A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques

Khaled El Emam* and Cecilia Álvarez**

Introduction

On 10 April 2014 the Article 29 Working Party issued an opinion, 05/2014, on data anonymization techniques[1] (henceforth the 'Opinion'). The Opinion was anticipated by many organizations as an authoritative guidance on the methods to use for anonymizing personally identifying information. The purpose of the current paper is to provide a critical appraisal of the Opinion, and to interpret some of its recommendations in the context of prior Working Party opinions, regulations in other jurisdictions, and current best practices.

We review what we consider to be the key themes in the Opinion, and for each provide commentary and an appraisal. The themes are not presented in an order of priority or importance, but were sorted to ensure a logical flow to our arguments.

Below is a summary of key points we make in this article:

- It needs to be made clear that acceptable re-identification risk is not zero risk.
- Privacy ethics councils are necessary to oversee data uses. That is the practical way to manage inference risks.
- Automated methods to protect against attribute disclosure (learning something new from the data) ought not be recommended as they will reduce data quality significantly, and have rarely, if ever, been used in practice for that reason.
- Anonymization is considered a permitted/compatible use.
- The data destruction stipulation in the Opinion needs to be deleted from any recommendations as this will not work in practice.
- Adversaries should be precisely defined as an anticipated data recipient and consider specific attacks.
- The stipulation in the Opinion to protect against incorrect re-identification should be deleted from any recommendations as this will not work in practice.

Summary

- The Article 29 Working Party opinion 05/2014, issued on 10 April 2014 on data anonymization techniques, has provided clarification and important interpretation guidance on some topics, but did not advance understanding with some other critical topics.
- The key recommendations found in the Opinion need to be appraised from a practical standpoint to determine whether they are consistent with current best practice in the disclosure control community, can be practically implemented, and continue to ensure that privacy is protected in a defensible way.
- We provide a broader perspective on the issues raised in the Opinion to help interpret them in a practical manner which will facilitate the disclosure of high-quality data while continuing to safeguard the privacy of EU citizens.

* Khaled El Emam is founder and CEO of Privacy Analytics Inc., a spin-off company founded in 2007 to commercialize the research efforts of the Electronic Health Information Laboratory (EHIL) of the University of Ottawa and CHEO Research Institute. Privacy Analytics provides organizations with enterprise software to safeguard and enable data disclosure for secondary purposes. Canada Research Chair in Electronic Health Information, University of Ottawa and Children's Hospital of Eastern Ontario Research Institute, Ottawa, Ontario, Canada. kelemam@uottawa.ca.

** Counsel, Uria Menéndez, Madrid, Spain.

[1] Article 29 Data Protection Working Party, *Opinion 05/2014 on Anonymisation Techniques*, 10 April 2014.

- Pseudonymous data should not be treated as anonymous data, as the Opinion states.
- Linkability for the purpose of linking records that belong to the same subject to create a longitudinal profile must be permissible. Prohibiting that will ensure that most data (longitudinal data) are useless.
- Theoretical techniques that have not been demonstrated to work broadly in practice should not be recommended, especially by regulatory authorities or by bodies that are treated as representative of regulators.

Critical Appraisal

Objectives of the Opinion

The stated objective of the Opinion was to document the 'effectiveness and limits of existing anonymisation techniques',[2] and as such it does not provide a methodology for data controllers or data processors to follow to anonymize their data. It addresses specific legal and policy issues and descriptively reviews a number of technical methods for anonymization. As such, there is no step-by-step guidance in terms of how to go about data anonymization.

The Opinion does acknowledge the value and importance of anonymization. It also notes the importance and benefits of data sharing to individuals and society (which we interpret to also mean the economies of EU countries).

It also makes clear that 'anonymised data do fall out of the scope of data protection legislation'[3] and 'Once a dataset is truly anonymised and individuals are no longer identifiable, European data protection law no longer applies'.[4] This means that, within the context of data protection legislation, no further data protection obligations would apply to the anonymizing organization or to the anonymized data recipient once the anonymized data are processed further.

Are We Aiming for Zero Risk?

Key Issues

In the analysis in this section we demonstrate that:

- there is a lack of clarity in the Opinion about the concept of the acceptable risk of re-identification, and that the Opinion alludes to multiple approaches that are not consistent with each other, and the Opinion alludes to zero risk as being the acceptable risk,
- zero risk is not consistent with the European Data Protection Directive 95/46/EC, and not consistent with notions of identifiability in other jurisdictions,
- there are significant practical disadvantages to following a zero risk approach, including the potential amplification of privacy risks to EU citizens and the attenuation of other socially and economically beneficial outcomes, and
- there are generally no legal requirements or regulatory expectations of achieving zero re-identification risk in anonymized data.

It is therefore important to be absolutely clear about the concept of acceptable risk, and move away from a narrative around zero risk.

What the Opinion States

In order to 'fall out of the scope of data protection legislation',[5] data must be properly anonymized prior to release. But what qualifies as *proper* anonymization? Practically, anonymization needs to be viewed as a risk management exercise because, as the Opinion notes, that there is 'residual risk of identification inherent in [anonymisation techniques]',[6] 'A risk factor is inherent in anonymisation',[7] and in examples alludes to 'an unacceptable risk of identification'[8] and 'anonymised data sets can still present residual risks to data subjects'.[9] Therefore, there is a spectrum of risk, let's say from Low to High, and along this spectrum there is acceptable risk and unacceptable risk. If data are anonymized to reach a level of acceptable risk, then it will still have residual risk. But this residual risk is low enough that it is considered acceptable.

However, the Opinion also presents an absolute definition of acceptable risk in the form of zero risk. For instance, there are characterizations of anonymization as a way to 'irreversibly prevent identification',[10] requirements that 'identification of the data subject is no longer possible',[11] 'the outcome of anonymisation as a technique applied to personal data should be, in the current state of technology, as permanent as erasure, i.e. making it impossible to process personal data',[12] and there is mention of 'irreversibly preventing the identification of the data subject',[13] stating that 'identification is no longer possible',[14] and 'identification has become reasonably

[2] Ibid 3.
[3] Ibid.
[4] Ibid 5.
[5] Ibid 3.
[6] Ibid.
[7] Ibid 7.
[8] Ibid 10.
[9] Ibid 4.
[10] Ibid 3.
[11] Ibid 6.
[12] Ibid.
[13] Ibid.
[14] Ibid 8.

impossible'.[15] In fact, the concept of 'reasonableness' combined with being 'impossible' is very challenging to even interpret.

Zero Risk is Not Practically Achievable

To be clear, if the acceptable risk threshold is zero for any potential recipient of the data, as alluded to by such statements, with very limited exceptions, then no existing technique can achieve that objective and it will not be possible to anonymize, nor for that matter, share data, without consent or another of the legitimate grounds listed in Article 7 of the European Data Protection Directive 95/46/EC (the 'Directive'). Under that interpretation, it would then not be clear how this Opinion can be consistent with the Big Data world that we are living in and how it can stimulate the private sector to invest and implement privacy protective techniques that help advance best privacy practices for Big Data and the business models built on that.

More specifically, since a zero risk requirement cannot be met, there is little incentive to invest in ways to anonymize data. This would result in one or more of three possible outcomes: (a) when not clearly legally mandated otherwise, personally identifying information will be processed (ie 'minimal necessary' requirements or 'limiting principles' for data use will be interpreted more permissively), (b) since obtaining consent will not be practical many data flows will simply cease with potentially significant negative impacts on commerce and analytics that are societally and economically beneficial, or (c) there will be increasing pressures to change legislation to allow the sharing of personal information by adding more exceptions or by making more uses primary purposes (eg by making research a primary purpose rather than a secondary purpose for collecting, using, and disclosing health data)—which actually presents a much higher privacy risk to EU citizens.

It can be argued that we are reading too much into this imprecise language in the Opinion. However, in practice we see privacy professionals talk about zero risk and refer to these types of reports and opinions as justification for advocating only zero risk. Therefore, the precision of the language, irrespective of the intent, is important and has consequences.

Zero Risk Is Not a Requirement

In fact, the Directive does not anticipate nor stipulate that the risk of re-identification be zero. The Directive uses a 'likely reasonably' standard in its definition of 'personal data' in its Recitals:

> (26) Whereas the principles of protection must apply to any information concerning an identified or identifiable person; whereas, to determine whether a person is identifiable, account should be taken of *all the means likely reasonably to be used either by the controller or by any other person to identify the said person* [emphasis added]

This is also the standard defined in the Spanish personal data protection regulations by including the concept of disproportionate effort in terms of time or activities. Indeed, an identifiable person is described as:

> any person who may be identified, directly or indirectly, through any information regarding his physical, physiological, psychological, economic, cultural or social identity. *A natural person shall not be deemed identifiable if such identification requires disproportionate periods of time or activities*[16] [emphasis added]

As noted in the Opinion, some jurisdictions such as France have not followed the Directive route, since they do not have a reasonableness requirement in their data protection laws with respect to identifiability and, therefore, in such cases the expectation may be that an organization needs to assume zero risk, with the potential consequences noted above.[17]

These divergent approaches in the EU Member States would need to be harmonized if the proposal for a regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data of 25 January 2012 (the 'EU Regulation Proposal') is approved. The European Commission and the European Parliament seem to agree on the 'likely reasonable' standard, even though the amendments proposed by the European Parliament to the EU Regulation Proposal provide additional elements to account for the context. It refers to the costs of and the amount of time required for re-identification, but also takes into consideration both available technologies at the time of the processing and technological development:

> (23) The principles of data protection should apply to any information concerning an identified or identifiable natural person. To determine whether a person is identifiable, account should be taken of all the means *reasonably likely* to be used either by the controller or by any other person to identify or single out the individual directly or indirectly. *To ascertain whether means are reasonably likely to be used to*

[15] Ibid.

[16] Article 5.1(o) of the Spanish Royal Decree 1720/2007, developing the Basic Law 15/1999, on the protection of personal data.

[17] Article 29 Data Protection Working Party (n 1) 6.

> *identify the individual, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration both available technology at the time of the processing and technological development*. The principles of data protection should therefore not apply to *anonymous* data, *which is information that does not relate to an identified or identifiable natural person. This Regulation does therefore not concern the processing of such anonymous data, including for statistical and research purposes*. [emphasis added]

In other non-European jurisdictions, a 'reasonableness' standard is also used to define identifiable information. For example, under the Personal Health Information Protection Act in Ontario, Canada[18] identifying information is defined as information that identifies an individual or for which it is reasonably foreseeable in the circumstances that it could be utilized, either alone or with other information, to identify an individual. Under the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule of 1996 in the US[19] information that is not personally identifiable is interpreted as having acceptably low risk or very small risk of re-identification. Indeed, two methods can be used to satisfy the HIPAA Privacy Rule's de-identification standard: the Expert Determination method and the Safe Harbour method. The Expert Determination method entails the application of statistical and scientific principles in order to determine that 'the risk is very small' that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information. The Safe Harbour method entails the removal of 18 types of identifiers of the individual or of relatives, employers, or household members of the individual and to ensure that the covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.

The US Federal Trade Commission has also set forth a standard of 'reasonable linkability' accompanied by two safeguards: a public commitment not to re-identify the data by (i) the company which carries out the de-identification and (ii) the downstream recipient(s) through contractual commitments imposed by the first one:

> The Commission believes there is sufficient support from commenters representing an array of perspectives – including consumer and privacy advocates as well as of industry representatives – for the framework's application to data that, while not yet linked to a particular consumer, computer, or device, may reasonably become so. There is significant evidence demonstrating that technological advances and the ability to combine disparate pieces of data can lead to identification of a consumer, computer, or device even if the individual pieces of data do not constitute PII. Moreover, not only is it possible to re-identify non-PII data through various means, businesses have strong incentives to actually do so.
>
> In response to the comments, to provide greater certainty for companies that collect and use consumer data, the Commission provides additional clarification on the application of the reasonable linkability standard to describe how companies can take appropriate steps to minimize such linkability. Under the final framework, a company's data would not be reasonably linkable to a particular consumer or device to the extent that the company implements three significant protections for that data. (. . .) *Accordingly, as long as (1) a given data set is not reasonably identifiable, (2) the company publicly commits not to re-identify it, and (3) the company requires any downstream users of the data to keep it in de-identified form, that data will fall outside the scope of the framework.*[20] [emphasis added]

Summary

Care needs to be exercised when discussing acceptable re-identification risk. A more precise way to describe anonymous data is that which 'has a very small risk of re-identification'. In practice, factors such as the reasonable effort to be undertaken and resources to re-identify, and the skills and motivations of the adversary, would be taken into account when deciding on what is acceptable risk, as well as other factors such as the sensitivity of the data and potential harm if there is a successful re-identification attack.[21] Therefore, at least all of the factors noted in the Opinion would be accounted for in a methodology to set acceptable risk levels.

Is It Necessary to Block All Inferences from Data?

Key Issues

In the analysis in this section we demonstrate that:

- the Opinion treats anonymization and inferences from data as a similar set of issues, but they are

[18] H Perun, M Orr, and F Dimitriadis, *Guide to the Ontario Personal Health Information Protection Act* (Irwin Law, 2005).

[19] Department of 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* (Department of Health and Human Services, 2012).

[20] Federal Trade Commission, *Protecting Consumer Privacy in an Era of Rapid Change*, March 2012.

[21] K El Emam and L Arbuckle, *Anonymizing Health Data: Case Studies and Methods to Get You Started* (O'Reilly, Sebastopol 2013); Khaled El Emam, *Guide to the De-Identification of Personal Health Information* (CRC Press (Auerbach), Boca Raton 2013).

orthogonal issues, and conceptually treating them the same makes it challenging to address the privacy risks, and

- algorithmic methods are suggested in the Opinion to manage the risks from inappropriate inferences, but these methods reduce data utility significantly and are rarely, if ever, used in practice.

It is therefore important to make those distinctions clear and develop appropriate mechanisms to manage the privacy risks from inappropriate inferences and decisions from data.

What the Opinion States

The Opinion considers whether information can be inferred concerning an individual (or by extension a group of individuals) as a criterion to evaluate the effectiveness of anonymization techniques. Inferences from data can be discriminatory, stigmatizing, creepy, or surprising (these are terms that are often used in the literature and the media to characterize the risks from inferences).

The Opinion has taken the position that eliminating inferences from the data is necessary in order to avoid such decisions. It states:

> even though data protection laws may no longer apply to [anonymised] data, the use made of datasets anonymised and released for use by third parties may give rise to a loss of privacy. Special caution is required in handling anonymised information especially whenever such information is used (often in combination with other data) for taking decisions that produce effects (albeit indirectly) on individuals.[22]

The Opinion then proceeds to describe two algorithmic techniques that can be used to limit inferences in data: t-closeness and l-diversity.

The Difference Between Attribute vs Identity Disclosure

The Opinion combines two types of disclosure that are usually not combined together. We shall examine these two types below.

One type of disclosure is *identity disclosure*. This is when an adversary is able to assign a correct identity to a record. When discussing identifiability one is referring to that type of disclosure—that of assigning an identity. An anonymized data set is one where the probability of identity disclosure is very small.

The second type of disclosure is called *attribute disclosure*. With this type of disclosure an adversary learns something new about individuals from an analysis of the data. Here we are talking about inferences from the data. Inferences can be simple, such as ‘all 57 year olds in the data set have had heart attacks’. This inference has absolute certainty in that ‘all’ 57 year olds have the attribute of a heart attack. An inference can be more complex where multiple variables are used, for example, to predict the probability of a patient being re-admitted to a hospital or the probability of being diagnosed with a particular type of cancer. In this case, the inference is achieved through a statistical or machine learning *model*. For instance, in the cancer risk model, the variables used may be where the patient lives, gender, race, age, and other diagnoses that a patient has (co-morbidities).

Model-based inference is often not absolute and there is some uncertainty or inaccuracy. For our purposes, we will consider attribute disclosure to have occurred if a model can be constructed with a high certainty or a high accuracy (predictive or descriptive).[23]

In the following we will use the cancer probability inference to illustrate various points, and assume that the model built from the data is highly accurate.

Inferences from Models Are Orthogonal to Anonymization

A model can be built from anonymized data: a data set that has a very small risk of identity disclosure can be used to build that cancer diagnosis model. The same model can also be built from identifiable data. In fact, the risk of identity disclosure is orthogonal to model building.

Once a model is built we can start drawing inferences and learning new things. Inferences are used to make decisions. Decisions can be about groups of individuals or specific individuals. Group decisions can be made without knowing the identity of any cancer patients, for example a health authority may develop cancer screening guidelines for all men when they reach a certain age based on the results of the model. Decisions can also be about individuals – for example individuals of a certain race and age who live in high risk areas may be personally visited by a nurse to discuss lifestyle choices.

The individual-level decisions can be targeted at patients who were not even in the data set. Once that model is constructed it can be used at some future point to predict cancer diagnoses for other patients, even those who are not born yet and could not conceivably be in the data set.

All of these distinctions are important because they affect how we deal with privacy risks.

[22] Article 29 Data Protection Working Party (n 1) 11.

[23] We will not define ‘high’ more precisely here because that definition will not affect the logic of our argument. We will assume that it can be defined in a context-specific manner.

The challenge is that sometimes group or individual decisions are discriminatory, creepy, surprising, or stigmatizing.[24] For example, a property valuation firm may reduce the value of all homes in the high-risk areas because there may be a pollution source causing the higher rates of cancer. In this case, all individuals living in those areas suffer an economic harm because the model was used to make a broad group decision. Alternatively, a bank may impose higher interest rates on loans for specific individual customers of a certain gender, race, and age and living in high-risk areas.

The same model can be used to make socially acceptable and socially beneficial decisions as well as to make stigmatizing or discriminatory decisions. In our example of the cancer diagnosis model, it can be used to develop improved health care services to high-risk communities, or it can be used to discriminate against these communities. The model is not the problem, it is the decisions that are made from the model. The determination of whether a particular decision is appropriate or not will be subjective and contingent on prevailing social norms.

Automated Algorithmic Methods Are not Appropriate for Blocking Inferences

The two algorithmic techniques mentioned in the Opinion, t-closeness and l-diversity, are not used in practice (the authors do not know of a single real-world application). The reason is that techniques which modify the data to limit inferences significantly diminish the analytic utility of the data. There are two reasons for this:

- These techniques assume that all models from the data will be used to make inappropriate decisions, and therefore the data needs to be modified to ensure that no models can be built. For example, the data would be modified so no cancer diagnosis models can be built since it is possible to make inappropriate inferences and decisions from such models.
- Because of the above, many useful models cannot be built from the data and therefore the data become quite useless for analytics purposes.

The determination of whether an inference from a data set or a decision from a model is appropriate or not is a subjective decision. This is why it is not amenable to automation.

Governance Mechanisms to Manage Risks from Inferences

To protect individuals from inappropriate decisions, it is important to manage the risks from the *use* of the models. An appropriate solution to attribute disclosure then is to put in place governance mechanisms that oversee the development and use of the models. Let's call this *privacy ethics*. A group of individuals within a data controller or data processor would advise the business about whether the model and its uses are discriminatory, stigmatizing, creepy, or surprising. Let's call this a *privacy ethics council*. The ethics review process has been in use for a long time in the research community and has worked quite well to ensure ethical data collection, analysis, and decision-making. We are proposing to replicate a lighter version of that type of review more broadly. A privacy ethics council would have a lay person representing the data subjects, a privacy expert, an ethicist, a person representing the business, and a person representing the brand (public relations). This council needs to be independent in order to give un-coerced advice.

An earlier opinion by the Article 29 Working Party provides some good criteria that such a council can consider to determine whether the model and its use would be appropriate.[25] The criteria were introduced in the context of evaluating compatibility (see discussion later in this article), but they are also relevant here:

- The relationship between the purposes for which the data have been collected and the purposes for model-based decision-making.
- The context in which the data have been collected and the reasonable expectations of the data subjects as to their further use.
- The nature of the data and the impact of the model-based decisions on the data subjects.
- The safeguards applied by the controller to ensure fairness in decision-making and to prevent any undue impact on the data subjects.

It should be noted that the application of such criteria is going to be subjective, and it may not always be possible for a data controller to know in advance all the possible models and decisions that can be made with an anonymized data set that is shared. For example, how would an ethics council know a priori if a cancer diagnosis model would be used for discriminatory purposes? It would be a problematic outcome if they erred on the conservative side because then they would likely not share any cancer data due to a possibility of some data processor using the data to make discriminatory decisions. In such cases, conditions of use may accompany the anonymized data to manage those risks.

[24] In many jurisdictions decisions need to be fair, but they can still be creepy.

[25] Article 29 Data Protection Working Party, *Opinion 03/2014 on Purpose Limitation*, 2 April 2013.

The combination of anonymization techniques that address identity disclosure only and governance mechanisms in the form of an ethics council would address the risks from identity and attribute disclosure.

Summary

Inappropriate inferences and decisions from data are a real privacy risk. However, given the subjectivity involved in making that determination, it is more constructive to put in place governance mechanisms to manage such risks. These mechanisms are a form of privacy ethics review and detailed criteria for setting them up and managing them are needed.

Is Anonymization a Compatible Use?

Key Issues

The Opinion, in conjunction with earlier opinions from the Working Party, clarifies that anonymization is a compatible use.[26] This means that a controller is not required to obtain consent to anonymize a data set.

Analysis of Compatibility

In order to anonymize data, it is necessary for an anonymization engine to ingest personal data, apply anonymization techniques to it, and then output anonymized data. The input is personal data. The Opinion notes that anonymization is a form of 'further processing' of that personal data.

With some exceptions (such as but not limited to historical, statistical or scientific use), the Directive states in Article 6(1)(b) that further processing of personal data must not be incompatible with the specified purpose of the data collection. A question that could be posed then is whether anonymization is a compatible processing activity or not.

In its opinion 2/2013, the Article 29 Working Party provided criteria for deciding whether further processing is compatible or not as noted above.[27] One of the criteria for making this decision is to evaluate 'the safeguards applied by the controller to ensure fair processing and to prevent any undue impact on the data subjects'. It did not consider anonymization a compatible or incompatible use as such but only a useful safeguard (among other criteria) that should be taken into account in order to determine whether further processing is compatible.

Article 6(1)(e) of the Directive also refers to anonymization when it notes that information should be kept for no longer than is necessary for the purposes for which the data were collected or for which they are further processed in a form that permits identification. This is the basis of the quality principle and the cancellation right under Article 6(1)(e) of the Directive: personal data must be deleted (or cancelled,[28] depending on how the 'deletion' duty has been implemented in the Member States) when the original legal basis is finished or exhausted. And this deletion or cancellation could be achieved through anonymization. Therefore, in this specific case, anonymization is something different or something more than a compatible use: it is a compulsory processing activity that enables one to comply with the data retention duties. Similarly, as the Opinion notes, the e-Privacy Directive also requires one to keep personal data in an identifiable form for no longer than is necessary, in this case the transmission of a communication, with limited exceptions (eg invoicing purposes or the provision of value added services with consent).

Even when the quality principle is not invoked, anonymization should be deemed 'compatible' by its own nature. Indeed, the Article 29 Working Party has specified, in its earlier Opinion 3/2013, on purpose limitation, that anonymization is to be adopted as a safeguard to ensure fair processing and prevent any undue impact on the data subjects. In this Opinion, anonymization is deemed an instance of further processing of personal data that eliminates or reduces the risk of incompatibility with the original purpose of processing of the anonymized data:

> (...) in this case, prior to its use/disclosure for the secondary purpose, the data is *effectively anonymised*. Therefore, although the two purposes are different, and provided the anonymisation is performed adequately (so the information no longer constitutes personal data or falls into a borderline zone with very low risks of re-identification) this reduces any concerns regarding incompatible processing. (. . .) *if complete anonymisation cannot be ensured* and some risks remain, this should be disclosed - as a rule, and unless an exemption under Article 13 could apply, informed consent will be required. [emphasis added]

This is an important clarification because if there was a requirement to obtain consent from individuals just to anonymize data then that would erect a significant practical barrier to the use and disclosure of data for secondary purposes. Furthermore, there is compelling evidence that consent results in bias,[29] which in certain

[26] Article 29 Data Protection Working Party (n 1) 7.

[27] Article 29 Data Protection Working Party (n 24).

[28] In Spain, for instance, the duty to delete *ex officio* or the right of erasure do not entail the immediate elimination of the data. Data must be cancelled and kept inaccessible during the time there are liabilities that could be claimed. The end of the relevant statutory period entails the definitive deletion.

[29] El Emam (n 21); Khaled El Emam and others, 'A Globally Optimal K-Anonymity Method for the De-Identification of Health Data' (2009) 16 J Am Med Inform Assoc 670–82.

circumstances can affect the outcome of the analysis.[30] Introducing bias into data would not be in the interest of any of the stakeholders.

Legitimate Grounds: a Processing Activity vs. Processing Purposes?

The concept of 'processing' under the Directive is broad and includes, among other activities, the data collection, the data storage, and the data disclosures. There is no doubt, as noted, that anonymization is a processing activity.

In principle, any processing of personal data subject to any EU data protection regulations implementing the Directive relies on one of the legitimate grounds set forth in the Directive.[31] For data controllers,[32] Article 7 of the Directive regarding non-sensitive data (which includes financial data) and Article 8 of the Directive regarding sensitive data (consisting only of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, health, or sex life) exhaustively list these legitimate grounds.

The legitimate grounds are not limited to the data subject's consent, and this is particularly important for anonymization. Furthermore, there are situations in which there is not a likely reasonable opportunity to obtain a free and meaningful consent. It may be impossible to obtain for practical (and economic) reasons if the volume of data is significant, if the contact details are not updated, or the data protection legislation of the EU Member State only accepts a consent that is explicit or challenges its validity on the basis of an 'imbalanced relationship' between the organization and the data subject (as can happen within the employment context[33]). Therefore, attention should not be focused on consent as the unique or primary legitimate ground but rather on (i) the level of risk of re-identification; (ii) the potential adverse impact on the data subject of the purpose of use of the result of the 'anonymization' process; and (iii) the safeguards that must be adopted to remove or mitigate the adverse impact without destroying the value of the data.

Among the legitimate grounds for data controllers, the most relevant for projects involving anonymization would be the so-called legitimate interest ground which is set forth in Article 7(f) of the Directive, which reads as follows:

> Member States shall provide that personal data may be processed only if: (. . .)f) processing is necessary for the purposes of the legitimate interests pursued by the controller or by the third party or parties to whom the data are disclosed, except where such interests are overridden by the interests for fundamental rights and freedoms of the data subject which require protection under Article 1 (1).

Indeed, the Opinion 6/2014 of the Article 29 Working Party includes anonymization as one of the safeguards that would permit one to avoid an undue impact on the data subjects.[34] If this is the case, there will be more chances for the legitimate interest of the controller (or the third party) to prevail in the balance test that is required under Article 7(f).

In any event, in order to determine which ground or grounds could be applicable, the key element is not the processing activity but rather the processing purpose. Many data protection authorities and scholars mix the processing activity itself with its purpose, in particular, when Big Data projects are analysed. However, what must be analysed for the purposes of legitimacy under the Directive are not the collection, storage, disclosure, or the anonymization themselves but the purpose of each of these activities if personal data are (still) involved.

[30] K El Emam and others, 'A Review of Evidence on Consent Bias in Research' (2013) 13 Am J Bioethics 42–44.

[31] As opposed to the Opinion 3/2013 on purpose limitation which states that 'a new legal basis alone cannot legitimize an otherwise incompatible further use', the EU Regulation Proposal (Article 6.4) sets forth that 'where the purpose of further processing is not compatible with the one for which the personal data have been collected, the processing must have a legal basis at least in one of the grounds referred to in points (a) to (e) of paragraph 1'.

[32] The Opinion does not address the specific position of a data processor versus a data controller. In many EU jurisdictions, the data processor shall only process the personal data for rendering the agreed services to the data controllers and must return the data or destroy them (at the data controller's choice) at the end of these services. Could a data processor be ever legitimized to take the decision to anonymize these personal data and use the result for its own purposes without the controller's consent?

[33] This is the case, among others, of France or Germany. This approach has been endorsed in the documents of the Article 29 Working Party. See, among others, the Opinion 8/2001 on the processing of personal data in the employment context:
The Article 29 Working Party takes the view that where consent is required from a worker, and there is a real or potential relevant prejudice that arises from not consenting, the consent is not valid in terms of satisfying either Article 7 or Article 8 as it is not freely given. If it is not possible for the worker to refuse it is not consent. Consent must at all times be freely given. Thus a worker must be able to withdraw consent without prejudice.
An area of difficulty is where the giving of consent is a condition of employment. The worker is in theory able to refuse consent but the consequence may be the loss of a job opportunity. In such circumstances consent is not freely given and is therefore not valid.
The situation is even clearer cut where, as is often the case, all employers impose the same or a similar condition of employment.

[34] *Here it is important to highlight the special role that safeguards may play[67] in reducing the undue impact on the data subjects, and thereby changing the balance of rights and interests to the extent that the data controller's legitimate interests will not be overridden.*
(67) Safeguards may include, among others, strict limitations on how much data are collected, immediate deletion of data after use, technical and organisational measures to ensure functional separation, appropriate use of anonymisation techniques, aggregation of data, and privacy-enhancing technologies but also increased transparency, accountability, and the possibility to opt-out of the processing. See further in Section III.3.4(d) and beyond.

Since the desired result of the anonymization is to obtain de-identified data, it is even more obvious that the relevant purpose is not the anonymization process itself but the purpose of the use of the information obtained once this anonymization process is finished. Therefore, the focus should not be on which 'legitimate grounds of data processing' an anonymization activity should rely on, but rather to ensure that the anonymization activity is properly done, ie that the risk that the data involved could be re-identified by the intended recipients is small and this risk is regularly re-assessed. And this re-identification risk would actually depend on the techniques used, the security and contractual measures adopted, and the intended recipients of the anonymous data.

Assuming that there is a small risk of re-identification according to the above, there is no reason to prohibit or limit the anonymization or to impose the data subject consent as a precondition. What must be analysed in terms of legitimacy are the actual uses of the anonymized information and this analysis would only be relevant (for data protection purposes) when this information is actually used to take decisions regarding data subjects who are identified or could be singled out by the data controller at hand (including but not limited to direct marketing based on patterns/trends resulting from the specific use of anonymized data). Otherwise, the data protection regulations should not apply.

If the intended uses (may) impact or harm the dignity of the human beings in a way that is undesirable or unacceptable in a democratic society, they should not be performed (it will no longer be a question of obtaining a data subject consent). In this context, processing activities that (may) lead to discrimination on the basis of patterns/trends/correlations/anonymous profiling based on race or ethnic origin, political opinions, religion or beliefs, trade union membership, sexual orientation, or gender identity[35] should be examined cautiously, in particular, if the discrimination is used against the individuals.

A practical mechanism to operationalize this decision-making is to set up a privacy ethics council within the organization to oversee these uses, as discussed earlier. In fact, under the legitimacy context such a council would be needed to perform and provide evidence that these factors are being considered and weighed by the organization. This also highlights our earlier point that such considerations cannot be exercised through computational techniques.

Is It Necessary to Always Destroy Original (Identifiable) Data?

Key Issues

The Opinion stipulates that original personally identifiable data must be aggregated or destroyed for a derived data set that has a very small risk of re-identification to be considered anonymized. That the original identifiable data merely exists means that no amount of anonymization would be acceptable.

What the Opinion States

The Opinion states that if there is an original data set with identifiable information, these data are then anonymized to create an anonymized data set, and if the original identifiable data set exists, then the created data set is still not considered anonymized. The mere existence of an original identifiable data by the controller renders any anonymized data set still personal information. The Opinion states:

> when a data controller does not delete the original (identifiable) data at event-level, and the data controller hands over part of this dataset (for example after removal or masking of identifiable data), the resulting dataset is still personal data. Only if the data controller would aggregate the data to a level where the individual events are no longer identifiable, the resulting dataset can be qualified as anonymous. For example: if an organisation collects data on individual travel movements, the individual travel patterns at event level would still qualify as personal data for any party, as long as the data controller (or any other party) still has access to the original raw data, even if direct identifiers have been removed from the set provided to third parties. But if the data controller would delete the raw data, and only provide aggregate statistics to third parties on a high level, [. . .] that would qualify as anonymous data.[36]

A key qualification here is the reference to 'event-level' data. We are interpreting this to mean individual-level data where the transactions or events pertaining to individual data subjects are itemized. This is typically what one would see in an individual-level longitudinal data set. Tabular data that cannot be reduced to an individual-level would not fall under this definition.

It Is Not Practical to Destroy Original Data

The implications of this interpretation are quite severe because some projects and programs will still need the

[35] This is aligned with the amendments of the European Parliament to Article 20 of the EU Regulation Proposal: '3. Profiling that has the effect of discriminating against individuals on the basis of race or ethnic origin, political opinions, religion or beliefs, trade union membership, sexual orientation or gender identity, or that results in measures which have such effect, shall be prohibited. The controller shall implement effective protection against possible discrimination resulting from profiling. Profiling shall not be based solely on the special categories of personal data referred to in Article 9'.

[36] Article 29 Data Protection Working Party (n 1) 9.

original data to conduct their business. For example, consider a hospital that wished to provide anonymized data for research. The hospital needs to retain the original data because that original data are required to treat the patients. To destroy or aggregate the original data would not make any sense. In the context of clinical trials, contemporary transparency initiatives mean that more data for approved drugs and medical devices will be made available to external analysts and researchers.[37] There are data retention regulations on clinical trials source data (for example, five years in the EU)—achieving clinical trials transparency and meeting regulatory obligations would not be possible. There are similar data retention requirements on national statistical agencies for census data. The implications of the above requirement mean that original census data would have to be destroyed or aggregated if they are shared. The same scenario would play out for businesses that need to retain data to serve their customers, but wish to create anonymous data to perform secondary analysis on that data to improve their customer service or to create new products. Requiring them to dispose or aggregate the original data is not practical.

Such a requirement creates strong disincentives to anonymize data. When the original data still have a value for other processing activities of this same organization (and based on a specific legitimate ground) or there is a legal duty to retain personal data for a certain period of time (eg in the context of clinical trials), this organization cannot destroy the original data—that is just not a realistic option. Under the circumstances, if anonymization of the data is considered a negative activity because it requires data destruction then why would organizations even try to anonymize data? To the extent permitted, subsequent processing of data will then be performed on personally identifying information. This increases the risk dramatically for the data subjects because, in practice, the processing of their data will be conducted on their identifiable data. This, arguably, decreases privacy protections in a significant way and significantly amplifies risks for EU citizens. When there is no authority to disclose and process identifiable data, then these societally and economically beneficial analytics will have to stop.

In an earlier opinion the Article 29 Working Party[38] emphasized the importance of 'likely reasonably' in the definition of identifiable information in the Directive. In that case, if it is not 'likely reasonably' that the data recipient would be able to re-identify the anonymized data because they do not have access to the original data, then that anonymized data would not be considered identifiable for this recipient. That would seem to be a more reasonable approach that is also consistent with interpretations in other jurisdictions.

The 'likely reasonably' lack of access by the actual recipient of the anonymized data may be attained in different ways, including through contractual commitments as suggested by the FTC when building its concept of reasonable linkability. Further clarification can be achieved by looking at the Expert Determination de-identification method in the HIPAA Privacy Rule, which only considers 'an anticipated recipient'.[39] It is unlikely that an unanticipated recipient would get access to the original data and therefore there is no stipulation to manage the re-identification risk from both anticipated and unanticipated recipients (eg by destroying the original data).

The European construction of the definition of personal data should therefore not be absolute and take into account the context and the role of each recipient: a recipient having access to both the source data and the anonymous data would still be deemed a data controller as opposed to a recipient having access only to the anonymized data.

[37] Steve Olson and Autumn S Downey, *Sharing Clinical Research Data: Workshop Summary* (National Academies Press, Washington 2013) <http://www.ncbi.nlm.nih.gov/books/NBK131772/>; European Medicines Agency, 'Developing the EMA's Policy on Access to Clinical-Trial Data', CT Data Group 1, *Protecting Patient Confidentiality* (EMA, London 2013); European Medicines Agency, *Release of Data from Clinical Trials* (EMA, London 2013), <http://www.ema.europa.eu/ema/index.jsp?curl=pages/special_topics/general/general_content_000555.jsp&mid=WC0b01ac0580607bfa>.

[38] Article 29 Data Protection Working Party, Opinion 4/2007 on the Concept of Personal Data, 20 June 2007: *'Recital 26 of the Directive pays particular attention to the term 'identifiable' when it reads that 'whereas to determine whether a person is identifiable account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person.' This means that a mere hypothetical possibility to single out the individual is not enough to consider the person as 'identifiable'. If, taking into account 'all the means likely reasonably to be used by the controller or any other person', that possibility does not exist or is negligible, the person should not be considered as 'identifiable', and the information would not be considered as 'personal data'. The criterion of 'all the means likely reasonably to be used either by the controller or by any other person' should in particular take into account all the factors at stake. The cost of conducting identification is one factor, but not the only one. The intended purpose, the way the processing is structured, the advantage expected by the controller, the interests at stake for the individuals, as well as the risk of organisational dysfunctions (e.g. breaches of confidentiality duties) and technical failures should all be taken into account. On the other hand, this test is a dynamic one and should consider the state of the art in technology at the time of the processing and the possibilities for development during the period for which the data will be processed. Identification may not be possible today with all the means likely reasonably to be used today. If the data are intended to be stored for one month, identification may not be anticipated to be possible during the 'lifetime' of the information, and they should not be considered as personal data. However, it they are intended to be kept for 10 years, the controller should consider the possibility of identification that may occur also in the ninth year of their lifetime, and which may make them personal data at that moment. The system should be able to adapt to these developments as they happen, and to incorporate then the appropriate technical and organisational measures in due course".*

[39] 45 CFR 164.514(b)(1)(i).

Summary

While Data Protection Authorities may vary in their interpretations, there is a clear dysfunctionality that will be introduced if all data controllers have destroyed their source data to use it for secondary purposes or if any recipient is considered a data controller or a data processor just because the source data still exist on Earth even if they cannot likely reasonably access it. More sophisticated methods for managing risk need to be considered.

Is It Necessary to Protect Against Any Third Party?

Key Issues

In the disclosure control literature, the entity that attempts to re-identify a data set is referred to as an 'intruder', 'attacker', or 'adversary'. We will use the term 'adversary' here. The adversary is considered to be 'any third party' as stipulated in the Opinion. However, protecting against any third party means that:

- we must make the worst possible assumptions about the context of the data release, and
- the real context where the data have an intended recipient and where there may be many other controls in place would have to be discounted.

The implication is that a re-identification risk assessment ought to ignore the context, which is inconsistent with other statements within the Opinion which require the context to be accounted for.

What the Opinion States

The Opinion, consistent with the Directive, notes that the adversary is the *data controller or any other third party*. One interpretation is that the original organization which has the identifiable data and creates the anonymized data should not be able to re-identify the data. Because of the dysfunction that this would create as noted above, we assume that that is not the intended interpretation. Rather, the intention is the *data recipient* should have a very small risk of re-identifying the data they receive. However, this data recipient can be any other third party.

The Opinion at the same time *does* emphasize the need to take into account contextual elements, for example:

> consideration of all relevant contextual elements – e.g., nature of the original data, control mechanisms in place (including security measures to restrict access to the datasets). . . .[40]

Protecting Against All Possible Third-Party Adversaries

Protecting against each and all possible third parties at any time, who are not necessarily the intended recipients of the data, is problematic and unrealistic. Such a requirement eliminates the need for any risk management because it compels the data controller to always make the worst possible assumptions even if they are not relevant to the specific context.

For example, consider a hospital that is disclosing anonymized data to a pharmaceutical company, Kronk Pharma, and Kronk has set up a very secure and audited environment following best-known practices. The probability of a deliberate, inadvertent, or accidental re-identification is very small because of all of the controls that have been put in place. If the data controller needs to consider all possible third parties as adversaries, then the organization must consider Professor Slocum (a fictitious name for the purpose of the example) as an adversary. Professor Slocum is known to be able to re-identify health data sets and has demonstrated that a number of times. She performs these re-identification attacks and then publishes them. However, the likelihood of Professor Slocum getting that particular data set from Kronk or through any other means is very, very, small. If the hospital has to anonymize the data always assuming Professor Slocum as the adversary then the real data disclosure context is not taken into account. The value of all of the controls that Kronk has put in place is severely diminished when performing this risk assessment.

If we must consider *any third party*, then there is only one context, and it is the context that assumes no controls are being put in place apart from the technical modifications to the data. Other guidelines for anonymization make clear that other controls, not just technical controls, are needed to manage the risk of re-identification.[41]

A possible interpretation of this requirement, although not explicitly stated, is that 'any third party' shall mean one that has the same context as the data recipient and is a 'motivated intruder'. The 'third party' context would be the same as the actual context and a context-specific risk assessment and anonymization can be performed. This would be consistent with current actual practices in the disclosure control community. The 'motivated intruder' concept is developed in the Code of

[40] Article 29 Data Protection Working Party (n 1) 25.

[41] Health System Use Technical Advisory Committee and the Data De-Identification Working Group, *"Best Practice" Guidelines for Managing the Disclosure of De-Identified Health Information* (Canadian Institute for Health Information, 2010); Information Commissioner's Office, *Anonymisation: Managing Data Protection Risk Code of Practice* (Information Commissioner's Office, 2012); Department of 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.*

Practice on anonymization by the ICO,[42] ie the UK data protection supervisory authority (and mentioned in the Opinion 3/2013 on purpose limitation of the Article 29 Working Party), as follows:

> The 'motivated intruder' is taken to be a person who starts without any prior knowledge but who wishes to identify the individual from whose personal data the anonymised data has been derived. This test is meant to assess whether the motivated intruder would be successful. The approach assumes that the 'motivated intruder' is reasonably competent, has access to resources such as the internet, libraries, and all public documents, and would employ investigative techniques such as making enquiries of people who may have additional knowledge of the identity of the data subject or advertising for anyone with information to come forward. The 'motivated intruder' is not assumed to have any specialist knowledge such as computer hacking skills, or to have access to specialist equipment or to resort to criminality such as burglary, to gain access to data that is kept securely. Clearly, some sorts of data will be more attractive to a 'motivated intruder' than others. Obvious sources of attraction to an intruder might include: finding out personal data about someone else, for nefarious personal reasons or financial gain; the possibility of causing mischief by embarrassing others; revealing newsworthy information about public figures; political or activistic(sic.)purposes, eg as part of a campaign against a particular organisation or person; or curiosity, eg a local person's desire to find out who has been involved in an incident shown on a crime map.

The risk from a motivated intruder is considered in currently used anonymization methodologies in the guise of a deliberate re-identification attack.[43]

Summary

To ensure that the context of the data sharing is taken into account, specific re-identification attacks need to be considered: deliberate (motivated intruder), inadvertent, and accidental. This allows the data controller to define more precisely the types of adversaries to consider in their risk assessment, and is narrower than 'any third party'.

Is Incorrect Re-identification a Risk that Must Be Managed?

Key Issues

The Opinion considers the potential for an incorrect re-identification to be a problem with anonymization techniques. However, it is not practical to have anonymization techniques that can guarantee that no incorrect re-identifications is possible, and no existing methods can meet that standard.

What the Opinion States

The Opinion notes that:

> In some cases, a wrong attribution might expose a data subject to significant and even higher level of risk than a correct one.[44]

The implication is that anonymization techniques that cannot protect against this would be somehow considered inferior or unacceptable.

Anonymization Techniques that Protect Against Incorrect Re-identification Are not Practical

Incorrect re-identification is a challenging problem to protect against. An adversary can just take a telephone book and assign random names from the telephone book to records in a data set—in this case the likelihood of an incorrect re-identification would be very high. The adversary can then go and discriminate against these individuals who match—there is nothing stopping an adversary from doing incorrect re-identification and that does not require much skill or effort.

A re-identification attack can have only three outcomes on a data record: re-identify the record correctly, re-identify the record incorrectly, or failure to re-identify a record. If anonymization is appropriate and done properly, then the probability of either an incorrect re-identification or a failure of re-identification is high. For example, let there be a record with an ID of 3 in a data set, and the real data subject is Alan Doe. However, record 3 has values that also matched with Bob Smith. If the adversary then assumed that the data subject was Bob Smith because of the match, this would be an incorrect re-identification. If record 3 does not match any real person then there was a failure of re-identification. In the context of disclosure control both of these outcomes are considered good outcomes because they are protective against identity disclosure. Both outcomes ensure that record 3 is not assigned the identity Alan Doe.

An adversary only gains value from the data if there is correct re-identification. Either of these two negative outcomes create a deterrent from attacking the data because they take away from this attacker value.

If anonymization techniques have to ensure that the probability of incorrect re-identification is zero then the data would need to be distorted considerably—resulting in data sets with limited utility. This would mean that the anonymized data do not match any real person

[42] Information Commissioner's Office, *Anonymisation: Managing Data Protection Risk Code of Practice*, November 2012.

[43] El Emam and Arbuckle (n 21); El Emam (n 21).

[44] Article 29 Data Protection Working Party (n 1) 13.

whatsoever. None of the known anonymization techniques can provide such assurances. Even fake data may match a real person by chance—but this would be prohibited if incorrect re-identification was disallowed.

Protecting against incorrect re-identification is just not practical to do. Penalizing organizations for the potential of incorrect re-identification creates a strong disincentive to anonymize data sets because this sets a standard that cannot be met and that is arguably of limited protective value. Expecting anonymization techniques to not allow incorrect re-identification is not consistent with contemporary disclosure control practices.

Summary

This stipulation of not allowing incorrect re-identification is not practical and is not necessary to protect against identity disclosure.

Does Pseudonymous Data Equal Anonymous Data?

Key Issues

Pseudonymous data are not considered to be anonymous data in the Opinion, and it is still treated as personally identifying information.

What the Opinion States

The Opinion makes clear that pseudonymous data are not considered anonymous data. To clarify this point further, fields in a data set are typically classified as direct identifiers, indirect (or quasi-) identifiers, and 'other'. Direct and indirect identifiers can be used to re-identify the records (link the records to the correct data subject). Pseudonymization is a set of techniques that are often used to protect the direct identifiers that are unique, such as social security numbers or credit card numbers.

Pseudonymous Data Still Has a High Probability of Re-identification

If there are direct identifiers in a data set, then pseudonymization is a powerful approach to protect some of those identifiers. However, by itself it is not sufficient to ensure that the risk of re-identification is very small. The reason is that pseudonymization is not applied to the indirect identifiers. Most known and successful re-identification attacks were conducted on pseudonymous data.[45]

Some organizations have made a distinction between personal information, anonymous information, and a third category of pseudonymous information. That third category is treated as personal information that is less risky but not quite anonymous information. In an earlier Opinion, the working party was moving in that direction when they stated:

> Retraceably pseudonymised data may be considered as information on individuals which are indirectly identifiable. Indeed, using a pseudonym means that it is possible to backtrack to the individual, so that the individual's identity can be discovered, but then only under predefined circumstances. In that case, although data protection rules apply, the risks at stake for the individuals with regard to the processing of such indirectly identifiable information will most often be low, so that the application of these rules will justifiably be more flexible than if information on directly identifiable individuals were processed.[46]

The EU Proposed Regulation allows for this to some extent in Article 4:

> (2a) 'pseudonymous data' means personal data that cannot be attributed to a specific data subject without the use of additional information, as long as such additional information is kept separately and subject to technical and organisational measures to ensure non-attribution;[47]

However, in neither of these views was pseudonymous data considered anonymous, and neither provides any special exceptions to pseudonymous data. The theoretical underpinnings and the empirical evidence are consistent in considering pseudonymous data as having a high risk of re-identification.

Summary

The treatment of pseudonymous data in the opinion is consistent with other interpretations by the Working Party and the disclosure control community.

Linkability Within a Database

Key Issues

The ability to link multiple records that belong to the same data subject within the same database is important to create longitudinal data sets. The Opinion views such a capability negatively.

[45] K El Emam and others, 'A Systematic Review of Re-Identification Attacks on Health Data,' (2011) 6 PLoS ONE, <http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0028071>.

[46] Article 29 Data Protection Working Party, *Opinion 4/2007 on the Concept of Personal Data*, 2007.

[47] European Parliament legislative resolution of 12 March 2014 on the proposal for a directive of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data by competent authorities for the purposes of prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and the free movement of such data: <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P7-TA-2014-0219+0+DOC+XML+V0//EN>.

What the Opinion States

The Opinion uses the *Linkability* criterion to evaluate anonymization techniques. It states

> Linkability, which is the ability to link, at least, two records concerning the same data subject or a group of data subjects (either in the same database or in two different databases).[48]

Linkability is considered undesirable here. However, if linkability of multiple records that belong to the same data subject 'within the same database' is not desirable, then the Opinion is moving in the direction of prohibiting longitudinal data.

Prohibiting Longitudinal Data Sets Is Not Practical

In order to create longitudinal trails of individuals, such as multiple visits to a clinic or multiple treatments during a hospital stay, it is absolutely critical to be able to link the records that belong to the same individual. If anonymization techniques break linkability within the same database, then many large data sets would be useless.

The Opinion also refers to the linkability criterion in the context of pseudonymization, where pseudonymization allows the linking of records that belong to the same data subject within the same database without having to retain original unique identifiers. While the Opinion considers that a disadvantage, it is one of the key benefits of pseudonymization. The most common use-case for pseudonymization is to allow the linking of records that belong to the same individual in the same database.

Summary

There are strong methods for anonymizing longitudinal data.[49] It is not reasonable to require anonymization techniques to eliminate the longitudinal patterns in data sets—that would be disastrous for analytics.

Other Issues

The following are a number of other points made in the Opinion that deserve additional discussion:

- The differential privacy technique is described in the Opinion as a candidate for anonymizing data.[50] It should be noted that differential privacy has some important technical and practical limitations,[51] and it has rarely been used in practice to anonymize data. In fact, the limitations are sufficiently severe that broader application is unlikely until significant additional research is conducted. At this point in time differential privacy is not a practical method for anonymizing real data.
- The Opinion notes that the strength of encryption schemes varies over time. For example, the recommendation not to rely on 'release and forget'[52] implies the expectation that there is a potential for re-identification risks to change over time. However, the Opinion then adds that an 'anonymisation process should not be limited in time'.[53] This is a strong requirement—why would it be acceptable for our encrypted data to be potentially decrypted after some time (see[54]) but not acceptable for there to be a risk of re-identification in the future? In reality we as a society have accepted that encryption has the potential to be broken in the future but we still very carefully and diligently encrypt all of our sensitive data anyway. We accept that there is a risk of new technological developments that defeat current encryption methods. When data are encrypted and transmitted, there is a possibility of the data stream being captured and stored, and decrypted at some future point in time. Anonymization will also be limited in time but we should still anonymize our data using the best techniques available today. This concept of continuous evaluation of risks over time is spelled out clearly in opinion 4/2007 of the Working Party when they state that one should:

  consider the state of the art in technology at the time of the processing and the possibilities for development during the period for which the data will be processed. Identification may not be possible today with all the means likely reasonably to be used today. If the data are intended to be stored for one month, identification may not be anticipated to be possible during the "lifetime" of the information, and they should not be considered as personal data. However, it they are intended to be kept for 10 years, the controller should consider the possibility of identification that may occur also in the ninth year of their lifetime, and which may make them personal data at that moment.[55]
- A claim is made in the Opinion that there has been very limited progress made in anonymization since 2006: 'very little progress has been made since the

[48] Article 29 Data Protection Working Party (n 1) 11.

[49] El Emam and Arbuckle (n 21).

[50] Article 29 Data Protection Working Party (n 1) 15.

[51] Fida Dankar and Khaled El Emam, 'Practicing Differential Privacy in Health Care: A Review', (2013) 5 Trans Data Privacy 35–67; Jane R Bambauer, Krish Muralidhar, and Rathindra Sarathy, 'Fool's Gold: An Illustrated Critique of Differential Privacy' (2013) 16 Vanderbilt J Entertainment Technol Law 55.

[52] Article 29 Data Protection Working Party (n 1) 24.

[53] Article 29 Data Protection Working Party (n 1) 29.

[54] Elaine Barker and others, Recommendation for Key Management – Part 1: General (Revision 3) (NIST, 1012).

[55] Article 29 Data Protection Working Party, *Opinion 4/2007 on the Concept of Personal Data*, 2007.

well-known AOL incident (2006)'.[56] This is a rather broad and bold statement and is not concordant with the reality of the large body of work on disclosure control that has been published over the last 8 years. For example, consider a recent literature review of techniques used to anonymize health data, and this pointed to a rich and diverse computational and statistical disclosure control body of work covering that period,[57] and there are at least two recently formed journals dedicated to disclosure control (*Transactions on Data Privacy*, and the *Journal of Privacy and Confidentiality*).

Conclusions

The Opinion attempted to address a complex topic in the context of multiple and possibly inconsistent interpretations among the member states. The final result has provided clarification and important interpretation guidance on some topics, but did not advance understanding with some other topics, and made stipulations that are not consistent with other Working Party opinions, other regulators, and the disclosure control literature. One of our purposes here was to provide a broader perspective on the issues raised in the Opinion to help interpret them in a manner that can be practically implemented, while still ensuring that the privacy of citizens is protected in a defensible way. Some elements of the Opinion cannot, however, be considered reasonable because they could potentially reduce the privacy of EU citizens by encouraging more sharing of identifiable personal information where permitted, stop beneficial analytics on data, and because they are just not consistent with contemporary practices in the disclosure control community. Other anonymization guidelines that readers should consider are those from the US,[58] the UK,[59] and Canada.[60]

Acknowledgements

We would like to thank Bradley Malin, Ruth Boardman, Francis Aldhouse, Kim Gray, Mark Elliot, and Christopher Millard for providing us feedback on an earlier version of this report.

doi:10.1093/idpl/ipu033
Advance Access Publication 13 December 2014

[56] Article 29 Data Protection Working Party (n 1) 31.

[57] Aris Gkoulalas-Divanis, Grigorios Loukides, and Jimeng Sun, 'Publishing Data from Electronic Health Records While Preserving Privacy: A Survey of Algorithms'. (2014) 50 J Biom Inform 4–19, doi:10.1016/j.jbi.2014.06.002.

[58] Department of 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*.

[59] Information Commissioner's Office, Anonymization: Managing Data Protection Risk Code of Practice.

[60] Health System Use Technical Advisory Committee and the Data De-Identification Working Group, '*Best Practice*' *Guidelines for Managing the Disclosure of De-Identified Health Information*.