another hand, researchers suggest looking for inspiration at the tested de-identification methods used in research for decades to handle big data sets. A key ingredient for the effectiveness of these methods is factoring in the impact of time on privacy – the age of the data, the period of collection and the frequency of collection (Subsection I.3).

The "Risk-based approach" to anonymisation was further explored by authors who put efforts into classifying data throughout the de-identification spectrum. A new category of anonymized data that could allow broader uses of pseudonymous data was identified and defined – "flexible pseudonymous data" (Subsection II.1). A machine learning process called "differential testing" was proposed to be able to distinguish between acceptable and unacceptable inferences made from pseudonymised data, after the authors explained that the ability to perform inferences is the key issue with respect to both privacy and utility of data (Subsection II.2). Finally, a case study was presented as example of a risk based approach to anonymisation applied in practice – the disclosure of Clinical Study Reports made by pharmaceutical companies in Europe (Subsection II.3).

"New perspectives" were also proposed, ranging from a systemic approach referring to multi-dimensional interventions (technical and administrative/regulatory responses) that can effectively combine to create practical controls for countering widespread re-identification threats (Subsection III.1), to an Impact Assessment Framework for data intensive technologies that takes into account moral standards, ethical values and the needs of communities (Subsection III.2), to analyzing the significant role anonymisation can play in the ever more complex interaction of data protection law and competition law (Subsection III.3).

Finally, the last contributions looked closely into the provisions of the GDPR and their significance for the anonymisation/identifiability debate. One of the questions looked into was whether there are sufficient incentives under the GDPR for controller to anonymise and pseudonymise the data they process (Subsection IV.1). The concept of a fluid line between personal data and anonymised data was introduced. It was claimed that identifiability is not the only key component of the concept of personal data, context in which the personal data are processed being another important component. The authors brought arguments from the recent case-law of the CJEU to support this idea (Subsection IV.2). Furthermore, a strong argument was made that de-identification techniques are fundamental to compliance with the GDPR. Looking closely to key GDPR provisions, including Articles 11, 12(2) and 6(4) it was argued that de-identification brings significant incentives for data controllers to comply with a series of key requirements, such as notice, data retention and data security (Subsection IV.3).

Concluding, the anonymisation/identifiability debate seems to significantly shift towards a risk-based approach understanding, which includes paying more attention to the spectrum of identifiability and to identifying concrete compliance mechanisms with privacy and data protection law for processing pseydonymised data.