Accepted; in Press: Computer Law & Security Review (2018) Functional Anonymisation: Personal Data and the Data Environment

Functional Anonymisation: Personal Data and the Data Environment

MARK ELLIOT,[a1] KIERON O'HARA,[b] CHARLES RAAB,[c] CHRISTINE M. O'KEEFE,[d] ELAINE MACKEY,[a] CHRIS DIBBEN,[c] HEATHER GOWANS,[e] KINGSLEY PURDAM,[a] KAREN MCCULLAGH[f]

[a]University of Manchester
[b]University of Southampton
[c]University of Edinburgh
[d]CSIRO, Canberra
[e]University of Oxford
[f]University of East Anglia

Abstract: Anonymisation of personal data has a long history stemming from the expansion of the types of data products routinely provided by National Statistical Institutes. Variants on anonymisation have received serious criticism reinforced by much-publicised apparent failures. We argue that both the operators of such schemes and their critics have become confused by being overly focused on the properties of the data themselves. We claim that, far from being able to determine whether data are anonymous (and therefore non-personal) by looking at the data alone, any anonymisation technique worthy of the name must take account of not only the data but also their environment.

This paper proposes an alternative formulation called *functional anonymisation* that focuses on the relationship between the data and the environment within which the data exist (their *data environment*). We provide a formulation for describing the relationship between the data and their environment that links the legal notion of personal data with the statistical notion of disclosure control. Anonymisation, properly conceived and effectively conducted, can be a critical part of the toolkit of the privacy-respecting data controller and the wider remit of providing accurate and usable data.

Key words: anonymisation, deidentification, deanonymisation, statistical disclosure control, data environment, ADF, DDF, functional anonymisation, release-and-forget, obscurity

1. Introduction

Superficially, the notion of anonymisation[2] is straightforward: if information contains the identity of a person, then other facts about them can be revealed by the dissemination of that information, and this may breach that person's privacy. For example, sentence (1) discloses information about Jane.

(1) Jane is a 39 year old female, suffering from diabetes, who presented herself for treatment on 24th May.

[1] Corresponding author. CCSR and Social Statistics, Humanities, Bridgeford Street, University of Manchester, Manchester M13 9PL, mark.elliot@manchester.ac.uk.

[2] In some jurisdictions (for example the US, Canada and Australia) the term 'de-identification' is used to mean what anonymisation means in the EU context. In this paper we will use the term anonymisation throughout.

If we can isolate and remove (or replace) that part of the information that contains the person's identity, then that person will not be identifiable from the information, and his or her privacy will no longer be at risk in this context. In our example, we could replace sentence (1) by (2).

(2) A 39 year old female, suffering from diabetes, presented herself for treatment on 24th May.

Clearly (1) is more specific and has more content than (2), but (2) still retains much of the information that is in (1). The decrease in information content in moving from (1) to (2) may be compensated for by the gain in privacy protection. There may also be practical benefits, in that people may be more willing to provide accurate or sensitive information if they trust that their privacy will be protected (Oswald 2014). In many fields of public policy such as health, privacy protection is consonant with the public interest in the use of high-quality sources of data, but it can also be a barrier to research.

However, although the basic idea of anonymisation seems straightforward, the procedure is easier said than done (or, rather, done effectively). For example, sentence (2) might easily reveal the identity of the referent, if one knew a little extra information: for example, that Jane was the only woman of that approximate age who presented herself for treatment that day; someone who knew only that about Jane would thereby learn that Jane had diabetes from (2).

Indeed, it can be formally shown that anonymisation can always, in theory, be reversed, as long as there is some informational content remaining in the data (see for example Dwork 2006). An adversary[3] attempting such a reversal could have access to an unpredictably wide range of information: for example, some of the information that we wish to protect by anonymisation might have been published on social media by Jane herself.

This does seem to lead us a worrying conclusion, that the only way to be certain of countering the threat of re-identification is to turn the information into noise (e.g. turning all of the values in a database to randomly generated ones). Does this mean that anonymisation is doomed to failure, and thus, legally or ethically, that the anonymiser has no justifiable practical basis for anonymisation? Are we condemned never to redeem any of the value of medical data, which is inherently associated with individuals at the micro-level, because – in theory – an opportunity might emerge for an adversary to re-identify the individuals in question? Or is there a trade-off to be made – as argued for example by Cavoukian and El Emam (2011) and Rubinstein and Hartzog (2016) – between the social (or commercial) value of sharing data, and some risk of identifying people, even if that trade-off has consequences for personal privacy?

It is difficult to answer these questions without making the concept of anonymisation more concrete. We argue in this paper that (i) there are various interpretations of 'anonymisation', and also of the related notion of the risk of re-identification; (ii) what is deemed to be an acceptable level of risk will affect understandings of anonymisation and (iii) that anonymisation itself is a complex

[3] We use the term "adversary" throughout the paper to refer to an agent who attempts to re-identify an individual population unit within a de-identified dataset and the term "attack" to refer to the attempted re-identification. Synonymous terms that are found elsewhere in the literature are "intruder" (e.g. Elliot and Dale 1999), and "snooper" (e.g. Duncan et al 2011).

process requiring attention to far more than the data. This line of reasoning leads us to posit that, contrary to a series of influential commentaries, anonymisation, properly conceived and effectively conducted, can be a critical part of the toolkit of the privacy-respecting data controller and the wider remit of providing accurate and usable data.

The question of identifiability, which underlies anonymisation, is prominent in data protection legislation. In the US, 'personally identifiable information' (PII) has a narrower scope, referring to information *maintained by a federal agency* that can be used to trace an individual's identity or that is linkable to an individual (see McCallister et al 2010). The definition is supported with examples of such identifiers: names, addresses (including email addresses), and other known identifiers such as the Social Security number. The European Union's (EU) data protection regime incorporates a category of 'personal data', defined as data from which the subject of the data is identifiable, either on its own or in tandem with auxiliary pieces of data.[4] This creates a legal, as well as an ethical, driver for anonymisation. If the removal or perturbation of information can transform personal data into non-personal data, or PII into non-PII,[5] then the information itself is outside the scope of data protection or privacy regulation, thereby reducing the constraints on the use of that information.

Anonymisation can therefore be seen through the lens of data protection law. If we look at the EU General Data Protection Regulation (GDPR) (2016/679), important and sometimes onerous restrictions are imposed on personal data, defined (article 4) as:

> any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

Note that identification can be direct or indirect. GDPR Recital 26 includes the explanation:

> To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.

A risk-based specification ('means reasonably likely to be used') casts the net wider than the data themselves, because someone wishing to identify the said person is reasonably likely to use whatever resources come to hand, provided that the cost of using them is not too great. Recital 26 specifically asserts:

> The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or

[4] Directive 95/46/EC, Art 2(a); GDPR 2016/679, Art 4(1).

[5] In this paper, except where flagged otherwise, we use the term 'personal data' to mean data from which people are identifiable, and therefore risky in a privacy sense, covering both EU personal data and PII.

> to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable.[6]

In this context, anonymisation has become an important part of the data-sharing toolkit, not merely an ethical means to protect privacy, but a procedure that adds to the business case for data sharing. Phrasing such as 'data rendered anonymous' in the GDPR implies that anonymisation is a procedure that could be applied to data, an algorithm that takes a privacy-breaching dataset as input and produces a dataset from which individuals could not be identified.

EU data protection legislation covers the EU and associated countries, but there are concepts similar to 'personal data' in place elsewhere, such as PII in the USA, and some non-EU countries, for example Singapore, have implemented legislation which draws very heavily on the text of the precursor to GDPR, the EU Data Protection Directive 95/46/EC (DPD).

The remainder of this paper contains four sections and an Appendix. In section 2 we outline the anti-anonymisation position by examining perhaps the most influential anti-anonymisation argument, that of Paul Ohm (2010). Section 3 brings into the discussion a subtler set of techniques, *statistical disclosure control*, that have a long history particularly within the realm of official statistics. In section 4, we define another concept of anonymisation - *functional anonymisation* - which is intended to anonymise data taking into account the context in which they exist, which, following Elliot and Mackey (2013), we refer to as the *data environment*. We then make the case that functional anonymisation is the only robust and effective means of anonymising data. In the concluding discussion, we set out some of the advantages and also the limitations of functional anonymisation, in particular looking at the new and uncertain (Stevens, 2015) legal context provided by the GDPR. The Appendix provides a first sketch of what a formal or semi-formal model based on functional anonymisation might look like. The argument of the paper is intended to be understood independently of the illustration given in the Appendix.

2. The Anti-anonymisation argument: is the promise of anonymisation broken?

The value of anonymisation as a concept and a practical process has been questioned in recent years (Ohm 2010, Rubinstein and Hartzog 2016, Narayanan and Shmatikov 2010). The analysis is often marked by disappointment, exemplified by the title of Ohm's seminal work, 'Broken Promises'. In this section we set out the essence of the anti-anonymisation argument. We first set up the simple anonymisation model, which is the subject of the anti-anonymisation attack. We then describe some examples of failures of this simple model before outlining Ohm's core argument in detail.

2.1 Anonymisation: a simple model

A simple view of anonymisation holds that one merely has to determine an algorithm which, when applied to data, ensures that those data are anonymous with no possibility of re-identification. Under this view, anonymity would be a *state of the data* and one could tell whether data are anonymous simply by inspecting them. If one determines them to be personal in legal terms, then

[6] Stevens (2015) argues that the new definition of 'identifiable' in the GDPR may well include data previously considered to be anonymous.

the dataset is reduced or altered so as to prevent re-identification of any data subjects; the anonymised dataset should not then threaten privacy but still retain enough utility to justify sharing.

One version of this view of anonymisation is that it would be enough to remove the information that might be identifying or disclosive. This would clearly involve removing names and other direct identifiers such as UK National Insurance numbers and US Social Security numbers, and less obviously other information that could lead an adversary to a data subject, such as addresses, places of work, job titles, telephone numbers, car registrations, next of kin, and so on.

A variant of this approach is *pseudonymisation,*[7] by which the same effect is intended while retaining the possibility of linking data about the same person. Replacing a name consistently with a pseudonym (for example, a randomly-generated number) that cannot be traced back to the subject will enable someone to determine when two pieces of data are about the same person, without knowing who that person is.

In more sophisticated variants, certain types of information might be aggregated; for example, in medical data to change someone's exact age to a range may not affect utility very much, while suppressing information that increases the risk of re-identification. Data might also be perturbed in various ways, hopefully ensuring that the dataset's statistical properties (e.g. means and variances) are preserved even though individual values may no longer be assumed to be accurate.

Now, these types of data manipulation do have a role to play, and they will be discussed in more detail in section 3 in the context of a risk based approach called statistical disclosure control. However, for now the main point to note is that under the simple view anonymisation is an algorithm that replaces disclosive data with a potentially less useful but still valuable abstracted version that is deemed to be non-disclosive. This raises two issues. The first is that 'disclosiveness' is not simple to define, and that discoveries may be made that mean that apparently anonymised data has not, in fact, been anonymised. A well-known example is the perhaps surprising discovery that 87% of Americans are uniquely identified by the combination of their ZIP code, birthdate and sex (Sweeney 2000). Such a discovery may affect our assessment of the anonymisation of individual datasets, but does not necessarily impinge upon the sustainability of the simple view. Anonymising effectively may be harder than first thought, and disclosiveness may be a more elusive concept than initially imagined, but that does not mean that anonymisation conceived as a kind of function applied to the data is impossible; perhaps the function has to be more complex than originally imagined in order to be effective.

[7] Recital 26 of the GDPR states: "The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person". This introduced the concept of pseudonymisation into European law. Although pseudonymised data do not directly identify a living individual, they do contain references to individuals' identifiers and may thus indirectly identify an individual if combined with other information. This raises the potential problem of pseudonymised data being used for identification purposes, which makes them in effect personal data. See Stevens (2015) for a detailed discussion of the implications of the GDPR.

The second issue is that, under the simple view, the person who determines the methods and purposes of data processing (in EU law the data controller), and therefore the person accountable for data protection breaches, discharges his or her responsibilities by anonymising. Anonymised data could then be shared beyond the constraints of data protection or privacy law.[8] This appears attractive for three reasons. First of all, it enables some value to be extracted from the data. Secondly, it imposes no costs on the organisation holding the data beyond the anonymisation effort itself. Thirdly, because it takes the data out of the scope of data protection constraints, it removes the need for data deletion[9] and purpose limitation,[10] enabling its use in unanticipated contexts, and thereby implicitly supporting the current 'big (or broad) data' paradigm (Hendler 2013).[11] This leads to a mode of dissemination of anonymised data sometimes called 'release and forget' (Ohm 2010, 1712), meaning that the data controller 'makes no attempt to track what happens to the records after release'. This does however raise further potential disclosure risks.

2.2 Three well-known failures of simple anonymisation

This simple model of anonymisation is somewhat unrealistic. Several high profile, successful re-identification attempts have made headlines in the past decade. In this section, we describe the three examples which have been most damaging to the naïve view.

In 2006, AOL released the search histories of 650,000 users (20 million records in all) over a 3-month period, without redactions.[12] Users were pseudonymised by replacing user names with numbers. The purpose for this was to achieve a public good; the records were made available online (where they remain) for academic and commercial research into aspects of information retrieval. However this public-spirited initiative backfired, as *The New York Times* investigated the data to discover the identities of some of the subjects, and one such person agreed for her identity to be published in a story (Horowitz et al 2010).

One problem, apparently unanticipated by AOL, is that search terms themselves are very identifying. If a person is not famous, they are most likely to be the person searching for their own name; people also search for their home towns and other such close associations. A second problem is that search terms disclose very sensitive information,[13] as people search for medical conditions, sexual content,

[8] In data protection law, this is not technically true, as the data controller will usually retain the original data and therefore still be able to identify persons in the anonymised dataset. This point discussed in more detail in section 3.1.1.

[9] This refers to the idea that once the purpose of data collection has been served, the data (from which individuals can be identified) should be erased.

[10] The idea that the purpose for the use of data must be specified when it is collected, that a data subject's consent (if required by law) is relative to that purpose, and that if the data are used for another purpose, consent must be collected anew from the data subjects.

[11] Broad data are data that come from heterogeneous sources, with varying provenance, reliability and data models, aggregated as a result of Web-scale data search and discovery. The size of the dataset need not be very big, but the complexities are considerable, and data protection issues may loom even larger than with big data.

[12] See Arrington (2006) for more details.

[13] The term 'sensitivity' is used in data protection legislation with a defined itemised meaning. In this paper however, we use the word in a broader sense, intending to convey that the disclosure of the data could easily result in harm to the data subject.

political content, and so on, and that pseudonymised data would allow these sensitive searches to be linked with individuals identified from other searches.

In the same year, the (then) DVD rental company Netflix released data for commercial reasons, as training data for the Netflix Prize competition. The prize, $1m, was to be given to the first researchers who could improve on Netflix's own recommendation algorithm by 10% based on the training data. The dataset covered 500,000 pseudonymised subscribers, with 100m ratings together with dates. To preserve privacy, the data were perturbed. However, within a few days, researchers had uncovered personal data from the dataset about some subscribers, by combining with data from auxiliary sources such as the Internet Movie Database (IMDb), on the assumption that someone rating movies on Netflix might well rate the same movies with similar marks on IMDb. The combination of ratings a person gives to films (especially those that are not blockbusters) is unique or near-unique (Narayanan and Shmatikov 2008). Although Netflix disputed the validity of the analysis, it settled a class action from subscribers, and cancelled the announced sequel to the prize in 2010.

Perhaps the highest profile of these examples was the release of data in 2013 about journey details of New York cabs, following a Freedom of Information request. Here the information was released of all the cab journeys made, including times, fares and start and finish points, pseudonymised by a hash of the cab license and medallion numbers. Unfortunately, the badly designed hash preserved much of the structure of these numbers, and anyone with knowledge of how they were designed could reverse-engineer them. The result was that an adversary could work out, from paparazzi photos of celebrities getting into cabs, where the photos were taken. Just from the data, independently of the identity of the cab, one could also look for the destinations of all cab journeys from significant origins (for example, Larry Flynt's 'adult entertainment' club between midnight and 6am). Such was the granularity of the data that someone looking for the movements of a specific person could search for journey origins very close to their address or place of work (Tockar 2014).

These are real-life examples. However, mathematicians and computer scientists have also been able to show that this situation is endemic (Dwork 2006). If an adversary already has information about an identifiable data subject, and can match that information to information within the anonymised dataset, then links can be inferred between further information in the dataset and the data subject. To take a simple example, if the adversary knows in advance that Jane Doe is 42, lives at a particular address, is female and 5' 6", and if the dataset includes an anonymous individual who is 35-44, lives within the same postcode, is female and 1.68m, has a salary of £35,000, drives a Ford Fiesta, votes for the Liberal Democrats and has £10.43 outstanding library fines, the adversary can infer, with a pretty high probability of being correct, that Ms Doe has a salary of £35,000, drives a Fiesta and so on.[14] This new information may be of direct value to the adversary, but even if not, it adds to his or her database about Ms Doe, and can potentially be used to make even more accurate matches against other datasets.

[14] In the statistical disclosure literature – which we will be discussing later – this is the population-uniqueness problem; see for example Dalenius (1986) or Marsh et al. (1991) for early discussion.

Quite clearly, when someone anonymises data, they cannot rule out the possibility that an adversary has, or will come to have, such auxiliary information. Indeed, anonymisation understood as an algorithm (naïvely) applied to data alone is unable to take the possibility of such information into account.

2.3 Ohm's argument

Ohm argues that 'it is naïve to assume that the adversary will find it difficult to find the particular piece of data needed to unlock anonymized data'. Indeed, he argues that this can 'often' be done 'with astonishing ease', and that 'it is startlingly easy to reidentify people in anonymized data' (Ohm 2010, 1724, 1701, 1730).

Ohm's powerful rhetoric, featuring modifiers like 'often', 'astonishing', 'startlingly', implies that it would be folly to rely on the difficulty of the task or the obscurity of the data standing in the way of success for an adversary, and leads to his recommendation of 'aggressive pessimism', 'given the avalanche of information now available on the Internet, and … the rise of blogs and social networks' (Ohm 2010, 1725). Pessimism is justified because (a) re-identification techniques are simple with 'a fast computer and widely available software like Excel or Access', (b) increasing financial rewards motivate more adversaries, and (c) 'the AOL release reminds us about the power of a small group of bored bloggers' (Ohm 2010, 1730).

Moreover, Ohm and others challenge the assumption that certain data fields are more revealing of identity than others, arguing that that any data field might, under the right circumstances, be revealing. This means that the approach of the US HIPAA Privacy Rule,[15] for example, of enumerating a series of identifiers constituting PII is misguided in two ways. First, it could never be a complete list, because the techniques of re-identification may hinge on any piece (or pieces in combination) of information, including non-sensitive ones such as movie ratings (cf. also Schwartz and Solove 2011, 1831-1835). Second, even if completeness were possible, the list is bound to evolve over time. As Narayanan and Shmatikov put it, laws that enumerate identifiers 'focus solely on the types of data that are commonly used for authenticating an individual, as opposed to those that violate privacy, that is, reveal some sensitive information about an individual. This crucial distinction is often overlooked by designers of privacy protection technologies' (Narayanan and Shmatikov 2010, 24). In contrast, in the DPD (and by extension, the GDPR), if all data fields are equally revealing, we have the opposite problem, that the net is drawn too widely: because the techniques of re-identification encompass the use of *any* relevant information, much more information should be defined as personal data. On these arguments, the US HIPAA Privacy Rule is fatally damaging to privacy, letting too much information through, while the DPD and GDPR are fatally damaging to information flow, and the consequences are drastic. 'At the very least, regulators must reexamine every single privacy law and regulation' (Ohm 2010, 1740, and cf. Schwartz and Solove 2011, 1873-1877).

Ohm argues that we cannot instead treat privacy breaches as a tort, punishing those who harm, because harms would soon overwhelm the system. As the number of data points associated with us increases, then, in effect, any release of information looks like a privacy breach. He points out the problem of accretion (Ohm 2010, 1746), where each new (step towards) identification decreases the protection of an individual that anonymisation can provide, and moves us closer to what he calls 'the

[15] See for example https://www.hhs.gov/hipaa/for-professionals/privacy/index.html [accessed 26/7/2017].

database of ruin', where re-identification and accretion 'join the data from all of the databases in the world together into one, giant, database-in-the-sky' (Ohm 2010, 1748). 'As soon as Narayanan and Shmatikov tied an IMDb username to Netflix rental data, they created an inferential link in the chain, and no regulator can do anything to break that link' (Ohm 2010, 1750). Narayanan and Shmatikov (2010) themselves agree that the notion of a special class of identifying information is 'increasingly meaningless as the amount and variety of publicly available information about individuals grows exponentially' (Narayanan and Shmatikov 2010, 25). All information goes together to create a rich and eventually identifying picture of the individual, and it will therefore be hard, if not impossible, to separate the torts from the harmless revelations.

Ultimately, Ohm asserts, 'data can be useful or perfectly anonymous but not both' (Ohm 2010, 1704). The trade-off between privacy and utility is not a happy one: 'even modest privacy gains require almost complete destruction of the data-mining utility' (Brickell and Shmatikov 2008, 70). On this strongly binary view of the trade-off, technological developments cannot help. We should reject release-and-forget (Ohm 2010, 1755), but this will not save the situation with respect to privacy law. If a data controller remains active in monitoring the data he or she has shared or released (we discuss how this might happen below), then costs will inevitably rise, and even then there are still techniques to game an anonymisation system when the data controller stays involved. 'These alternatives do not make up for the broken promises of release-and-forget anonymization' (Ohm 2010, 1756). Nor will improved regulation help (banning re-identification, for example), because it would be too easy to evade detection and so constraints would not be credible (Ohm 2010, 1758-1759).

As a way forward, Ohm argues for a move 'from math to sociology' (Ohm 2010, 1761), looking at the nature of the threat and the history, practices and traditions of the industry or sector generating the data. He recommends attention to Nissenbaum's (2010) notion of 'contextual integrity', and consideration of the risks, sensitivity of the information and its social utility. Law, for Ohm, should encompass both the US sector-specific approach and the EU's generality, with 'top-up' laws in sectors which are not well-protected by general laws.

3. Beyond the simple model

Much of Ohm's critique is based upon well-documented events that have led to serious repercussions, including those we narrated in section 2.[16] In this paper, we consider not only whether his critique of anonymisation is justified, but also whether his assumptions about the inadequacy of law are correct. In short, can Ohm's criticism be met only by a complex and major restructuring of privacy law (the way forward he proposes he calls 'difficult but necessary' (Ohm 2010, 1776); Schwartz and Solove (2011, 1865) call it 'dramatic')? Or, rather, can anonymisation *practice* be reconfigured to take into account these criticisms?

[16] Having said that, most of these attacks were academic exercises intended to demonstrate the weakness of anonymisation techniques, not attacks by adversaries motivated by malign purposes. Of course, were such a malign attack to be successfully prosecuted, we might not learn of it at all, depending on how the reconstituted personal data were used by the adversary. Attacks that are intended to demonstrate weakness are much more likely to be attended by publicity.

These questions are more than merely academic. Cavoukian and El Emam (2011, 1) respond, for example, by arguing that 'the fear of re-identification is greatly overblown', and that anonymisation 'remains a crucial tool in the protection of privacy'. Their concern is that if anonymisation is considered to be a poor protection, then it will not be used, with the paradoxical result that shared data might be less protected than before.[17] If Ohm's claim that data are either useful or anonymous but not both were true, then it would seem that sharing data to extract their potential value requires a reckless attitude to risk. Indeed, taken at face value, such a claim removes the imperative to reduce the utility of data to make them safer, as to do so will not produce any corresponding gain in privacy protection. Yet this flies in the face of the experience of, for example, National Statistical Institutes, for which anonymisation does provide *some* protection. An analogy: a house with doors and windows is useful, but never perfectly secure, because the means of ingress for the owner also make it possible for a robber to break in. True, but that is not to say that locks, bolts and alarms have *no* role to play in securing the house. It would be a very odd person who said 'I must have a door in my house, and so there is no point in locking or even closing it, as a burglar could break in anyway'. How does this analogy apply to data? Consider Table 1, which might typically be released by a census agency. Now consider an adversary who has complete information about everyone in the table except one person – call her Jane Smith – about whom they only know that she is a female living in Anytown. We might represent the adversary's knowledge as Table 2. Given this knowledge the adversary can subtract Table 2 from Table 1 to arrive at Table 3. We can quickly see that by doing this the adversary learns that Jane Smith is divorced as this is the only logically possible solution.

Now it might be argued that rounding or some other modest manipulation should be used to protect data in Table 1, and we might well agree. But if you are rounding, then you are anonymising, which is precisely the point. The alternative is to say that, because an adversary with the knowledge in Table 2 might exist, aggregated census tables should not be released, but this would seem a perverse and unintended conclusion based on hypothetical premises. There are a number of things that might be said about the meaning of Tables 1 to 3. However, the key point is that such a scenario *could* arise in theory, but it is an extremely unlikely configuration in practice. Critically, what is the plausible scenario whereby an adversary would have all but one piece of information about a population of over 50,000 people, and that information exactly maps onto the equivalent information collected on the census? Obviously, this is just one possible configuration; there are many. However, we are not arguing by this extreme example that we should not consider any threats; merely that we should focus on the empirically plausible rather than the logically possible ones. Aside from being a pragmatic approach this is precisely what is required of us by law: 'To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used' (GDPR Recital 26).

Table 1: Hypothetical cross-tabulation of sex and marital status census data for Anytown.

	Sex		
Marital Status	Male	Female	Total
Single	11105	11100	22205

[17] See also Schwartz and Solove (2011, 1877-1879), who point out that their suggested revision of the concept of PII provides incentives for data controllers to consider subjects' privacy.

	Sex		
Married	12190	12191	24381
Divorced	4633	4128	8761
Widowed	1150	1192	2342
Total	29078	28611	57689

Table 2: Representation of a hypothetical adversary's knowledge of the population in Table 1.

	Sex		
Marital Status	Male	Female	Total
Single	11105	11100	22205
Married	12190	12191	24381
Divorced	4633	4127	8760
Widowed	1150	1192	2342
Total	29078	28610	57688

Table 3: The result of subtracting Table 2 from Table 1.

	Sex		
Marital Status	Male	Female	Total
Single	0	0	0
Married	0	0	0
Divorced	0	1	1
Widowed	0	0	0
Total	0	1	1

Returning to the other side of the debate, Rubinstein and Hartzog (2016) express concern that the debate about anonymisation and re-identification in the wake of AOL, Netflix etc. has not produced any adjustment or improvement in policy, merely a polarised debate. Hartzog (2014) has argued that, in a world where sharing is trivially simple and the practice of social networking has graduated online leaving far more traces than before, the complexity of our preferences about who gets to see what information about us requires *ad hoc* protections that may, in intent, be far more modest than universal measures. However, given the clear willingness of many people to share information, images and records, and to conduct certain aspects of their lives in a public or semi-public way online, and given the need for some data such as census aggregate statistics to be a matter of public record, anything beyond modest, *ad hoc* measures may result in inappropriate and chilling effects. All this suggests that a principled risk-based approach is the right one and so in the remainder of this section we consider the feasibility of such a risk-based view of anonymisation.

3.1 Orthodox approaches to anonymisation

Let us call the data that an organisation wants to share the *original dataset*. If, in order to share data, they perform some anonymisation techniques on the original dataset, then let us call the

resulting dataset the *derived dataset*. Note that we do not call it the 'anonymised dataset', because this would presuppose the success of the anonymisation process.

The question is: how close to the original dataset can the derived dataset be? Given the derived dataset, a motivated adversary may set the reconstitution of (parts of) the original dataset as his or her goal. The original content is the target for the adversary,[18] and the more effective the anonymisation, the further from the original his or her attempts are likely to be.

3.1.1 Formal anonymisation and identifiers

The bare minimum that needs to be done to protect identities from the adversary is to clear the original data of direct identifiers (or replace them with pseudonyms), a procedure which we will call *formal anonymisation*. There are several different types of direct identifier, including at least the following.[19]

- **Unique:** these are specially created for a specific administrative purpose, and associated directly with an individual, such as US Social Security number or UK National Insurance number.
- **Digitised biometrics:** these involve some technological record of a measurement of a human characteristic associated with an individual via a database, such as a genome, an iris scan or gait.
- **Associational:** these are labels or data strings naming objects that have strong and enduring if not permanent associations with individuals, such as mobile phone numbers, static IP addresses or car registrations.
- **Transactional:** these are labels or data strings associated with individuals within the scope of a particular transaction, such as dynamic IP addresses, cookies or an email alias.
- **Social:** more traditional identifiers, such as names and addresses, usually serve social rather than formal purposes of identification, as they are not designed for uniqueness (certain names are extremely common), and addresses generally only become unique when a system is imposed by, for example, a national postal system.

A formal anonymisation procedure clearly helps make the identification of individuals harder, and deals with the first half of the EU definition of personal data in that, following removal of direct identifiers, an individual cannot be re-identified from the data themselves. However, it clearly has no necessary effect on the second half of the definition, which implies that data are personal if someone can be identified indirectly from them, i.e. from the data in combination with other data (GDPR, Article 4). Most obviously, given an original dataset DP containing personal data, then applying an anonymisation function A to it which removes identifiers, results in a derived de-identified dataset A(DP) = DD. Yet assuming that DP still exists, individuals can still be identified in DD using DP as a key, and hence can still be identified indirectly from DD which, on some interpretations of the term, is still personal data. The fact that no-one in their right mind, in

[18] Strictly this is an overspecification. As Elliot and Dale (1999) point out, some re-identification attacks might not be concerned with the actual information content but with the secondary consequences such as the political effect of the public demonstration of the possibility of re-identification, as in the Netflix case. However, we are not here concerned with the specifics of the adversary's goals, and this simplifying description is sufficient to capture the essence of re-identification.

[19] These types overlap, but cover the space.

possession of DP, would go to the trouble of indirectly identifying individuals from DD when they could do it directly from DP is beside the point of this interpretation. In US law, by contrast, surprisingly many profiling practices have evolved to game the current system of PII regulated in sector-specific ways and by privacy policies; for example, some companies amass sufficient quantities of data to identify people, while they take special care not to learn, or store, their names (Schwartz and Solove 2011, 1848-1864).

A data controller relying solely on formal anonymisation would be considered to be extremely irresponsible; so irresponsible that very few conclusions could be drawn about the general efficacy of anonymisation from examples of such irresponsibility, in the same way that we do not base general statements about the utility and safety of cars on the behaviour of the small number of people who occasionally choose to drive them off cliffs. Yet examination of the main exhibits of the anti-anonymisation case – AOL, Netflix, New York cabs – shows that the data controllers were precisely this irresponsible in those cases. In the AOL case, the pseudonyms were numerical codes but the ability to link across the data enabled very revealing information to be deduced. Without pseudonymisation, sensitive searches (e.g. about medical issues) could not be linked back to the revealing searches (for names or home towns). In the Netflix case, more care was taken (data were perturbed, for instance), but even so it was not understood that a movie fan might easily behave consistently across different sites, such as Netflix and IMDb. In the NYC cab case, the hash was easily decodable if the adversary happened to know something about the typical structure of cab registrations, while the failure to anonymise or perturb the location data in the derived dataset provided further routes to identify individuals.

Given these examples of extremely poor practice, the resulting case against anonymisation is not so open-and-shut as it is sometimes presented to be. These egregious failures have rather more rhetorical than logical force. Indeed, as they are very well-known amongst data controllers, they may serve the paradoxically beneficial purpose of warning against complacency.

3.1.2 Irreversible anonymisation

As noted above, anonymisation will always be – in theory – reversible if the derived data have any utility at all, that is, if they possess any informational content. Only if the data are rendered useless – e.g., turning every cell in the derived dataset to zero – can anonymity work, but then of course there would be little point in doing it.

Ohm's claim, 'data can be either useful or perfectly anonymous but never both' (Ohm 2010, 1704), frames the notion of reversibility. Although he does not define perfect anonymisation, he says that '[m]any anonymization techniques would be perfect, if only the adversary knew nothing else about people in the world' (Ohm 2010, 1724), and that '[n]o matter what the data administrator does to anonymize the data, an adversary with the right outside information can use the data's residual utility to reveal other information' (Ohm 2010, 1752).

It follows from this that perfect anonymisation requires that the adversary has no auxiliary data since they could use such data to re-identify individuals within the dataset, and hence it seems reasonable to equate perfect anonymisation with an anonymisation process that prevents re-identification with 100% certainty – i.e., irreversible anonymisation. This squares with other accounts of anonymisation, such as that of the EU's Article 29 Working Party (Working Paper 216,

2014, pp.3 and 5), which says that 'anonymisation results from processing personal data in order to irreversibly prevent identification', and indeed emphasises that '[a]n important factor is that the processing must be irreversible'[20]. Yet we know that anonymisation is reversible (Dwork 2006), and so the technique is set up to fail.[21]

3.1.3 Obscurity

The notion in US law of *practical obscurity* is also relevant here (O'Hara and Shadbolt 2015). Practical obscurity recognises the different privacy status between, say, 1,000 documents each of which contains a single reference to an individual, and a single dossier containing all 1,000 of these references; the former is less threatening to privacy than the latter. In a non-digitised world, this made an enormous difference (both finding a reference in a paper document and finding the paper document itself was far more challenging for an adversary) but even in a digitised world information can be more or less obscure. It cannot be assumed that the *mere existence* of the 1,000 references in digitised form dissolves all distinction between documents and renders document boundaries (and the contexts in which such documents are held) irrelevant.

Schwartz and Solove (2011) argue that the notion of PII can fruitfully be broken down into information that relates to identified or identifiable persons, and describe a continuum of how close to being identified a person is. Hartzog and Stutzman (2013) have attempted to provide a finer-grained framework for classifying types of obscurity, facilitating a more nuanced response. Their analysis is based on four dimensions, which they recognise may not be exhaustive. First, there is *search visibility*: how easy is it to discover the information on a system? Second, there is *the nature of access*: is the information in the open, is its use covered by terms and conditions, or password protected? Third, there is *identifiability*: are individuals identifiable from the information, and if so how easily, and with what supplementary information? And fourth, there is *clarity*: can the information be easily comprehended, is it encrypted, or does it use complex technical terms or organisational jargon that may be hard for an outsider to understand?

Other dimensions may be discerned beyond Hartzog and Stutzman's categorisation. Ambrose (2013) discusses the *life cycle* of information; its importance and sensitivity (and the public interest in accessing it) varies dramatically over time. Although information online will not disappear to order, it does decay at a surprisingly high rate, related to its social relevance. A sixth dimension might be connected to the likelihood of a page being de-indexed by a 'right to be forgotten' request (O'Hara

[20] Another consequence of the Article 29 working party's view is: that the 'anonymised' data set remains personal data even to a third party who cannot access the original data nor has any other means reasonably likely to be used to reidentify the data subjects. This view (which is perhaps related to the proposition that anonymisation must be irreversible) seems to be based on the assertion that data are personal if the subject can be identified indirectly by any data, wherever and by whomsoever they might be held. We thank an anonymous reviewer for this point. As Elliot et al (2016) argue in detial, this view is untenable as it leads to perverse outcomes which make any form of data sharing problematic if not impossible.

[21] Setting anonymisation up to fail creates a space for *differential privacy* (Dwork 2006), which aims to provide a formal but probabilistic privacy framework based on specific (and arguably extreme or unrealistic) assumptions about what information a data user might have access to about the population represented in the data. The extremity of such assumptions means that the data's utility can often be diminished beyond a reasonable limit. When differential privacy techniques are applied to an analysis server, the effect is to rule out meaningful queries to the database (Muralidhar and Sarathy 2010). This is unsurprising – differential privacy aims to restrict access to data that differentiates population units, whereas data analysis (if it is to be worth doing) requires access to such data.

and Shadbolt 2015). A seventh dimension – beyond the issue of simply finding information pertaining to given entities – is the resources required to assemble information that has been gathered from multiple disparate sources into single coherent data entities. Elliot et al (2016) demonstrate that it is possible to gain sufficient information on some entities to make high probability linkages onto de-identified datasets (on the assumption that the adversary knows that the entity is present in the data), but the process is intensively manual; at the moment, data linkage technology is not sufficiently reliable to automate this at scale.

In all these ways, then, online information may be more or less obscure, and this obscurity will affect the ability of an adversary to assemble the dossier of information that will enable him or her to re-identify individuals. Taking obscurity into account has the added advantage that anonymisation is not being set up to fail; reversibility is quite properly acknowledged as a factor in making and evaluating anonymisation decisions. For instance, Elliot et al (2016b) simulated an attack on a pair of social survey datasets, taking into account the ease or otherwise of getting hold of relevant information, to develop a more realistic threat/risk model.

Ironically, this argument from obscurity is supported by an alternative paper by Ohm himself, when he argues that debate about computing issues is inappropriately dominated by a mythical 'superuser', someone who is able to maximise the power of technology, and who understands and exploits all available legal loopholes. He contends that this 'pathological characteristic' of debate has led to 'overbroad prohibitions, harms to civil liberties, wasted law enforcement resources and misallocated economic investment', and the paper calls for policymakers to 'stop using tropes of fear' (Ohm 2008, 1371). Yet his warning, described above, about 'a database of ruin', or 'one, giant, database-in-the-sky' (Ohm 2010, 1748) may itself flirt with the fallacy of the superuser, both in terms of its exaggerated rhetoric and the implicit postulation of a single actor or small group of actors able to use technology to avoid the constraints of obscurity described in this subsection. In order to create the 'database of ruin' using generic software and basic database operations such as inner joins (Ohm 2010, 1725-1727), many complex open research questions in data science would have to be solved, ranging from ontology alignment (Ehrig 2010) to provenance (Moreau 2010).

3.2 Statistical disclosure control and its limitations

Formal anonymisation –whether removal of direct identifiers or pseudonymisation – is, by itself, clearly inadequate to ensure the privacy of data subjects in all, or even most, circumstances. Irreversible anonymisation is a chimera. However, these are not the only tools available, even if we continue to consider anonymisation as an algorithm applied to a dataset containing personal data, to produce a derived dataset. Formal anonymisation is a minimum intervention, while irreversible anonymisation is an impossible ideal. That leaves the possibility of something in between these extremes.

Pragmatically, it is clear that neither formal anonymisation nor absolute (irreversible) anonymisation is a tenable position. Formal anonymisation attempts to stop identification being certain (risk < 100%), while absolute anonymisation makes the impracticable demand that identification risk = 0%. Yet these risk profiles are irrelevant outside of theoretical contexts. The aims of data sharing are ordinarily twofold: (i) making data available that are useful for end-users, while (ii) ensuring that confidentiality is protected, thereby respecting the privacy of data subjects and keeping data controllers compliant with data protection law. Formal anonymisation cannot achieve

(ii) unless it can be shown that the formal identifiers are the only route to disclosing information about an individual. Irreversible anonymisation cannot achieve (i). These types of anonymisation deal with relatively straightforward notions of risk. If we eschew their simplifying assumptions, the calculation of risk becomes an open research question. This brings us to *statistical disclosure control* (SDC, also sometimes called statistical confidentiality), which is both a set of tools for anonymisation and an active research field (see Duncan et al. 2011; Hundepool et al. 2012 for field reviews).

SDC is a deliberately more pragmatic option, congruent with the field of business risk management. On the premise that it is impossible to reduce re-identification risk to zero, it attempts to provide the means of controlling or limiting the risk of disclosure events.[22] The actions and choices of data controllers are embedded in a dynamic, complex and unpredictable world, and so outcomes are inherently uncertain. Given that, the data controller should gather the best information available and use it to optimise decisions to maximise benefits and minimise risk within the bounds of a reasonable cost.

While neither the simple nor absolute notions of anonymisation can bear the weight that is placed on them, SDC can, with some rigour, produce an assessment of the risk of re-identification based on statistical analysis. However, there is still a problem. Most work within SDC has focused exclusively on the statistical properties of the data to be released or shared, as this aspect of the disclosure risk problem is by far the most tractable. Sophisticated statistical models have been developed which anchor identification probability assessments in the data properties. However, these sophisticated models only cover half of the picture; data are not released or shared into a vacuum but into *an environment* and the properties of that environment will also have a significant impact on risk. Measuring the environmental component of risk is undoubtedly challenging, but consideration of that challenge can lead us to a more inclusive and effective understanding of anonymisation, which we now describe.

4. Functional anonymisation – a new approach

It was noted above that the proof that anonymisation is only irreversible at the cost of rendering the data useless relies on the fact that the auxiliary data available in the data environment to an adversary cannot be predicted. Yet it goes unnoticed in much of the literature that this undermines not only irreversible anonymisation, but also the premise of the simple model and SDC. If the failure of anonymisation is down to uncertainty about the auxiliary information, it follows that *one cannot tell from the data alone whether a dataset is anonymous*, for the obvious reason that the data alone say nothing about the auxiliary data.

It is clear that the risk of re-identification is a function of several components, only some of which are statistical properties of the data uncovered by SDC (relevant though these are). As several

[22] It should be noted here that SDC researchers distinguish between *identification* and *attribution* processes in a disclosure. The former indicates that agent X has found person Y in some (supposedly anonymised) data, while the latter indicates that agent X has learnt something new about person Y. These two processes often co-occur but need not. The distinction between these two processes is blurred in data protection law; thus in the *Anonymisation Code of Practice* the UK Information Commissioner says "Note that 'identified' does not necessarily mean 'named'. It can be enough to be able to establish a reliable connection between particular data and a known individual" (ICO, 2012, 21).

authors (Paass 1988; Elliot and Dale 1999; Mackey and Elliot 2009; 2013) have pointed out, other important relevant issues are often ignored, including:

1. The *motivation* of an adversary wishing to attack anonymised data in order to re-identify somebody within it (this will affect what happens and how).
2. The potential *consequences* of disclosure (which will affect the motivations of an individual to attempt a re-identification, and the cost-benefit analysis of the data controller).
3. How a disclosure might happen without malicious intent (the issue of *spontaneous identification*).
4. The *governance structures*, *data security* and other *infrastructural* properties surrounding the release/sharing of the data (this will affect the risk).
5. The *auxiliary data/knowledge* that could be linked to the data in question (without which disclosure or identification is impossible).
6. *Divergence* between the data in question and the other data/knowledge (even if they overlap in content, there may be differences and lack of fit due to alternative semantic encoding, error, quality, differences in measurement, differences in calculation, and so on).

In this section we augment the framework of SDC with these considerations, to create the concept of *functional anonymisation*. Our claim is that, far from being able to determine whether data are anonymous by looking at the data alone, any anonymisation technique worth the name must take account of not only the data but also their environment. This leads to the defining proposition of functional anonymisation:

> **Whether data are anonymous or not (and therefore personal or not) is a function of the relationship between those data and their environment.**

This in turn demands that we provide a well-formed description of what we mean by 'environment'.

4.1 The data environment

The term 'data environment' coined by Elliot et al, was first defined instrumentally as 'the set of all possible data that might be linked to a given dataset' (2010). Yet this is not the only factor, and so the environment needs extension to include the context in which any item of data exists (Mackey and Elliot 2013). This includes the set of (formal or informal) structures, processes, mechanisms and agents that either (i) interact with the derived dataset; (ii) control interactions with those data; or (iii) provide interpretable context for those data. Most critics of anonymisation agree on the importance of context, although they tend to focus on the technology (for instance, 'the line between PII and non-PII is not fixed but rather depends upon changing technological developments'; Schwartz and Solove 2011, 1818, 1885; but cf. 1847 for a more inclusive statement).

Are we able to pin down these structures, processes, mechanisms and agents any further? A data environment usually consists of four key elements, and a description of a data environment that includes these four elements is usually adequate for discussing, planning or evaluating the functional anonymisation of the original dataset. These elements are:

- Other data
- Data users
- Governance processes

- Infrastructure

The first element may seem obvious, but is necessary. The environment in which the derived dataset is placed is likely to contain much more data interrelated in a complex series of ways, and we call this the other data. This is the missing factor in the other types of anonymisation we discussed earlier, which are not sensitive to context and which focus on the properties of the derived dataset alone. Once the derived dataset is disclosed within the environment, then certain linkages with the other data become possible, which will strongly affect the ability of others to re-identify subjects within the derived dataset.

The second element, the data user,[23] motivates and operates on/in the data environment. This is also arguably a necessary component, as without users there would be no reason for data to be shared – and no data sharing, no data environment. Data users capture data, move it, transform it, link it and analyse it, and ultimately they combine those data with other information. Through such operations the data environment is transformed, but data-user behaviour is also shaped by the structure of, and processes available to them within, the data environment.

The third element of the data environment consists of the governance processes over the environment that determine how the users' relationships with the data are managed. These will typically cover a range of behavioural restrictions including formal governance (e.g. data access controls, licensing arrangements, contracts, terms and conditions, policies which prescribe and proscribe user behaviour, and potentially sanctions for breaching agreements) through *de facto* norms and practices to socio-cognitive properties of users (e.g. risk aversion, prior tendency towards disclosure, etc.). It is arguable that the GDPR's consistent use of the phrase 'technical and organisational measures' encapsulates some or all of these governance processes.[24]

The final element is the infrastructure, including physical elements such as security systems, and the software processes that implement functional restrictions on how users can interact with the derived dataset.

We discuss these elements in a little more detail in the Appendix, although the elaboration of the key aspects of the data environment remains an open research question.

4.2 The structure of the data environment

It is a useful explanatory device to assume that the data environment is compartmentalised, that is, it is partitioned by (hard or soft) boundaries that prevent or at least limit the movement of data in and out. These boundaries are constructed out of the infrastructural element of the environment and managed/policed by the governance element. So, in practice, we can identify and act upon the environment in which a dataset resides with relative ease. Of course, the local data environment – that which is contained within the boundaries – is a small, if key, element of a more global context from which it cannot easily be detached. The data environments of different datasets will overlap

[23] By data users we mean any person who might have an interest in the data/information and what it might reveal; we are not using it in its more specific sense as a synonym for data analyst.

[24] For example, GDPR Article 4(5), on pseudonymisation and the use of additional information: "provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person".

and connect in complex ways, which will also impact on the identifiability of the people represented within them.

From another perspective, the global data environment has a structure that comprises local data environments and the relationships between them. There are more or less formal partitions, implemented by laws and regulations, software, security infrastructure, business models, licence agreements, firewalls (within and without organisations) and so on, which frame and even create local environments. So for example, an organisation collects data about its customers and holds them on a restricted access server; the restriction of access creates a partition and thereby a local data environment. The data users are agents who move in, out of and across data environments, storing and processing information, in effect turning themselves into sentient parts of these environments that change other data environments as they enter and leave them, bringing and taking with them data and metadata.

The total data environment is complex, dynamic and fluid, and the local data environments reflect that complexity, preserving it in an almost fractal way. However, in the interests of simplicity and practicality, we have bracketed it off, defining the local data environment in such a way as to emphasise the elements upon which the data controller can place operational controls, namely, the other data, the users, the governance and the infrastructure, at the points closest to the derived dataset. In most circumstances, these controls will impact significantly on the possibility (and cost) of re-identification of a data subject from the dataset. It follows from the earlier discussion that they will also be vital in determining whether the derived dataset contains personal data or not.

4.3 Anonymisation decision-making

What is needed is a notion of anonymisation that goes deeper than other conceptions, by encouraging the re-imagining of the data environment; but also goes beyond that fundamental proposition to implement functional anonymisation in practice. Such a practical position is described in the UK Anonymisation Network's Anonymisation Decision-Making Framework (ADF; Elliot et al 2016a).[25] The ADF suggests a series of issues that a data controller should address while considering how to maintain confidentiality of data subject whilst sharing or releasing the data:

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

[25] The ADF has been adapted for the different cultural and legal environment in Australia, with some differences in detail and a new name, the De-identification Decision-Making Framework (DDF – O'Keefe et al 2017), and see http://data61.csiro.au/en/Our-Work/Safety-and-Security/Privacy-Preservation/De-identification-Decision-Making-Framework.

Clearly not all of these points are straightforward; many are extremely complex. The Framework is heuristic in nature and specifically does not claim to be an anonymisation algorithm. The point we emphasise here is that functional anonymisation is not conceived as an algorithm applied to an original dataset to produce a derived dataset (although it may involve the application of such an algorithm as part of its toolkit). In the first place, the original dataset is understood as operating in a data environment. A derived dataset is likely to be created (this is the essence of component 7 in the above list), but care is also taken to engineer an environment in which the derived dataset is low-risk. Note also that functional anonymisation clearly eschews the release-and-forget ethos (components 8-10); we discuss this point further below. The aim of functional anonymisation, as described in the ADF, is for the data controller to reduce the risk of a data breach to an acceptably low level.[26]

To reiterate, anonymity cannot be 'read off' from the data, and so we need to consider risk factors pertaining not only to properties of the data (as in SDC), but also from the data environment. Such risks include the potential motivation for attempting to re-identify people from the data, (and whether, for example, this goal is achievable by other, less resource-intensive means); the consequences of re-identification; the governance of the data (who gets to see it, under what conditions, and are these enforceable?); the provenance of the data (how have they got here, and through whose hands?); relevant auxiliary data (e.g. related time series data, open data, commercially-available data in the same domain); and the quality of the data (ironically, poor quality data may be safer from a privacy point of view). These various risks can only be modelled, although clearly not all motivations can be anticipated. But this is not sufficient to prevent a data controller building a useful threat model nor is it a justification for not doing so.

Given these factors, there are many operations that can be applied to the original dataset to build a derived dataset: for example, aggregation, removing fields or variables, sampling, suppressing unique values, perturbation, or microaggregation, as well as medium-specific ideas such a pixellating faces or detecting and disguising verbal tics. But it is also important to shape, or at least to understand, the environment in which the derived dataset will sit. An open data environment will be extremely permissive, and leaves no residual element of control. This demands a very secure derived database, and it has been argued that anonymisation and publication are inappropriate in a totally open environment (Rubinstein and Hartzog 2016). We return to that contention below.

The data controller is not a passive occupant of the data environment. Environments can be engineered to ensure greater control. Access controls can be used to restrict access to trusted analysts, to provide constraints or to ensure the possibility of enforceable sanctions. Query control can be used to restrict the questions asked of the dataset, and therefore the information it will

[26] 'Acceptable to whom?' one might ask; good question. Key stakeholders here include the data subjects, who have a personal privacy interest; the data-holding organisation itself, which has a reputational interest and an interest in reducing its liabilities, and presumably will already have an appetite for risk that could be high or low; and the regulators, who need to decide, in the event of a data breach, whether the data controller was negligent. The acceptable level of risk will depend on the outcome of debate between these various stakeholders, but ultimately it will be the data-holding organisation that will make the final decision as the data controller, albeit possibly subject to accountability requirements and legal challenge. It is also worth noting that any engagement activity will be partial and imperfect but will still undoubtedly be better practice than not engaging at all.

reveal.[27] Data might only be accessed in secure environments, or with particular software or hardware, and analysis might be restricted, for example via project approval processes, or publication agreements.[28]

Under functional anonymisation, data controllers are understood to have a number of clear responsibilities. They need to understand how and why a privacy breach might occur, understand the consequences and understand the environment. Importantly, they need to address the risk of a breach occurring: a risk-based approach must always countenance the possibility of an unfavourable outcome, by definition. There should be a plan in place for when the worst happens, and the plan should include informing all relevant stakeholders, including data subjects from the original dataset and those who might suffer reputational damage or other harms as a result of the breach.[29]

5. Discussion

It has been argued (Ohm 2010, Narayanan and Shmatikov 2010) that the mere creation of a derived dataset is not enough to allow a data controller to assert that that dataset is 'anonymised' such that subjects cannot be re-identified within it. It is clear, from the various formal and empirical investigations into anonymisation, that this argument is correct. It does *not*, however, follow that a derived dataset is thereby unsafe. Our argument here is that, once we consider the importance of the data environment, and given that it is possible to describe the data environment conscientiously in such a way that threats and protections within it can be made clear, the notion of risk (or for that matter safety) is not applicable to a decontextualised derived dataset at all. Anonymisation should not be understood as an *algorithm* that is applied to a dataset to achieve a derived dataset with a measurable level of risk, but an ongoing *process* intended to keep the risk of re-identification from the data down at an acceptable level.[30]

5.1 Aligning the response to the threat

A number of things follow from this, a few of which we have space to discuss here. The first point is that there is a presupposition in the argument that data are valuable, that the value often increases when they are integrated with other datasets, and that external agents may often do this more effectively. Hence data sharing can, at least in some cases, be socially, scientifically or commercially valuable.

Hence functional anonymisation has two functions to perform. It is intended to render data safe (whereby the risk of re-identification is negligible), but also to preserve the utility of data. The nature of tensions between these two will vary. Indeed, as many have argued, sometimes utility is increased along with privacy protection; for example, if the privacy-preservation quality of the data increases confidence in it, then it may lead to greater willingness of data subjects to provide

[27] Differential privacy is a type of query control in the broad sense that it is a standard for a data process. Rather than an alternative to anonymisation, we see it in its proper guise as a tool in the functional anonymisation toolbox.
[28] For a description of the range of potential controls, see O'Keefe and Rubin (2015).
[29] The GDPR regulates data breaches in Recitals 85-88 and in Articles 33, 34, 82 and 83,
[30] We do not here put a specific number on what is acceptable as this will be contextualised. Elliot et al use the concept of negligibility – to denote a risk that reasonable person would ignore. So on my journey home I ignore the risk of being hit by a meteor (it's negligible) but I don't ignore the risk of being hit by a car (it's not).

accurate data in the first place. But if the data are anonymised into a state of minimal usefulness, this is just as powerful a failure as if re-identifications were trivially possible from the data.

This leads to the obvious recommendation of aligning the remedy to the magnitude of the threat. Sensitive data may need a lot of work to protect privacy, and may not be releasable even in derived form, except under very restrictive conditions. Data for which a threat cannot easily be discerned, or which provide information that would be easier to extract from other sources, may not justify very onerous functional anonymisation procedures. Schwartz and Solove (2011, 1879-1883) argue that individuals might have different rights within fair information practices, depending on the obscurity of their identity within the data.

This relates in turn to an often-unconsidered aspect of the re-identification situation, which can best be expressed as the question 'can the adversary achieve their goals (more easily) by other means?' By focusing on the data rather than the entirety of the situation, proponents of an absolutist approach fail to take account of this functional constraint on risk. If my goal is to find some information about X and I can achieve that with lower costs and/or higher probability of success by taking action A than action B, then it is reasonable to assume that I will take action A (unless I am unaware of it as an option). So, if action B is 'carry out a re-identification attack on dataset D' then we can assume that that attack will not take place if there is a more effective and/or efficient means for me to achieve my goal. Using the terminology of Marsh et al. (1991), the probability of an attempt being made is lower. We explore this in more detail in the Appendix.

5.2 Forget release-and-forget

A strong argument stemming from Ohm (2010) is that release-and-forget is absolutely untenable as a philosophy for sharing data derived from personal data, and it should be clear that in this we concur. Our reasoning, however, is different. Ohm's point is based on the theoretical limits for simple/naïve anonymisation that have been proven mathematically, and the inevitability of eventual exposure. 'In the arms race between release-and-forget anonymisation and re-identification, the re-identifiers hold the permanent upper hand' (p. 1752).

Our alternative argument to the same conclusion is that the data environment (agents, auxiliary data, governance and infrastructure) is not a static entity.[31] Auxiliary data will change and grow. Infrastructure and governance will change as computer security evolves. The capacities of agents will also change over time, as new tools and methods become available. In particular, adversaries will learn new tricks, and will try new attacks. Since anonymisation is, as we have argued, a context-relative process, so data that is anonymous now (i.e. the risk has been managed down to an acceptable level) will not necessarily be anonymous in the future. Hence, release-and-forget, which implicitly assumes a static data environment, is indefensible.

Functional anonymisation is not therefore a cheap option that removes the responsibilities of data controllers at the moment of sharing. It sets a requirement for ongoing data stewardship, and implies future resource commitments. The extent of this commitment will vary, and again should be aligned with the threat and the risk. It is evident that many positive suggestions for moving beyond

[31] In principle, it could evolve to become more privacy-friendly or less. In practice, we would expect the environment to evolve in a less friendly direction, all things being equal, because more data will become available, and better data mining technologies are likely to appear.

release-and-forget put forward elsewhere, for example access controls (Ohm 2010, Narayanan and Shmatikov 2010, Schwartz and Solove 2011, Elliot et al 2016a) are entirely congruent with the ideas put forward in this paper.

The Framework does give ground for some optimism that action can be taken, on the assumption that any kind of data, even open data, can be understood as obscure to some degree. Rubinstein and Hartzog (2016) argue that open data should never be truly 'open' if they are derived from personal data but equally in the event of a re-identification from anonymised open data, it would still be possible to lower the risks, e.g., by taking the data down or introducing access controls. If stakeholders were tolerant of that situation, and if it could be established that the offending data had not proliferated online, the formerly open data may then have an increased degree of obscurity that will re-establish some protection (cf. Hartzog and Stutzman 2013). Rubinstein and Hartzog make the point that making data fully accessible without controls is 'not what matters most' (2016, p. 722), and is therefore taking a needless risk.

5.3 Functional anonymisation as data protection by design

Recital 78 of the GDPR states that 'In order to be able to demonstrate compliance with this Regulation, the controller should adopt internal policies and implement measures which meet in particular the principles of data protection by design and data protection by default'. This is explicitly specified to include pseudonymisation. The promotion of data protection by design is a clear aim of the GDPR (e.g., in Recitals 78 and 108, and Article 47). However, the meaning, of 'data protection by design' is perhaps less clear (though see GDPR Article 25) than the chronologically prior notion of privacy by design (PbD), which is characterised by seven principles (Cavoukian 2011), and which has been outlined in much more detail. On Cavoukian's account, PbD should be:

1. Proactive, not reactive, preventative, not remedial.
2. Privacy as the default setting.
3. Privacy embedded into design.
4. Full functionality. Positive sum, not zero sum.
5. End-to-end security. Full lifecycle protection.
6. Visibility and transparency.
7. Respect for user privacy. Keep it user-centric.

Space precludes a full discussion, but functional anonymisation meets these seven requirements. Given the ambition of the GDPR, further engagement with the detail of anonymisation might have been valuable (SDC, for example, is absent from its text), to co-opt functional anonymisation as a means of embedding PbD into data controllers' practice.

A risk-based approach allows a dynamic definition of personal data, with data being personal in some data environments and not others and that status being variable over time depending on the evolution of the environment. This would allow a return to the relatively clear motive for anonymisation, to take data out of the scope of data protection regulation, which would have two advantages for the data controller.[32] First, it would link the notion of personal data with a process

[32] Of course, there are some implications for data subjects that are not all necessarily positive. If data are no longer personal, they are no longer governed by GDPR or other data protection laws. Those laws include

that could directly manage it. Second, it would provide a clear rationale and incentive for data controllers to invest resources in using functional anonymisation. It would produce these two advantages while at the same time ruling out the release-and-forget approach to anonymisation.

6. Conclusion

In this article we have addressed the question of the extent to which the nature of the environment within which data are held (their *data environment*) can affect the possibility of data subjects being re-identified in datasets, and therefore the question of whether, and when, personal data can be rendered non-personal. In our discussion of anonymisation we have described how data that are *irreversibly anonymised* are unlikely to be useful and data that have only been *formally anonymised* are likely still to be personal. Influential arguments by Ohm (2010) and Narayanan and Shmatikov (2010), attack a simple notion of anonymisation, and underestimate the resources available for data controllers to lower the re-identification risk. The same effect as Ohm's suggested radical change in both US and EU law could in fact be produced by methodological changes to the practice of anonymisation. These could play a similar role in law to ideas like 'reasonable behaviour' in negligence law (Schwartz and Solove 2011, 1884).

We have developed the concept of *functional anonymisation* that ties together notions of disclosure risk with those of the data environment. Following Mackey and Elliot's (2013) initial work, we have proposed that a data environment can be understood by a small number of parameters: other data in the environment, the skills, knowledge and motivations of the persons who are present in the environment, governance structures and processes, and the infrastructure in which the data resides; and we have argued that those parameters are to a large extent controllable by the data controller. Control of these parameters can result in major changes to the identifiability or otherwise of data subjects in any dataset. As argued above, anonymisation is not an algorithm that is applied to a dataset to achieve a measurable level of safety, but an ongoing *process* intended to keep the risk of re-identification from the data down to an acceptable level. As Narayanan and Shmatikov state, 'any system for privacy-preserving computation on sensitive data must be accompanied by strong access control mechanisms and non-technological protection methods such as informed consent and contracts specifying acceptable uses of data' (Narayanan and Shmatikov 2010, 26). However, we would emphasise that a holistic methodology, which specifies anonymisation as a combination of technological and non-technological methods applied in an integrated fashion, is likely to be more effective than an approach that casts anonymisation as exclusively technological, with non-technological processes bolted on as an afterthought.

In essence then, functional anonymisation is the practice of reducing the risk of re-identification through controls on the data and its environment so that it is at an acceptably low level. This leaves open the question of how small a risk would be deemed to be sufficiently low for us to regard a

information rights (e.g., subject access, remedies, complaints procedures, role of supervisory authorities in enforcing compliance, etc.). The data subject can still be harmed (or subject to other information risks) by data processing and usage even if the law no longer considers the data to be 'personal', and may have fewer and weaker legal protections. Also, if data controllers are deemed not to be using personal data, they might escape legal requirements for accountability and transparency for their data stewardship because data protection laws no longer apply to what they are doing. One might therefore have to rely on ethical precepts to which users might or might not adhere. However, this is beyond the remit of this article.

dataset as non-personal. This is ultimately a policy decision that is outside the scope of this article. However, we note that these sorts of decision are made all the time in human societies (e.g., what is the expected number of plane crashes that is deemed to be sufficiently low so that we can accept the cost?). The reason for these policy decisions is that we want the social goods that are the upside of the decision (available plane travel). Recall that the aim of anonymisation is not solely to produce safe data (and it cannot produce risk-free data); it needs also to produce useful data. One candidate for understanding these that has gained some traction in a UK context is the principle of negligibility.[33] A negligible risk is one *that a reasonable person[34] would ignore*. For example, returning to the example of household security, when Prudence left for work this morning she locked up her house because the risk of her house being burgled was not one that she chose to ignore, and so she took action to reduce that risk down to a level that she believed to be negligible. On the other hand, she took no action on her way to work to manage the risk of being struck by a meteor. She did that because the risk of that event is already negligible, even though the costs, were it to happen, would be extreme.

The potential benefits of the use of data about individuals for our society and those individuals themselves and others, are huge. Yet the potential costs in privacy loss are also considerable, and of concern, both for the individual and for society, given that privacy has a social-good and public-interest value alongside its value as an individual right. This is under-appreciated in conventional policy discourse but should nonetheless be considered seriously (Regan 1995; Raab 2012). Functional anonymisation is a practical framework for delivering the desired benefits without throwing away the concept of information privacy altogether.

Acknowledgements

This work was partially supported by a grant from the Simons Foundation, and by the EPSRC project SOCIAM, grant no. EP/J017728/2.

Several of the author(s) thank the Isaac Newton Institute for Mathematical Sciences, University of Cambridge, for support and hospitality during the programme Data Linkage and Anonymisation where work on this paper was undertaken. This work was also (partially) supported by EPSRC grant no EP/K032208/1.

References

Meg Leta Ambrose (2013). 'It's about time: privacy, information life cycles, and the right to be forgotten,' *Stanford Technology Law Review,* 16(2) 369-422.

[33] The phrase *de minimis non curat lex* – (translation from Latin: the law does not concern itself with trifling matters) expresses the view that in law, some risks are so small that we have no reason to take action against them even if such action could be taken at no or negligible cost.

[34] The phrase 'reasonable man' is used to denote a hypothetical person in society (formerly rendered in the UK, in somewhat more gendered terms, as the man on the Clapham Omnibus*: Hall v Brooklands* [1933] 1 KB 205) who exercises average care, skill and judgment in conduct and who serves as a comparative standard for determining liability.

Michael Arrington (2006). AOL proudly releases massive amounts of user search data, TechCrunch, available at: https://techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/ [accessed 04/08/2017].

Article 29 Working Party, Opinion 05/2014 on Anonymisation Techniques, available at: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf [accessed 04/08/2017].

Justin Brickell and Vitaly Shmatikov (2008). 'The cost of privacy: destruction of data-mining utility in anonymized data publishing', in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 70-78.

Ann Cavoukian (2009, rev. 2011) Privacy By Design: The 7 Foundational Principles, revised version, Office of the Information and Privacy Commissioner of Ontario, Canada, available at: https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf [accessed 04/08/2017].

Ann Cavoukian and Khaled El Emam (2011) Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy, available at: http://www.ontla.on.ca/library/repository/mon/25006/310614.pdf [accessed 04/08/2017].

Tore Dalenius (1986) Finding a Needle In a Haystack. *Journal of Official Statistics* 2(3), 329-336.

George T. Duncan, Mark Elliot and Juan-José Salazar-González (2011). *Statistical Confidentiality: Principles and Practice*, (New York: Springer).

Cynthia Dwork (2006). 'Differential privacy', in Proceedings of the 33rd International Colloquium on Automata, Languages and Programming - Volume Part II, 1-12 (Springer-Verlag Berlin, Heidelberg).

Department for Digital, Culture, Media and Sport, (2017). A New Data Protection Bill: Our Planned Reforms, Statement of Intent, (07.08.17) available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/635900/2017-08-07_DP_Bill_-_Statement_of_Intent.pdf [accessed 07/08/2017].

Chris Dibben, Mark Elliot, Heather Gowans and Darren Lightfoot, (2015). Chapter 3: The data linkage environment, in Katie Harron, Chris Dibben and Harvey Goldstein (eds.), *Methodological Developments in Data Linkage,* (London: Wiley) 36-62.

EU(1995) Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, *Official Journal L 281*.

Marc Ehrig (2007). *Ontology Alignment: Bridging the Semantic Gap,* (New York: Springer).

Mark Elliot and Angela Dale (1999). 'Scenarios of attack: the data adversary's perspective on statistical disclosure risk,' *Netherlands Official Statistics,* 14 (Spring), 6-10.

Mark Elliot, Susan Lomax, Elaine Mackey and Kingsley Purdam. (2010). 'Data environment analysis and the key variable mapping system' in Josep Domingo-Ferrer and Emmanouil Magkos (eds.), *Privacy in Statistical Databases,* (Berlin Heidelberg: Springer), 138-147.

Mark Elliot, Elaine Mackey, Kieron O'Hara and Caroline Tudor (2016a). *The Anonymisation Decision-Making Framework*, Manchester: UKAN.

Mark Elliot, Elaine Mackey, Susan O'Shea, Caroline Tudor and Keith Spicer (2016b). 'End User Licence to Open Government Data? A simulated penetration attack on two social survey datasets', *Journal of Official Statistics,* 32(2), 329-348.

Woodrow Hartzog (2014). 'The value of modest privacy protections in a hyper social world', *Colorado Technology Law Journal,* 12(2), 333-351.

Woodrow Hartzog and Frederic Stutzman (2013). 'The case for online obscurity', *California Law Review,* 101(1), 1-50.

James Hendler (2013).'Broad data: exploring the emerging web of data', *Big Data* 1(1), 18-20.

Adam Horowitz, David Jacobson, Tom McNichol and Owen Thomas (2007) '101 Dumbest Moments in Business: The year's biggest boors, buffoons and blunderers', *CNN Money,* available at: http://money.cnn.com/magazines/business2/101dumbest/2007/full_list/index.html [accessed 04/08/2017].

Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer and Peter-Paul De Wolf (2012). *Statistical Disclosure Control,* (Chichester, UK: John Wiley & Sons).

ICO, (2012). Anonymisation: code of practice available at: https://ico.org.uk/media/1061/anonymisation-code.pdf [Accessed 04/11/2017].

Information and Privacy Commissioner of Ontario *Privacy By Design: The 7 Foundational Principles*, revised version, Toronto: Information and Privacy Commissioner of Ontario, available at: https://www.ipc.on.ca/images/resources/7foundationalprinciples.pdf. [Accessed 04/11/2017].

Elaine Mackey and Mark Elliot (2009) 'An application of game theory to understanding statistical disclosure events', in *Proceedings of Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality* (Bilbao, Spain, 2-4 December 2009), available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2009/wp.40.e.pdf [accessed 04/08/2017].

Elaine Mackey and Mark Elliot (2013).'Understanding the data environment', *XRDS*, 20(1), 36-39.

Catherine Marsh, Chris Skinner, Sara Arber, Bruce Penhale, Stan Openshaw, John Hobcraft, Denise Lievesley and Nigel Walford (1991). 'The case for samples of anonymized records from the 1991 census', *Journal of the Royal Statistical Society series A*, 154, 305-340.

Erika McCallister, Tim Grance and Karen Scarfone (2010) Guide to Protecting the Confidentiality of Personally Identifiable Information (PII): Recommendations of the National Institute of Standards

and Technology, NIST Special Publication 800-122, Gaithersburg MD: National Institute of Standards and Technology, available at: http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-122.pdf [accessed 04/08/2017].

Luc Moreau (2010). 'The foundations for provenance on the Web', *Foundations and Trends in Web Science*, 2(2-3), 99-241.

Krish Muralidhar and Rathindra Sarathy (2010) "Does differential privacy protect Terry Gross' Privacy?" in Josep Domingo-Ferrer and Emmanouil Magkos (eds.), *Privacy in Statistical Databases,* (Berlin Heidelberg: Springer), 200-209.

Arvind Narayanan and Vitaly Shmatikov (2008). 'Robust de-anonymization of large sparse datasets', in Proceedings of the 2008 *IEEE Symposium on Security and Privacy,* 111.

Arvind Narayanan and Vitaly Shmatikov (2010). 'Myths and fallacies of "personally identifiable information,"' *Communications of the ACM*, 53(6), 24-26.

Helen Nissenbaum (2010). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*, (Stanford, CA: Stanford University Press).

Kieron O'Hara and Nigel Shadbolt (2015). 'The right to be forgotten: its potential role in a coherent privacy regime,' *European Data Protection Law Review*, 1(3), 178-189.

Kieron O'Hara, Nigel Shadbolt and Wendy Hall (2015) A Pragmatic Approach to the Right to be Forgotten, Global Commission on Internet Governance, paper 26, available at: https://www.cigionline.org/sites/default/files/gcig_no26_web_1.pdf [accessed 04/08/2017]..

Christine M. O'Keefe and Donald B. Rubin (2015). 'Individual Privacy versus Public Good: Protecting Confidentiality' *Health Research, Statistics in Medicine 34*, 3081-3103. DOI: 10.1002/sim.6543

Christine M. O'Keefe, Stephanie Otorepec, Mark Elliot, Elaine Mackey and Kieron O'Hara (2017). *The De-Identification Decision-Making Framework*, CSIRO Reports EP173122 and EP175702, http://data61.csiro.au/en/Our-Work/Safety-and-Security/Privacy-Preservation/De-identification-Decision-Making-Framework.

Paul Ohm (2008). 'The myth of the superuser', *University of California Davis Law Review*, 41, 1327.

Paul Ohm (2010). 'Broken promises of privacy: responding to the surprising failure of anonymization', *UCLA Law Review*, 57, 1701-1777.

Marion Oswald (2014). 'Share and share alike? An examination of trust, anonymisation and data sharing with particular reference to an exploratory research project investigating attitudes to sharing personal data with the public sector', *SCRIPTed*, 11(3), DOI: 10.2966/scrip.110314.245.

Gerhard Paass (1988). 'Disclosure risk and disclosure avoidance for microdata', *Journal of Business and Economic Statistics*, 6(4): 487-500.

Charles D. Raab (2012). 'Privacy, Social Values and the Public Interest', pp. 129-151 in Andreas Busch and Jeanette Hofmann (eds.) 'Politik und die Regulierung von Information' ['Politics and the

Regulation of Information'], *Politische Vierteljahresschrift* Sonderheft 46, (Baden-Baden: Nomos Verlagsgesellschaft).

Priscilla Regan (1995). *Legislating Privacy: Technology, Social Values, and Public Policy*, (Chapel Hill, NC: The University of North Carolina Press).

EU (2006). 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), *Official Journal of the European Union*, L 119/1

Ira S. Rubinstein and Woodrow Hartzog (2016). 'Anonymization and risk', *Washington Law Review*. 91(2), 703-760.

Paul M. Schwartz and Daniel J. Solove (2011). 'The PII problem: privacy and a new concept of Personally Identifiable Information', *New York University Quarterly Review of Law*, 86, 1814-1894.

Leslie Stevens, (2015). 'The Proposed Data Protection Regulation and its Potential Impact on Social Sciences Research in the UK' *European Data Protection Law Review*, 1(2), 97-112

Latanya Sweeney (2000). Uniqueness of Simple Demographics in the U.S. Population, *Laboratory for International Data Privacy Working Paper*, LIDAP-WP4.

Anthony Tockar (2014). 'Riding with the stars: passenger privacy in the NYC taxi dataset', *Neustar Research blog*, 15th Sept, 2014, available at: https://research.neustar.biz/author/atockar/ [accessed 04/08/2017].

Appendix: formalisms for describing the data environment

Formally describing the data environment is non-trivial, although it is possible to sketch means for doing this, to give a sense of what formalisms data controllers might be able to reason over in a fully developed system. That is the aim of this appendix, which is not compulsory for the reader who is not interested in this kind of formal detail. In this appendix, we will consider three types of environment which are commonly created for sharing derived datasets: (i) open data; (ii) a community of users created by licensing; and (iii) controlled safe settings. We begin with open data as it is the simplest to parameterise and we will then discuss the variation arising from the other two types of environment. The fundamental point is that however low a risk threshold one deems sufficiently low to make data functionally anonymous, the judgment about whether data meet that threshold will depend on the environment as much the data themselves.

Open data

When data are made open they are in effect released into the global data environment with minimal restriction. Any agent can in principle access the data and process them. Because in most jurisdictions open data are necessarily presumed to be non-personal data (indeed, publishing personal data in the open without consent or enabling legislation would not be legal across the EU), there are no legal restrictions on their processing.

Given the open environment, the auxiliary data in the environment are necessarily broad-ranging. An agent U_i has simultaneous access to four forms of auxiliary information (I) which could be used to re-identify population units within an open derived dataset D.

1. Data held within partitions[35] to which U_i has access (I_P).
2. Information available to U_i as personal knowledge – either their own or that of a third party (I_K).
3. Other open data (I_O).
4. Readily available data (I_R).

These are likely to have different properties, and may require different methods of control.

As there are no governance constraints on U_i, the set $\{I_P,I_K,I_O,I_R\}$ forms the *theoretical total data environment* (TE_i) for U_i. The attributes A_i of U_i (skills, metadata knowledge, resources and partition constraints on U_i) act as the single functional constraint on that theoretical environment expressed formally in equation 1 (where FE_i is the *functionally-constrained data environment*):

$$(1) \quad FE_i = A_i(TE_i)$$

(1) should be interpreted to mean that A_i acts as a moderator for TE_i such that the information content of FE_i is always less than TE_i. For the purposes of re-identification we consider the data environment with respect to both U_i and D, as shown in equation 2, where ∩' signifies that the intersection of D and $A_i(TE_i)$ is probabilistic due to potential data divergence between the data and their environment. In other words, the data in dataset D might overlap with the data in TE_i, but there may be inconsistencies (perhaps due to error, differences in measurement, or different methods of calculation) between the two. Hence the intersection cannot be specified with 100 percent probability.

$$(2) \quad FE_{iD} = A_i(TE_i) \cap' D$$

Equation 2 relates directly to the probability of an identification given an attempt to do so, as set out by Marsh *et al*. (1991) in equation 3:

$$(3) \quad pr(identification) = pr(attempt).pr(identification \mid attempt)$$

The probability that an attempt will be successful given that it has taken place is captured by the attributes of the user in combination with the empirical intersection between the data that are being attacked and those data's environment. Substituting into equation 3 leads us to equation 4:

$$(4) \quad pr(identification)_{iD} = pr(attempt).A_i(TE_i) \cap' D$$

Equation 4 makes no assumption about the costs and benefits to the data user about re-identifying, other than that the re-identification is worthwhile. But if we wish to assess the risk as accurately as possible by gathering as much relevant information as we can, we also need to take account of the

[35] Partitions are soft or hard divisions of the data environment. So a limited access server is a partition as is a set of rules governing use of a dataset or service.

agent's level of motivation to attempt to re-identify somebody. This will give us a more complex estimate for the probability of an attempt being made that we can substitute into equation 4.

Discussion and evaluation of the many different models of motivation are beyond the scope of this article, so in lieu of that important exercise we will assume simply that there is some indicator function *M*, which for a given user will indicate whether they will attempt to re-identify somebody in the data. That indicator function will itself be dependent on the derived dataset (*D*) the user's functional attributes (A_i), and some utility (*U*) that the user expects to obtain from the re-identification. Furthermore, because the likelihood of the user's attempting a re-identification will be affected by their perception of the likelihood of success, *M* will also depend on $A_i(TE_i) \cap' D$. Hence we get equation 5:

$$(5) \quad Pr(identification)_{iD} = M_i(U_i, Ai, D, A_i(TE_i) \cap' D).A_i(TE_i) \cap' D$$

Finally, we can generalise this across all agents. The function *E(I)* is the expected number of identifications and *N* is number of agents with access to the open environment, as shown in equation 6:

$$(6) \quad E(\mathrm{I}) = \sum_{i=1}^{N} M_i(Ui, A_i, D, A_i(TE_i) \cap' D).A_i(TE_i) \cap' D$$

We have not here specified the indicator function *M* and given the aforesaid complexity.[36] However, it is reasonable to assume that *M* is roughly proportional to its parameters, so that a functional increase in any of those would lead to an increase in M. For instance, if the user gains more re-identification skills, or if the utility for the user of re-identifications increase, then it is reasonable to assume the he or she will be more motivated to attempt re-identification. In those limiting cases where attempting re-identification is trivially easy, we can assume that *M*=1.

To sum up, the series of equations (1)-(6) shows us that it is possible to express formally what one might expect the risk of attempts at re-identification to be, what parameters would be relevant to the calculation, and how varying those parameters would cause the risk to change. If we make the simplifying assumption that the data environment is open, so that access and disclosure controls are not in place, we can develop an equation for the expected number of identifications based on the number of agents, their motivations, skills, resources and the auxiliary data that they have access to.

It will be evident that this is not going to be an exact science soon. Nevertheless, a data controller could reason about the data environment, and use formalisms of this kind to understand and communicate the risk with a little more confidence. Such an equation as (6) will not produce a clear number – 'the risk is 0.43' – even if we could understand what that meant in real-world terms, but the data controller would be able to reason about the effects on the expected number of re-identifications of a change in one or more of these parameters. For instance, if the data controller could be confident that the original dataset would have utility for only a small number of agents, then she might argue that the risk calculation need only include those potential users, and could effectively overlook the long tail of uninterested agents.

[36] Any attempt to do so would be pre-theoretical in that there would not be a parsimonious set of principles for specifying this. One obvious extension to this would consider *how hard* a given user will try to carry out the re-identification, but this is an unhelpful elaboration for current purposes.

In the remaining two sections, we will take these ideas and consider how they might work in more restrictive environments, showing how changes to the parameters might affect data controllers' reasoning.

Licensing: creating a community of trusted users

Licensing is a means of soft environmental partitioning, introducing some friction into the flow of information. Minimally, licensees undertake not to pass on the data to an unlicensed third party. Often there are additional sets of prescriptions (typically data security expectations) and proscriptions (typically an undertaking not to attempt re-identification of a data unit).

This moderates the 'Wild West' of the open data environment. Firstly, the number of agents is smaller. If the licensing regime was 100% functional then the number of agents would equal the number of licensed users. On more realistic assumptions, allowing for some 'leakage', the number of agents is still likely to be considerably reduced from the global open data environment. Against this it is reasonable to suppose the vast majority of those agents in the open data situation will have low values for *U* and *A*, and that most or all of this barely-relevant tail will be excluded by the licensing regime. Nevertheless, we can take *N* in the licensing situation as much smaller than *N* in the open data situation.

A second difference is in the governance provided by the licence. If we assume that there will be some consequence for the user for visibly breaking the licence conditions, then this will impact on U_i with the effect of lowering the value of M_i.

Even for those who have obtained a copy of the data outside of the licence, we can reasonably assume that they have 'behaved badly' and perhaps even illegally (e.g. hacking to obtain a copy of the data, which in the UK would be in contravention of the Computer Misuse Act 1990).[37] This is a different situation to that of open data where the licence is permissive and specifically does not prohibit re-identification. This element of bad faith alters the values of the parameters in equation 6. For many agents, it is likely to reduce *U*, although not in all cases. There will be cases where the licensing regime effectively reduces the number of people with access to the information, which will increase the competitive advantage of holding it, which may increase *U*. However that may be, even if we assume the mean effect on *U* of a licensing regime is zero, the skills and resources required for re-identification will be greater. Partition constraints may also increase – it may be that the auxiliary information required to achieve re-identification may itself only be available from a small number of sources. Identifying the re-identifier may have legal or reputational consequences.

In short, the environmental elements of the right hand side of equation 6 will have reduced value, reducing utility to agents, and therefore almost certainly bringing down the expected number of identifications. The implication of this is that for a given level of *E(I)* it should be possible, all things being equal, to release data such that $(TE_i) \cap' D$ is larger with licensed data than with open data,

[37] In a statement of Intent on a Data Protection Bill to give effect to derogations under the GDPR, the UK government has announced plans to create a new offence of intentionally or recklessly re-identifying individuals from anonymised or pseudonymised data. And, offenders who knowingly handle or process such data will also be guilty of an offence. The maximum penalty would be an unlimited fine (DCMS, 2017, 10).

which implies that *D* is either more detailed and contentful or less perturbed. Either way, the utility of *D* for licensed users should be enhanced.

Secure safe settings

Secure safe settings are a broad class of hard environmental partitions that use both soft and hard infrastructure and governance to control access. $A_i(TE_i)$ is directly limited, sometimes by physical infrastructure as well as rules. For instance, the data user may be required to work in a particular physical environment, in which access to the Internet is forbidden, monitored or limited, or the analytical output that can be removed from the secure environment is strictly controlled. This severely restricts the ranges of I_P, I_O, *and* I_R. Examples from the UK include the Secure Data Service,[38] and the Data Lab of Her Majesty's Revenue and Customs.[39]

In effect an agent that wished to carry out a re-identification would be relying on I_K. It is possible to imagine a user memorising information for a particular individual and then hunting for individuals with those attributes in the dataset but wide-scale cross-match attacks as described by Elliot and Dale (1999) are ruled out. Indeed some safe settings are so secure that the user cannot see the data, ruling out even those attacks.

The licence conditions associated with secure safe settings tend to be heavier and the consequences of carrying out a re-identification more severe. The probability of being able to do so undetected is also much smaller. So compared with simple licensing, safe settings have lower levels of both A_i (and therefore $A_i(TE_i)$) and U_i, and consequently lower levels of *E(I),* given a fixed derived dataset. How much lower will depend on the details of the governance and infrastructure of the safe setting. Virtual safe settings (such as the Australian Bureau of Statistics RADL[40] or the UK Data Service's Secure Lab[41]) will have higher levels of *E(I)* compared to on-site labs; settings where the user is able to view the data directly will have higher levels than those where this is not allowed (for example the English Census Longitudinal Study[42]).

[38] https://www.ukdataservice.ac.uk/use-data/secure-lab.
[39] https://www.gov.uk/government/organisations/hm-revenue-customs/about/research#the-hmrc-datalab.
[40] http://www.abs.gov.au/websitedbs/D3310114.nsf/home/CURF:+Remote+Access+Data+Laboratory+(RADL).
[41] https://www.ukdataservice.ac.uk/use-data/secure-lab.
[42] https://www.ons.gov.uk/aboutus/whatwedo/paidservices/longitudinalstudyls.