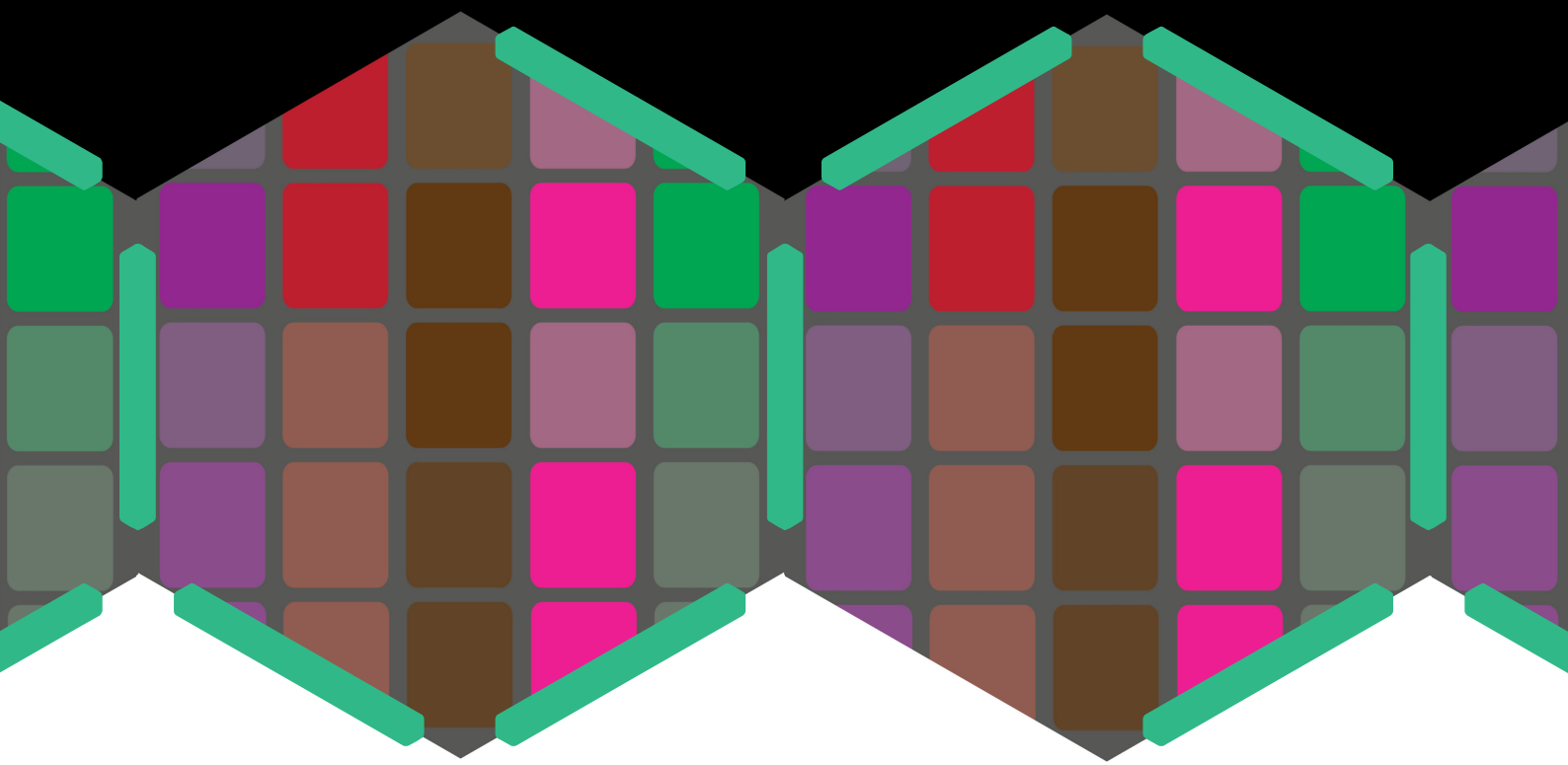
The De-Identification Decision-Making Framework

Appendices

Christine M O'Keefe, Stephanie Otorepec,

Mark Elliot, Elaine Mackey and Kieron O'Hara

18 September 2017



Australian Government

Office of the Australian Information Commissioner

Citation

CM O'Keefe, S Otorepec, M Elliot, E Mackey, and K O'Hara (2017) The De-Identification Decision-Making Framework. CSIRO Reports EP173122 and EP175702.

This work is an adaptation to the Australian context of the publication: M Elliot, E Mackey, K O'Hara, and C Tudor. The Anonymisation Decision-Making Framework. UKAN Publications, UK, 2016. http://ukanon.net/ukan-resources/ukan-decision-making-framework/

Author affiliations

Christine M O'Keefe, CSIRO, Australia
Stephanie Otorepec, Office of the Australian information Commissioner, Australia
Mark Elliot, University of Manchester, UK
Elaine Mackey, University of Manchester, UK
Kieron O'Hara, University of Southampton, UK

Licensing and copyright

Important disclaimer

Contents

Figures

Tables

Part I More detail on topics covered in the main text

Appendix A Terminology

A.1 De-identification, anonymisation and confidentialisation

The reader is probably aware that internationally, a range of terms is used to describe a series of related concepts and processes, including: 'de-identification', 'anonymisation', and 'confidentialisation'. Each of these terms is heard in Australia too, for example, confidentialisation is currently used by the Australian Bureau of Statistics (ABS) to refer to the statistical and non-statistical processes used to meets its legislative requirement to ensure that information "… shall not be published or disseminated in a manner that is likely to enable the identification of a particular person or organization", see the *Census and Statistics Act 1905* and the *Statistics Determination 1983*.

De-identified is the term which is used in the Australian Privacy Act, to refer to information which is no longer about an identified or reasonably identifiable individual. Anonymisation and confidentialisation, by contrast, do not appear in any relevant Australian legislation.

Both internationally and in Australia, the term de-identification is sometimes used differently – in particular, to refer exclusively to the removal of direct identifiers. This more restricted definition is not used in this book. As most readers will already be aware (and which this book explains in detail), the removal of direct identifiers alone would not generally result in information being no longer about an individual who is reasonably identifiable.

Further, the terms anonymisation, confidentialisation and de-identification are used more or less interchangeably in the Australian context. To avoid any misunderstandings, it is important to be aware of the differences and nuances in use of terminology in this field, and to check what definitions are applicable in any given document, conversation or scenario. In particular, there is a high degree of overlap between the use of the term of de-identification in this book (for example, where it refers to statistical processes such as data modification) and use of the term confidentialisation by the Australian Bureau of Statistics.

A.2 Usages of the term 'de-identification'

The term 'de-identification' has been used with a variety of different meanings in different contexts and at different times. Two of these usages have already been introduced in Appendix A.1. For the purposes of the discussion in this book, it is useful to distinguish four different usages, which we call 'types of de-identification' and give them the names: formal de-identification, absolute de-identification, statistical de-identification, and functional de-identification, adapted from Elliot et al (2015).

A.2.1 Formal de-identification

Data is formally de-identified if all direct identifiers have been removed from the dataset or masked in some way. Direct identifiers come in five forms:

1. **Intentional Unique Identifiers**: These are numbers that have been created with the explicit intention of identifying a person and for linking transactions. They are often used in multiple contexts and usually are associated with a person across their lifespan. Examples in Australia include: Tax File Number and Individual Healthcare Identifier.
2. **Digitised Unique Biometrics**: These are codifications of unique, or statistically very likely to be unique, physical characteristics of individuals, to be used intentionally as identifiers. Their use can be intrusive, and – because they are hard to disavow – are often used in security contexts. Examples include fingerprints, iris scans, gait recognition systems, DNA and handwritten signatures.
3. **Associational Unique Identifiers**: These occur where some object which itself has a unique identifier is (strongly) associated with a person. Examples are a telephone number (particularly a mobile phone number), credit card number, static IP address, and driver's license or car registration number. They are invariably non-permanent but can exist for a while. General Unique Identifiers or GUIDs, which are used by Windows OS to identify software components and indeed users, and which in some cases can be semi-permanent, also fall into this category.
4. **Transactional Unique Identifiers**: These are numbers which have been generated as part of some transactional process. They are not necessarily permanent. Examples are sessional cookies and dynamic IP addresses.
5. **Functional Unique Identifiers**: This category is a borderline one. Technically, they are a form of indirect identifier. The most straightforward example is full name and address. They will almost always be constructed out of more than one piece of information. They will also usually include the possibility of data twins (it might be that there are two people called 'John Henry Smith' living at address X), but these will be rare enough that we can treat them as if they are unique.

A.2.2 Absolute de-identification

For de-identification to be absolute (also called absolute or irreversible) there must be zero risk of an individual being identified within a dataset given whatever assumptions one wishes to underpin the guarantee. This is the meaning of de-identification that is often used within the security engineering literature (Ohm 2010).

Ohm (2010) asserts that one can have de-identified data or useful data but not both, and if one regards de-identification as an absolute process then he is correct. It may not be immediately obvious that this is true. So you might think for example that heavily aggregated data are 'absolutely de-identified'. However, a theoretical intruder who has almost complete knowledge of the population from which the aggregated data were drawn but who lacks one piece of information about one particular individual could utilise what they already know to discover the piece of information that they are lacking (this is called a subtraction attack which we will discuss further in Appendix B.4.3). You might argue that this is a contrived situation and we would entirely agree. The point here is not to suggest this approach is sensible – it is not – but rather to illustrate how Ohm's assertion is a logically necessary consequence of the notion that risk can be entirely removed from the process.

However, we would argue that de-identification should not be considered from this absolute standpoint.

Let us consider Ohm's position further using a simple analogy: we can have secure houses or usable houses but not both. If we assume that by secure we mean absolutely secure, then this is true. An absolutely secure house would lack doors and windows and therefore be unusable.[1] But that does not mean that all actions to make one's house more secure are pointless, and nor does it mean that proportional efforts to secure my house are not a good idea. The deadbolt on my door may not help if a burglar comes armed with a battering ram or simply smashes my living room window but that does not mean that my lock is useless, merely that it does not (and cannot) provide absolute security.

The above considerations also apply to personal information. The Privacy Act does not require de-identification to remove risk entirely, but rather demands that those sharing or disseminating data mitigate the risk of re-identification until it is very low. However, when choosing which de-identification solutions to apply in respect of a dataset so that the data can be shared lawfully, one should ensure that the data will still be of sufficient utility once those techniques have been applied.

So in general, absolute de-identification is not practical if one wants to share useful data – there will always be some risk associated with that activity. Risk management naturally suggests a statistical treatment and this brings us to the third type of de-identification.

A.2.3 Statistical de-identification

The notion of statistical de-identification is tied into a technical field called statistical disclosure control (SDC) which we discuss in more detail below. The basic tenet of SDC is that it is impossible to reduce the probability of re-identification to zero, and so instead one needs to control or limit the risk of disclosure events to a low level.[2] This brings the notion of de-identification into line with other areas of business risk management. One accepts that our actions and choices, responsibilities and constraints are embedded in a complex world which is impossible to predict in detail so one gathers the best information one can and optimises one's decisions to maximise the expected benefits and minimise the risks.

One could argue that both formal and absolute de-identification are simply special cases of statistical de-identification. Formal de-identification is a mechanism for reducing the probability of re-identification below the value 1 (certainty) and absolute de-identification is a mechanism for reducing it to zero. However, someone who releases or shares data that relates to individuals should have two goals:

- to release/share data in a form which prevents re-identification (and thereby protects privacy), and
- for those data to be useful.

[1] Artist Rachel Whiteread's concrete cast of the complete interior of a house makes this point quite nicely: https://en.wikipedia.org/wiki/House_%28sculpture%29 [accessed 30/5/2016]. Of course, as a work of art, this had no need for utility!

[2] It should be noted here that disclosure control researchers distinguish between types of disclosure, including *identification*, *attribute disclosure,* and *inferential disclosure*. See Appendix B.1.

It should be clear from the foregoing that formal de-identification will fail to achieve goal 1, while absolute de-identification will fail to achieve goal 2. Statistical de-identification recognises that there is a lot of ground in between these two extremes.

At this point it is worth introducing a de-identification standard called k-anonymity. In some ways this standard is an attempt to take the best features of the absolute and statistical approaches and combine them in a single standard which also combines risk assessment and control. Essentially, a dataset is regarded as k-anonymised if – on all sets of key variables – each combination of possible values of those variables has at least k records that have that combination of values. The parameter k is defined by the entity carrying out the de-identification, and common choices are 3 and 5.[3]

A.2.4 Functional de-identification

Unfortunately, assessing disclosure risk even with the simplest of data is far from trivial. Indeed, a whole research community has built up around the topic with its own journals and conferences. Much of the work in this field has focused on the statistical properties of the data to be released/shared, primarily because there is the perception that this aspect of the disclosure risk problem is by far the easiest to address. A great deal of headway has been made; sophisticated statistical models have been developed which have at least facilitated re-identification probability assessments anchored in the properties of the data.

However, as several authors (e.g. Paass 1988; Elliot and Dale 1999; Mackey 2009; Mackey and Elliot 2013, and Ritchie 2014) have pointed out, despite the advances in statistical disclosure control theory we are at best basing our measurement on only some of the determinants of the risk. There is a range of other issues to consider, including:

- The motivation of somebody wishing to attack de-identified data in order to re-identify somebody within it (this will affect what happens and how).
- What the consequences of a disclosure are (which will affect the motivations of an individual to attempt a re-identification).
- How a disclosure might happen without malicious intent (such as under spontaneous identification).
- How the governance processes, data security and other infrastructure for managing data access affect the risk.
- The other data/knowledge that might be associated with the data in question (without which disclosure/identification is impossible if the data have had direct identifiers removed or replaced).
- Differences between the data in question and other data/knowledge (often referred to as data divergence, see Appendix B.2.2).

[3] We refer a reader interested in the technical discussion to Samarati and Sweeney (1998) and Samarati (2001), the thorough critique by Domingo-Ferrer and Torra (2008) and the recent review in the context of privacy models by Domingo-Ferrer et al (2016).

Combining these considerations with the framework of statistical de-identification creates the fourth type, that is: functional de-identification. As well as looking at the data itself, functional de-identification requires consideration of the interaction of the data with the contextual factors which Mackey and Elliot (2013) refer to as the data environment. Recall that the data situation is the term meaning the data interacting with its environment. It is these concepts that we will be explaining in the course of this book.

Although we have presented functional de-identification as a separate type it does in fact overlap with other types and specifically it still requires the technical know-how that characterises statistical de-identification.

Appendix B Understanding disclosure risk

The task of understanding and assessing disclosure risk in a given dataset is very complex and a topic of current research activity. In this section, we present the ideas that are most important and useful for the practitioner engaged in de-identifying personal information. More detail can be found in one of the recent books (Duncan et al 2011 or Hundepool et al 2012).

Statistical disclosure control is a complex topic and it is not our intention to attempt to give a full discussion of all the possibilities and nuances. If you want to dig deeper we would recommend you read one of the recent books (Duncan et al 2011 or Hundepool et al 2012). Here, we sketch the ideas that are most important and useful for the practitioner engaged in de-identifying personal information.

B.1 Types of disclosure: Re-identification and attribute disclosure

Technically, there are two main types of disclosure: re-identification and attribute disclosure. Re-identification (or identity disclosure) is the process of associating an identity to some data. Attribute disclosure (or attribution) is the process whereby some piece of information is associated with a population unit.

The two processes can sound very similar but the distinction is quite important in terms of how disclosure risk is assessed for different types of data. In essence, identification means we find a person; attribute disclosure means we learn something new about a person. Although the two processes often occur simultaneously, they can in fact occur separately.

Accurate re-identification typically (but not always) leads to attribute disclosure, but attribute disclosure can happen without re-identification. For example, suppose I somehow know that one of five of the records in a dataset corresponds to you, but I don't know which one. This could happen if, for example, you told me that you participated in a particular survey for which I have access to the data, I also know that you live in Anysuburb, and there are five records in the dataset for individuals living in Anysuburb.[4] Suppose now that all of those five records include the information that the data subjects are bank managers. It follows that I now know you are a bank manager even though I have not associated your identity with a particular record. Without further information, the best I could do is guess which of the five records corresponds to you, with a 20% chance of being right.

Importantly, the attribute disclosure process may also - as a side effect - increase the risk of re-identification. In the example given above, I now know that one of only five records of the dataset corresponds to you, and so I may be motivated to try to find some additional information that would distinguish your record from the other four.

[4] This example is a bit artificial in that it further requires that any other information I know about you is not in the data. However it does serve to illustrate the general phenomenon in a simple manner.

In summary, when reducing the risk of disclosure via a de-identification process, data custodians must reduce the risk of both re-identification and attribute disclosure to an acceptably low level.

B.2 Building disclosure scenarios

A key component of a well-formed SDC exercise is the development of disclosure scenarios to ground risk analysis, through semi-formal specification of the risks. Put simply, until you know what could happen, you are stuck with only a vague idea that the data are risky, and quite apart from being a stressful state of affairs this does not get you anywhere in practical terms.

Broadly speaking there are two types of disclosure risk: inadvertent disclosure and disclosure occurring through deliberate action by an intruder.

B.2.1 Inadvertent disclosure and spontaneous recognition

A simple example will suffice to illustrate the notion of spontaneous recognition. Living next to me is a young married couple – very young in fact, both are sixteen. Unfortunately, the woman dies in childbirth leaving the man a 16-year-old widower with a baby.

Putting aside the sadness of this story, we suppose most people would agree with the assertion that this combination of a small number of characteristics is extremely rare. Why is that? Well, we all have an intuitive knowledge of the population, biased perhaps by our own circumstances but reliable enough to enable us to assert with confidence that 16-year-old married couples are unusual, 16 year old widowers are likely to be very rare and 16 year old widowers with a young child even more so. Might there be a good chance that my neighbour is the only one in Australia, or at least in my area?

Now suppose that I am using a dataset from which direct identifiers have been removed, and I come across a record of a sixteen year old widower with a young child who lives in my area. I might assume that it is my neighbour. This is spontaneous recognition: the unmotivated identification of an individual in a dataset from personal knowledge of a small number of characteristics.

Of course such judgements are subjective and subject to availability bias,[5] overconfidence effects and other forms of cognitive bias. So claims to have found someone in data can easily be misjudgements. Let us look at the example a little more objectively. At the 2011 Australian Census there were seventy two 16-year-old widowers in Australia (according to ABS Census TableBuilder). So my neighbour is not unique but an example of a rare combination of attributes. However, if one added in the fact that this person has a young child and included postcode then the probability of the data actually singling out my neighbour may be unacceptably high. So, theoretically, the risk of inadvertent and accurate spontaneous recognition is non-zero.

Other factors which will impact on the risk of spontaneous recognition are the size of the dataset, whether the user has response knowledge and who the users are, as follows.

[5] Availability bias refers to the fact that decision-makers give preference to information that is more recent, personally observed, and/or more memorable for any reason.

Response knowledge: We will talk about this in more detail shortly. But simply put, if I know you are in the dataset then I am more likely to spot your combination of characteristics and subsequently more likely to assume that it is you.

Dataset size can have a counterintuitive effect. If I know you are in the dataset (response knowledge), then a smaller dataset effectively decreases the size of the haystack so it increases the likelihood of coming across you.

Who the users are: With open data the users are potentially the whole world and if it is high utility data then the actual user base might be very large. The larger the user base the more likely a spontaneous recognition event will be. In some data situations there might be a relationship between the user and the data subjects (for example an academic doing research using data on students) and this can increase the risk.

One data situation where all three of these factors can come into play is the in-house survey and in particular the staff satisfaction surveys that are now commonplace in all sectors. The datasets tend to be small and drawn from a particular population with which the users of the data (the organisation's management) have a relationship. The users know that many (or even all) members of staff will be in the survey. In this type of data situation spontaneous recognition can be a serious possibility.

B.2.2 Deliberate attacks and the data intruder

In SDC, the agent who attacks the data is usually referred to as the data intruder (or, attacker, snooper, or adversary). As soon as you consider such a character as a realistic possibility rather than a shady abstraction, several questions immediately arise such as who might they be and what might they be trying to achieve by their intrusion? Considering such questions is an important first stage in the risk management process. Elliot and Dale (1999) have produced a system of scenario analysis that allows you to consider the questions of who, how and why. This method involves a system of classification which facilitates the conceptual analysis of attacks and enables you to generate a set of key variables that are likely to be available to the data intruder. This system was further developed by Elliot et al (2016a) for the purposes of the Anonymisation Decision-Making Framework. The classification scheme is as follows:

INPUTS

- **Motivation:** What are the intruders trying to achieve?
- **Means**: What resources (including other data) and skills do they have?
- **Opportunity**: How do they access the data?
- **Target Variables**: For a disclosure to be meaningful something has to be learned; this is related to the notion of sensitivity.
- **Goals achievable by other means?** Is there a better way for the intruders to get what they want than attacking your dataset?
- **Effect of Data Divergence**: All data contain errors and/or differences from reality. How will that affect the attack?

INTERMEDIATE OUTPUTS (to be used in the risk analysis)

- **Attack Type**: What are the technical details the statistical/computational method used to attack the data?
- **Key Variables**: What information from other data resources is going to be brought to bear in the attack?

FINAL OUTPUTS (the results of the risk analysis)

- **Likelihood of Attempt**: Given the inputs, how likely is such an attack?
- **Likelihood of Success**: If there is such an attack, how likely is it to succeed?
- **Consequences of Attempt**: What happens next if they are successful (or not)?
- **Effect of Variations in the Data Situation**: By changing the data situation can you affect the above?

This approach in scoping the who, why and how of an attack owes as much to criminology as it does to technical risk analysis.

In order to make sense of this scenario-classification scheme you need to understand a set of basic concepts: key variables, data divergence, and response knowledge. We will go through each of these in turn explaining how they fit into the scenario classification scheme as we go.

Key variables

The pivotal element in the scenario analysis is the identification of the key variables. These are the means by which the intruder achieves re-identification allowing the association of an identity with some target information. Key variables are those for which auxiliary information on the data subjects is available to the data intruder, including identifiers such as name and address. The intruder compares the values of the key variables in the target record with those in the auxiliary information, and if they coincide then a match is found and is usually claimed to be a re-identification of the target individual. See Figure 1 for a schematic view of how this works.

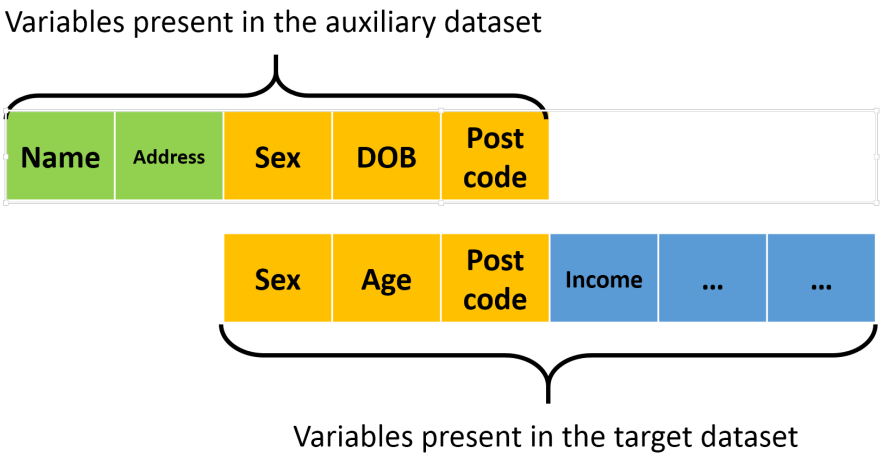


Figure 1 An illustration of the key variable matching process leading to disclosure. The key variables (in yellow) are Sex, Age and Postcode, adapted from Duncan et al (2011)

In Figure 1, the lower row represents the target record, with key variables Sex, Age, and Postcode, and target variables including Income. The auxiliary information represented by the upper row, comprising a file containing Sex, Age, Postcode, Name and Address. If the intruder finds a match between the values of the key variables Sex, Age, and Postcode in the target record with those in

the auxiliary information, then the intruder associates the value of Income with the matching Name and Address. Ideally, from the intruder's point of view, the coding method of a key variable must be the same on both the auxiliary information and the target dataset, or one must be able to be transformed into the other.

Essentially, there are four sources of auxiliary information:

1. datasets containing the same information for the same (or sufficiently similar) population,
2. information that is publicly available (e.g. in public registers or on social media),
3. information obtained from local knowledge (e.g. house details obtained via a real estate agent or by physical observation), and
4. information obtained through personal knowledge (e.g. things I know about my neighbours or work colleagues).

There are some similarities between the notions of key variable and of indirect identifier. The distinction is that a key variable is specific to a particular scenario (for example a particular combination of datasets) whereas the term indirect identifier is focused on the dataset itself and which variables could be used as identifiers in any scenario. So in effect the set of indirect identifiers is the set of all possible key variables across all possible scenarios. But – and this is crucial – one would very rarely (if ever) encounter a situation where one considered all potential indirect identifiers simultaneously as most scenarios will only involve a subset – the key variables for that scenario.[6]

Data divergence

Another crucial point in the scenario framework is consideration of data divergence. All datasets contain errors and inaccuracies. Respondents do not always supply correct data. Interviewers make mistakes in recording. Data coders transcribe incorrectly. Data items are missing. Missing or inconsistent values may be imputed using methods with no guarantee of accuracy. Data may be months or possibly years old before they are shared or released and characteristics will have changed since the data were generated. This is true of the target dataset as well as auxiliary information held by an intruder. The combination of these will introduce the possibility of error in any match.

Collectively, we refer to these sources of errors and inaccuracies in the data as data divergence. The term refers to two types:

- data-data divergence, or differences between datasets, and
- data-world divergence, differences between datasets and the world.

In general, both types of divergence can be assumed to reduce the success rate of matching attempts. However, where two datasets diverge from the world in the same way, which we call parallel divergence, then the probability of correct matching is unaffected. This would be the case,

[6] We note that sometimes the term *quasi-identifier* is used to refer to both key variables and indirect identifiers. This can cause confusion in practice, so we prefer to keep the terms separate.

for example, if a respondent has lied consistently or when two datasets both have out-of-date but identical data.

Taking data divergence into account in a coherent way is complicated and it tends to mean that orthodox risk measures overestimate risk (given the scenario). Elliot and Dale (1998) estimated that the effect in their particular study was to reduce the number of correct unique matches by as much as two thirds. This is one reason why it can be important to carry out penetration testing as well as Data Analytical Risk Assessments (see Appendix B.5.1).

Notwithstanding the above remarks, a paradox of data divergence is that it is not a reliable protection of privacy. First, for any particular individual-record pair there may be no divergence at all. Second, analysts are becoming increasingly sophisticated in dealing with finding matches in the face of 'fuzziness' – when we talk through the process of doing a penetration test in Chapter 3 you will see an attempt to tackle that issue.

So the best way to think about data divergence is that it provides you with a little extra protection – a margin of error rather like the reserve in a car's petrol tank. It is good as back up but not to be relied on.

Response knowledge

The issue of response knowledge can be captured by a single question: 'Do I know that you are in the data?' If the answer to that is 'yes' then I am said to have response knowledge of you in respect of that data.[7] In that case, one key element of uncertainty, whether you are even in the data at all, is removed. In practice, response knowledge can occur in one of two ways:

1. The intruder knows that
    a. the data correspond to a population and
    b. the target is a member of that population
2. The intruder has ad hoc knowledge about a particular individual's presence in the data (e.g. my neighbour told me that she had been surveyed).[8]

The second is relatively simple to understand and is particularly pertinent to an open data situation. The first is more complex as 1(b) can be nuanced. Consider a hypothetical de-identified dataset of the members of the Anytown Bicycle Club. Straightforwardly, I could know that you are in the club and therefore in the dataset. That is clear cut response knowledge but I could have other information which falls short of full response knowledge but is nevertheless informative. I could know that you live or work in Anytown or that you are an avid cyclist or perhaps that you are a compulsive club-joiner. All of these constrain the super-population that contains the Anytown

[7] Of course I might be wrong – perhaps the information which tells me you are in the data is out of date or misattributed. Technically, response knowledge should be called something like 'beliefs about particular population units' presence in a particular dataset' but it is not a very reader friendly formulation. This is part of a more general issue of data divergence and applies even to direct identifiers (I might think I know your name and address but I could be mistaken). We will discuss this general issue in more detail shortly

[8] Another theoretical possibility is that the intruder has inside knowledge of the data collection process. This would imply a complex security breach involving a situation where the intruder did not gain access to the raw data but did have access to a de-identified version of the data. Although this should not be discounted it is obviously quite obscure and the key problem here is the security breach, not the de-identification problem.

Bicycle Club population and that in turn increases the effective sampling fraction.[9] As we will see in Chapter 3 the sampling fraction is an important element of the risk.

B.3 Uniqueness

The example above brings us to uniqueness, one of the fundamental concepts in disclosure risk assessment, which underpins much of the research on disclosure risk analysis. A record is unique on a set of key variables if no other record shares its combination of values for those variables.

For disclosure risk purposes we need to examine two types of uniqueness on a set of key variables: population uniqueness—a unit is unique in the population data file such as a census; and sample uniqueness—a unit in the sample is unique within the sample file.

A simple example – using just two variables – should clarify the relationship. Imagine that there are twenty people living in Anytown, and we have some data on their ages and sexes as shown in Table 1.

Name	Age group	Sex
Johnny Blue	0-16	Male
Jenny Blue	0-16	Female
Sarah White	0-16	Female
Sam Brown	0-16	Male
Julia Black	0-16	Female
James Green	17-35	Male
Peter Grey	17-35	Male
Jemima Indigo	17-35	Female
Jim Blue	17-35	Male
Joshua White	17-35	Male
Joan White	17-35	Female
Jill Brown	17-35	Female
James Brown	36-64	Male
Jessica Black	36-64	Female
Joe Orange	36-64	Male
John Black	36-64	Male
Jacqui Purple	65+	Female
Julie Beige	65+	Female
Jane Azure	65+	Female
Jeffrey Magnolia	65+	Male

Table 1 Ages and sexes of all people living in Anytown

If you peruse Table 1 you will see that there are two people who have a unique combination of characteristics in the data, Jeffrey Magnolia and Jessica Black. Now imagine that we take a 50%

[9] The sampling fraction is the proportion of a population to be included in a sample. It is equal to the sample size divided by the population size.

random sample of this population. One possible sample is shown in Table 2; we have also replaced name with sample ID.

Sample ID	Age group	Sex
1	0-16	Male
2	0-16	Male
3	0-16	Female
4	17-35	Male
5	17-35	Female
6	17-35	Male
7	36-64	Male
8	36-64	Male
9	65+	Female
10	65+	Male

Table 2 Ages and sexes of a 50% sample of the people living in Anytown

If you peruse Table 2 you will see that we have 4 records that are unique in the sample; the ones with sample IDs 3, 5, 9 and 10. Only one of these (number 10, the one corresponding to Jeffrey Magnolia) is actually unique in the population. The other sample uniques have statistical twins in the population – units sharing the same attributes. So, for example, we cannot tell whether record 9 corresponds to Jacqui Purple, Jane Azure or Julia Beige. Jessica Black who is unique in the population is not in the sample.

In one form or another, these two concepts – population and sample uniqueness – form the basis of many of the disclosure risk assessment methods for microdata (files of records about individuals). If a unit is population unique then disclosure will occur if an intruder knows it is population unique. Much of the methodology in this area concerns whether sample information can be used to make inferences about population uniqueness.

The simplest inference is that given the sample file, if a record is not unique in the sample file it cannot be unique in the population, while a record that is unique in the population will be unique in the sample if it appears at all. This will not get the intruder very far but as we will see later not all sample uniques are the same.

B.4 Types of attack

B.4.1 Re-identification attacks through linkage

Re-identification through linkage is the most common form of attack. It is similar to but more general than the process shown in Figure 1. The presupposition is that a data intruder has access to some auxiliary information which contains direct identifiers for population units and a set of key variables which are also present on the target dataset. Linkage is conducted to compare records between the target and auxiliary information. If a match is found, then re-identification occurs and the target information is attributed to the individual. In principle, the target information could be any information not already known to the data intruder but in practice, in the scenario framework, we assume that the information has some value in terms of the intruder's goal.

Formal risk assessment for microdata releases usually requires us to understand the probability of the data intruder being able to find correct matches.

B.4.2 Attribution attacks

We illustrate an attribution attack that does not rely on re-identification. Consider the table of counts shown in Table 3.

Occupation	Annual income ($)			Total
	High	Medium	Low	
	>250K pa	40-250K pa	<40K pa	
Academics	0	100	50	150
Lawyers	100	50	5	155
Total	100	150	55	305

Table 3 Table of counts of income levels for two professions from hypothetical population of 305 individuals, adapted from Duncan et al (2011)

Suppose the population represented in this table is everyone at a workshop I am attending – and note that this is an important assumption. Over drinks, I overhear someone saying that they earned over two million dollars in the last quarter. Since I know the person must appear in the dataset, and have high income, then looking at the first column of Table 3 I can infer with certainty that person is a lawyer. This is positive attribution — the association of the attribute 'is a lawyer' with a particular person. Conversely, if I hear somebody talking about their students, I can infer that they do not have a high income. This is a negative attribution — the disassociation of a particular value for a variable from a particular population unit.

Note that, in effect, association and disassociation are different forms of the same process, attribute disclosure arising from zeroes in the dataset. The point to note is that the presence of a (non-structural[10]) zero in the internal cells of a table is potentially disclosive.

B.4.3 Subtraction attacks

We illustrate a subtraction attack, using Table 4. The population in this table differs from that in Table 3 in one respect—there is now one highly paid academic in the population represented by the sample. Given this table, I can no longer make the inferences that I could from Table 3 (at least not with certainty). However, what about myself? I am a member of the population represented in the table and we can assume that I know my own occupation and income! Suppose that I am a highly-paid academic. Given this extra piece of knowledge, I can subtract 1 from the high income

[10] A structural zero occurs when a combination of attributes is impossible. For example, the number of three year old married people would, in Australia, produce a structural zero because of Australian law. Non- structural zeroes appear where there are possible combinations of attributes which happen not to be instantiated. So there might happen to be no sixteen year old married people in Anytown in my data but the existence of such a person is possible.

academic cell in the table, which then reverts to Table 3 and I am back to the situation where I can make attributions from overheard partial information about particular individuals.

Occupation	Annual income ($)			Total
	High	Medium	Low	
	>250K pa	40-250K pa	<40K pa	
Academics	1	100	50	151
Lawyers	100	50	5	155
Total	101	150	55	306

Table 4 Table of counts of income levels for two professions from hypothetical population of 306 individuals, adapted from Duncan et al (2011)

We can extrapolate further. Consider a situation where I have complete information (in terms of the two variables Occupation and Annual income contained in Table 4) about multiple individuals within the population. Such information can be represented as a table of counts of the subpopulation of the individuals for whom I have complete knowledge. Under the assumption that identification information is available for both that subpopulation and for any auxiliary information I gain through overheard conversations (or other sources of data), I can subtract the whole of that subpopulation table from the population table before proceeding. In principle, this could lead to more zeroes appearing in the residual table. The 'low-paid lawyers' cell would be particularly vulnerable to further subtraction and this illustrates a further crucial point: whilst zero counts are inherently disclosive, low counts also represent heightened disclosure risk, because it is easier to obtain sufficient auxiliary information to enable subtraction to zero of small cell counts in comparison with high cell counts.

B.4.4 Inference attacks

Beyond the risk of subtraction to zero, there is another sense in which low cell counts constitute a risk. Consider again Table 4. Now recall that it is not possible, without auxiliary information about the population represented in the table, to make inferences about any given individual with certainty. However, imagine again that I overhear someone at the workshop boasting about their high income. Whilst I cannot say with certainty that this individual is a lawyer, I can say so with a high degree of confidence. From the table, the conditional probability that a randomly selected person is a lawyer given that they are a higher earner is greater than 0.99.

This is inference – the capability of a user of some data to infer at high degrees of confidence (short of complete certainty) that a particular piece of information is associated with a particular population unit. Such inferential capacity could also in principle be derived from statistical models and other statistical output.

Depending on circumstances, this inferred knowledge may be good enough to meet the data intruder's goals. Deciding in any categorical sense what level of certainty of inference constitutes a problem is impossible. The best approach for dealing with this issue is to understand whether an inference at a particular certainty level would be a success for the intruder and then whether that

inference would cause harm to a data subject. This reiterates the necessity of well-formed disclosure scenarios.

B.4.5 Differencing attacks

A differencing attack is possible with variables for which there are multiple different plausible coding schemes, where the categories in those coding schemes are not nested but instead overlap. For example, given a table with information on 20–25 year olds and a table with information on 20–24 year olds, then the difference between the two tables will reveal information about 25 year olds only. If the difference between cells contains a small number of individuals then a disclosure is more likely. As another example, there is a risk of differencing attack whenever information is released using two different and overlapping codings for a variable such as geography.

This situation may occur for tables or maps with different geographical codings potentially allowing more information to be revealed about individuals in the overlaps than intended from each individual table. Although it could happen with any variable the issue most commonly comes up with geography.

The result of this differencing is that whilst a table may be considered safe in isolation, this may not be the case for multiple tables when overlain with one another.

B.4.6 Complex attacks

The attacks mentioned above are the simple ones. There are more complex operations that a sophisticated intruder can try, for example: table linkage, mashing attacks, fishing attacks,[11] reverse fishing attacks and so forth. It is outside the scope of this book to go into the details of these but suffice it to say that all of these involve bringing together multiple data sources. In practice if one covers the simple attacks then the complex ones also become more difficult to execute.

You must also bear in mind that if you release multiple data products from the same personal information source into the same environment then you will be increasing the risk and you therefore need to proceed with caution. One way in which this comes up is where both microdata samples and aggregate population counts are released from the same underlying dataset. To give a simple illustration, let us return to our hypothetical sample dataset in Table 2 and add another variable, ‘has cancer’, to it, as in Table 5.

[11] Fishing attacks should not be confused with Phishing. Phishing is fraudulently obtaining personal authentication information (usually passwords) by pretending to be a third party (often a bank). A fishing attack on the other hand is the identification of an unusual record in a dataset and then attempting to find the corresponding entity in the world.

Sample ID	Age group	Sex	Has cancer
1	0-16	Male	No
2	0-16	Male	No
3	0-16	Female	No
4	17-35	Male	No
5	17-35	Female	No
6	17-35	Male	Yes
7	36-64	Male	No
8	36-64	Male	No
9	65+	Female	No
10	65+	Male	Yes

Table 5 Hypothetical 50% microdata sample of the people living in Anytown

Now if I know a person who is Male and 65+ who lives in Anytown then I might suspect that it is case 10, but it is a 50% sample so I cannot be sure that my acquaintance is even in the data. However, suppose that the data custodian also publishes Table 6 on its web site.

Sex	**Age group**				
	0-16	17-35	36-64	65+	Total
Female	3	3	1	3	10
Male	2	4	3	1	10
Total	5	7	4	4	20

Table 6 Cross-tabulation of people living in Anytown by age group and sex

On its own, Table 6 looks fairly innocuous – but by combining this with the microdata in Table 5, to which the data custodian has allowed me access, I am able to ascertain that my acquaintance has cancer. This example is obviously quite simplistic. With real data situations the interactions between different data products drawn from the same data source can be more subtle. To reiterate the take-home message here: be very careful if you are considering releasing multiple data products from the same data source.

B.5 Types of formal disclosure risk assessment

Broadly speaking there are two types of disclosure risk assessment: Data Analytical Risk Assessment and penetration testing. The two approaches have complementary advantages and disadvantages.

B.5.1 Data Analytical Risk Assessment (DARA)

This is sometimes referred to as (statistical) disclosure risk assessment. It covers a large range of techniques from the very simple (counting uniques or identifying small cells) to the more complex

involving constructing statistical or computational models.[12] What the techniques have in common is that they take the dataset in question as an analytical object, treating disclosiveness as a property of the data and attempting to identify the level of that property latent in the data.

Done well, DARA should be grounded in scenario analysis. However, even with this in place, there are several differences between the DARA analysis and what would happen in a real attack; most importantly that no auxiliary information is involved in DARA. Having said that, if the analyst is mindful that (no matter how sophisticated the techniques) they are only producing proxy measures for the real risk then DARA can be very informative.

In Chapter 3 we will present one approach that you can take to DARA.

B.5.2 Penetration testing

Another way of assessing disclosure risk, detailed in two UK-based case studies,[13] is what we refer to as penetration testing (also known as intruder testing). The idea of penetration testing is to replicate what a plausible motivated intruder might do (and the resources they might have) to execute a re-identification and/or disclosure attack on your data.

As noted in the OAIC's de-identification guidance, a 'motivated intruder' is someone who is relatively competent, who has access to external data resources such as the internet and public documents, and who is willing to make enquiries to uncover information (OAIC 2014c). They are not assumed to have specialist knowledge or advanced computer skills, or to resort to criminality. You can of course use a different set of assumptions about the type of knowledge skills and resources that an intruder has if to do so makes sense within your own scenarios. However, it may not always be reasonable to assume that intruders will obey the law.

There are essentially four stages to a penetration test:

1. data gathering

2. data preparation and harmonisation

3. the attack itself; and

4. verification.

Stage 1 tends to be the most resource-intensive whereas stages 2 and 3 require the most expertise. We go into these in more detail in Chapter 3.

There are three core advantages of penetration testing as a risk assessment method compared to DARA approaches:

1. It mimics more precisely what a motivated intruder could do.

2. It will explicitly take account of data divergence.

3. It is based on real data gathering and real external data.

---

[12] We will not go into the details of the modelling approaches here and would refer the interested reader to Hundepool et al (2012) for a recent technical review.

[13] http://ukanon.net/ukan-resources/case-studies/

In other words it is grounded. Against this, it has one important disadvantage: it will be tied very tightly to one particular exercise and therefore does not necessarily represent all of the things that could happen. This disadvantage is the flip-side of the advantages and indeed is an issue with all testing regimes: one trades off groundedness against generality and so in practice one should combine data analytical techniques with penetration testing rather than relying solely on either one.

Appendix C Reducing the risk of disclosure

In this section we review the various options you have to reduce the risk of disclosure from your data to a negligible level; in other words to carry out functional de-identification. These options fall into two groups, those focused on the data and those focused on the data environment. Normally you will need both, however environment-based controls do provide the potential to reduce the risk of disclosure significantly, possibly more so than can be achieved for the same utility impact by modifying the data itself.

Before we move on to discuss the solutions in detail, we first want to discuss the unavoidable trade-offs that you will need to make as part of your de-identification process.

C.1 Risk-utility and other trade-offs

Because de-identification is about producing safe, usable data, we need to understand the trade-off between the two (while still ensuring that our legal obligations are met). Often the information that makes data risky is what makes it of interest to bona fide analysts. However, that is not always the case and as we will see in Chapter 3, one of the important parts of functional de-identification is considering the use case. Why are you sharing or releasing these data and what information is necessary to achieve that end?

Let us look at the example of the ABS enabling secure access to linked 2006 Vocational Education and Training (VET) in Schools and 2011 Census of Population and Housing unit record level data.[14] Engagement with The National Centre for Vocational Education Research (NCVER) identified demand for this linked data to answer important questions around the post-school outcomes of students who have undertaken a VET in Schools program. The new linked dataset was of much higher utility than the individual datasets with researchers being able to analyse longer term outcomes than was previously possible; value also came from the detailed information available in Census data. There was a low disclosure risk since the dataset was only accessed from within the secure environment of the ABS DataLab. Resourcing constraints make it difficult to carry out bespoke projects such as this without funding from other agencies. Nevertheless, the needs of data consumers are important considerations when designing de-identification processes.

A second important trade-off is a three-way balancing of data environment risk (risk associated with issues like security, the number of users, governance, etc.), disclosure risk (the properties of the data, given the environment, which make it possible or not to re-identify a person) and the sensitivity of the data. As is hopefully clear by now, total risk in a data situation is a function of all three of these so that if one increases then the others must be decreased to compensate (if one is to maintain risk at the same functional level). So, for example, if you are comparing a dataset containing non-sensitive information with a second containing sensitive health information, then

[14] http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/4260.0Main+Features32006-2011

the environmental and disclosure controls on the latter should, all things being equal, be stronger than on the former. As another example, if one is thinking of releasing as open data a dataset that was previously only available under special licence, then one must increase the disclosure controls applied to the data and/or decrease the sensitivity of the data (by, for example, removing sensitive variables).

This is just common sense but it does suggest a useful insight. In a dynamic data situation, if the data in the original environment are regarded as sufficiently safe (this will normally be so) and if overall risk, taking those three components into account, in the destination data situation is no higher than in the original data situation, then the destination data situation can also be regarded as safe. This conceptualisation is called comparative data situation analysis and is particularly useful for data sharing.

Comparative analysis can also be useful if a gold standard dataset exists which has been shared or released in a similar manner to your intended release or share without problems. One such gold standard dataset is census microdata, the record level datasets released from a population census. In Australia, samples of census microdata have been released under end user license since 1981. The 1% Census Confidentialised Unit Record File (CURF) is only made available to employees of organisations where a Responsible Officer has signed an agreement with the ABS for use of these files. As part of the preparation work for these releases, extensive and detailed work on the disclosure control for the CURF is carried out before and after each census. Rich and deep technical analyses are conducted. Penetration tests are carried out. To date, there have been no re-identification issues (that anyone is aware of). So, if one has a comparable data situation to that of the release under licence of these microdata, one has an available comparison to a tried and tested data situation and can glean insight from the intensive work that was done.

C.2 Environment-based solutions

Environment-based solutions essentially control data users' interactions with the data in some way to reduce the risk in one or more (and usually all four) of the elements of the data environment (other data, agents, governance processes and infrastructure). The key point to remember is that one cannot make a judgement about whether data are de-identified or not without reference to their environment. The implication of this is that one can de-identify data by the operation of environmental controls as well as by the operation of controls on the data itself.

Following Duncan et al (2011), options for environmental controls can be broadly characterized as answering 'who', 'what', and 'where and how' questions:

- Who has access to the data?
- What analyses may or may not be conducted?
- Where is the data access/analysis to be carried out and how is access obtained?

These questions are interrelated – a decision about one has implications for the others. We will consider each in turn.

C.2.1 Who can have access?

The 'who' question is essentially all about agent control. In most basic quantitative terms the risk level of 10 people accessing your data is considerably smaller than if you open it up to 10,000. Beyond the simple additive effect of more people contributing some quantum of risk there is the additional effect that opening up access necessarily implies more relaxed governance. With 10 people it is possible to think about vetting procedures, but vetting 10,000 will inevitably be less sensitive and more routine.

This raises the question of how the data custodian identifies those classed as 'safe people'. Is there a method that establishes those individuals or organisations that the data custodian should trust and those that they should not? At a high level, organisations or individuals with a track record of good practice in data security and stewardship may be given greater data access rights than those without. Often, restricted access conditions stipulate that users must have specified credentials to get access to data. Here are some criteria by which a data custodian might assess a potential data user.

- Whether he/she is associated with some organisation that can assure compliance with the data custodian's data access requirements.
- If the proposed use is for research then the researcher is able to demonstrate the ability to do research of scientific merit.[15]
- Whether he/she has undergone some sort of accreditation or 'safe user' training.

In Australia, the ABS in collaboration with international experts has recently developed training for the accreditation of researchers. Attendance at this training and passing a subsequent test is a requirement for access to data using the DataLab.[16] Similarly, users of the Secure Unified Research Environment (SURE)17 are required to complete training prior to approval for access to the SURE facility.

As in all de-identification matters the key is proportionality. The degree to which *agent controls* should be applied will be related to the disclosiveness and sensitivity of the data and inversely to the degree of other environmental controls.

C.2.2 What analysis is permitted?

Governance control can constrain the projects that can be undertaken with the data. This may be in the form of categorically prohibiting certain types of analysis or may require a project approvals process. For example, the Population Health Research Network (PHRN)[18] requires that projects are

[15] It may not seem immediately obvious why this is in the list. However, remember that we are in the game of risk management and there has to be some benefit to counterbalance the risk. Sharing data (and therefore taking a risk) for a piece of work with no value would not meet this requirement.

[16] http://www.abs.gov.au/websitedbs/D3310114.nsf/89a5f3d8684682b6ca256de4002c809b/c00ee824af1f033bca257208007c3bd5!OpenDocument

[17] The Secure Unified Research Environment (SURE) is a secure computing environment that has been purpose-built for analysis using linked health and health-related data. See https://www.saxinstitute.org.au/our-work/sure/

[18] The Population Health Research Network (PHRN) has been established to build a nationwide data linkage infrastructure capable of securely and safely managing health information from around Australia. See http://www.phrn.org.au/

approved by a number of parties, namely, the relevant data linkage unit, the data custodian responsible for each dataset involved, and a Human Research Ethics Committee. The ABS defines 'safe projects' as those that are for statistical purposes only, and may ask the researcher to demonstrate the value to the public of the proposed research. ABS data cannot be accessed to undertake regulatory or compliance checks.

A related type of restriction is controlling the output. In a strongly controlled environment, some sort of output control is usually necessary. The intuition here is that if the data itself are disclosive (in an open environment) then analytical outputs will also have the potential to be disclosive. Outputs are after all simply a form of data and so, as we will discuss in Chapter 3, the publication of the results of analyses (which is usually what is intended) creates a dynamic data situation.

In all output checking, what is in essence being checked is whether it would be possible to recover (some of) the underlying data from the output. As a simple example, given tabular frequency data one typically would not permit unrestricted requests for multivariate tables of counts,[19] since a sequence of such requests can be used to recover the original data. If a user were to request such cross-tabulations then the request would have to be denied. However, the problem goes beyond this situation. For example, using any regression model in combination with residual plots and the right auxiliary information, it can be possible to recover some of the original data used to generate the model. At this point, developing valid output checking processes that could be automated is an open research question. Therefore, output needs to be checked manually by data centre staff with some expertise.[20]

One way to reduce the burden on the output checkers, used by, for example, the ABS DataLab[21] is to define a very conservative class of outputs as 'safe' and then leave it to the user to demonstrate that anything not on the list is also safe.

C.2.3 Where and how can access be obtained?

In many ways the where and how questions are the key drivers in determining the type of environment that you are working in. There are four modes of access that are currently used for disseminating data for use outside of organisational boundaries:

- Open access
- Delivered access
- On-site safe settings
- Secure virtual access

Open access

Open access (or what can be called unrestricted access) has always been used for publishing some census tabulations and high level administrative data. An instance of free access is the ABS

[19] These are cross-tabulations of two or more variables.

[20] Recently, some progress has been made with automating at least some of the functionality of output checking, see for example Thompson et al (2013) and O'Keefe et al (2016).

[21] http://www.abs.gov.au/websitedbs/D3310114.nsf/home/CURF:+About+the+ABS+Data+Laboratory+(ABSDL)

QuickStats,[22] which is intended as public use data. Since the ABS has comprehensively checked these data beforehand to ensure they are appropriately de-identified, there are no restrictions on who can access the data, or on what they can do with it. Also, there is usually no monitoring of users or what they are doing. Likewise, the administrative data from government agencies published on the data.gov.au website are openly available to the public. Another mechanism by which data can become open is a freedom of information request. FOI requests are in effect requests to release the information openly, and often websites such as OpenAustralia Foundation's https://www.righttoknow.org.au/ ensure that FOI requests are published.

Until 25 years ago, the dissemination medium of such open data was paper-based, usually in the form of thick volumes of tables. However, web delivery is now far more common, and this has opened up datasets for much wider use. In Australia there is significant pressure arising from both user demand and government policy to make more government data available openly. In the UK, this has led to the development of the Open Government License by the National Archive (National Archive 2016). This licence specifically does not apply to personal data/information. The point of this is to underline that open data environments are really only appropriate to data that are either not personal in the first place or have been through an extremely robust data focussed de-identification process that ensures with a very high degree of confidence that no individual could be re-identified and no disclosure could happen under any circumstances.

In Australia the Australian Government Public Data Policy Statement requires Australian Government entities to publish `appropriately anonymised` government data by default under a Creative Commons By Attribution licence unless a clear case is made to the Department of the Prime Minister and Cabinet for another open license.[23]

Delivered access

Delivered access is a more restricted form of access, in which access to the data is applied for and the data are delivered to the user, most commonly through an internet portal or possibly via encrypted email. The former is common in cases where delivery is potentially to a community of many. The latter is perhaps more common where the data situation is a single site-to-site share users (the Population Health Research Network (PHRN) Secure data transfer (SUFEX) service[24] is an example of this). It is important not to forget that the transfer medium is itself an environment, and that therefore one needs to model potential data transfer media as data environments as well in order to decide the appropriate means of transfer.

Usually, as in the example of the PHRN, the process of applying for a copy of the data requires the user to specify what they are to be used for and invariably they are required to agree to specified conditions on a licence for data access. We discuss such licences below.

---

[22] http://www.abs.gov.au/websitedbs/censushome.nsf/home/quickstats?opendocument&navpos=22

[23] https://www.dpmc.gov.au/sites/default/files/publications/aust_govt_public_data_policy_statement_1.pdf

[24] http://www.phrn.org.au/for-researchers/services-for-researchers/sufex/

On-site safe settings

On-site safe settings are regarded as the strongest form of restricted access, usually including a high level of security infrastructure control. The prospective data user applies for access to the data in a particular location — often in the offices of the data custodian or otherwise at a research data centre (RDC) that has been established by the data custodian.[25] For example, for some datasets, the ABS will stipulate that they may only be accessed in the ABS DataLab at an ABS office. Often, the users are required to analyse the data on a dedicated stand-alone computer and are restricted in the software that they may use. There are also often numerous governance controls in place. For example the user may:

- Not be permitted to take in data transport devices such as USB drives or mobile phones.
- Not be allowed to copy down anything that appears on the screen.
- Be required to log in and out of the facility.
- Only attend at pre-booked days and times.
- Be required to sign a user agreement stipulating that they will adhere to conditions of access such as those specified in 1 - 4 above and undertake not to attempt to identify any individuals in the dataset.

The user will be allowed to take away some analytical output, but usually only after it has been checked by output checkers for disclosiveness.

On-site safe settings may be considered as less than ideal by researchers because

- travel to one of these sites is expensive,
- the facility is only open at certain hours,
- computing facilities may be unfamiliar or inadequate,
- internet access may not be available, and
- it requires users to work in unfamiliar surroundings.

On-site safe settings also impose resource and opportunity costs on the data custodian, in terms of providing technical infrastructure and supervisory staff. However it is worth remembering that such arrangements facilitate research on vitally important but sensitive topic areas that might not be possible in other types of settings.

An alternative approach, used for example by the Office of National Statistics Longitudinal Study (Office of National Statistics 2016), is that the researcher submits the syntax for their analysis software to a dedicated unit which – if it approves it – then runs it. However, this rather awkward procedure is being superseded by virtual access systems.

[25] Examples are the ABS DataLab (http://www.abs.gov.au/websitedbs/D3310114.nsf/home/CURF:+About+the+ABS+Data+Laboratory+(ABSDL)), Statistics New Zealand Data Lab (http://www.stats.govt.nz/tools_and_services/microdata-access/data-lab.aspx), the Administrative Data Research Centres in the UK (www.adrn.ac.uk) and the US Federal Statistical Research Data Centres (http://www.census.gov/about/adrm/fsrdc/locations.html).

Secure virtual access

Virtual access is now widely regarded as the future of research data access. It combines many of the advantages of the physical safe setting with much of the flexibility of having access to a copy of the data from one's desktop.[26] There are two variants on the virtual access theme: direct virtual access and analysis servers.

Direct virtual access uses virtual remote network-type interfaces to allow users to view, interrogate, manipulate and analyse the data as if it was on their own machine. There are two critical differences between direct virtual access and delivered access. Firstly, output is typically checked in the same manner as in an on-site safe setting. Secondly, there is still no possibility of a user inappropriately linking the accessed dataset to another dataset (either because dataset uploads are not permitted or because external datasets are checked before upload and activity is audited) and this restricts the number and type of disclosure scenarios that the data custodian needs to consider.

Analysis servers go one step further than direct virtual access in not allowing direct access to a dataset while allowing the user to interrogate it. In such systems data can be analysed but not viewed. Usually, there is a mechanism for delivering the analysis (for example through uploading syntax files for common statistical packages or, occasionally, through a bespoke interface). The analysis server will return the results of the request for analysis, usually after they have been checked for disclosiveness. From the data custodian's viewpoint, the advantages over direct virtual access are twofold:

- because the user cannot see the data the risk of spontaneous recognition of a data unit is all but removed, and
- there is no risk of the screen being seen by somebody who is not licensed to use the data.

The disadvantage from the user's point of view is that it is more difficult to explore the data.[27]

Licensing

Another governance control tool a data custodian has is licensing, often used in conjunction with other restricted access mechanisms. Licences can be used as a pro forma to be signed by a set of users or in a bespoke data sharing agreement for site- to-site shares. Some common themes in such licensing agreements are:

- Specification of those permitted access (agent controls).
- Data security requirements (infrastructure controls).
- Restrictions on use, particularly prohibition against linking with other files and on deliberate re-identification (other data and governance controls).
- Requirement to destroy the data once the use is complete (governance controls).

---

[26] An intermediate hybrid approach is where safe rooms are installed at user institutions as a semi controlled medium for virtual access. So the user will have to go to the local safe room, but this will involve minimal travel and is therefore less restrictive than an on-site lab. This is being explored by the ABS as a method for allowing academic researchers to have access to administrative data within Australia.

[27] For a useful discussion of the different types of virtual access systems and why a data custodian might choose one over another see O'Keefe et al (2014).

The function of licensing is threefold:

- It clearly distinguishes between those individuals or organisations the data custodian trusts and those that it does not.
- It is a framework for specifying the conditions under which access can occur.
- It can specify sanctions or penalties should the individual/organisation transgress on those access conditions.

It is possible to have licensing at graded levels, with different users having access to data with different levels of disclosure risk (and therefore presumably different levels of other controls, as well as different levels of data utility). In Australia, the ABS makes a distinction between public and research levels of access, with licensing used to access research level data.

By including some infrastructure and governance controls, the licence allows the data custodian to maintain some control over the security of the data and can also provide guidance to the data user regarding good practice. If the data are being provided to the user at their site then various physical and computer security conditions might be required. Here is an example of a set of requirements that might be included in a licence for a single site-to-site share:

- Data must be stored in a dedicated secure data lab.
- There must be an independent locking system to the data storage area.
- There must be extra security at all possible primary and secondary points of entry, extra locks on doors, bars on windows, etc.
- Data must be stored on a stand-alone machine.
- Multiple passwords are required to access the data.
- Devices such as external disc drives/USB ports must be disabled.
- Output must not be removed from the data laboratory and must be destroyed when finished with.
- Entry to the data laboratory must be limited to particular staff.
- Log books must be kept of access.

As well as providing actual security, imposing such conditions may also be intended to change the mind-set of the user, who will hopefully react to them by being more security-aware. The flip side of this is that these conditions may place awkward obstacles in the researcher's usual research method.

Another type of commonly used licence condition asks the user to agree to restrictions on what they can do with the data – in particular, not linking it with other datasets that contain direct identifiers.

The third function of the licensing process involves specifying the sanctions that can be applied to users or their organisations for non-compliance with the licence conditions. In order to serve as deterrents for non-compliance, they must be enforceable. Typical sanctions are fines and removal of the right to access the data. For example, the Australian Bureau of Statistics Undertaking by an

Individual,[28] that binds individual users and must be signed by a Responsible Officer in an Organisation before access to microdata is granted, states: 'In signing this Undertaking, I understand and acknowledge that a breach of these terms and conditions may:

- be an offence under subsection 19(3) of the Census and Statistics Act 1905; and
- result in sanctions which may include, but are not limited to:
  - the Australian Statistician imposing restrictions on me accessing any Unidentified Information for a set period of time, including a life-long ban;
  - the Australian Statistician imposing restrictions on my Organisation's access to any Unidentified Information for a set period of time; and
  - a penalty under subsection 19(3) of the Census and Statistics Act 1905 of up to 120 penalty units ($21,600) or imprisonment for 2 years, or both, …'

The threat of sanctions will be taken most seriously if the data user or their organisation is subject to a security audit by the data custodian.

Overall, licensing can be a useful way to decrease disclosure risks for certain uses of a disclosive dataset by researchers, especially when explicit or implicit sanctions can be invoked.

C.2.4 Summary of environment-based controls

Environment-based controls provide the potential to reduce the risk of disclosure significantly, possibly more so than can be achieved by modifying the data itself, for the same utility impact. All of these controls affect at least one of the four elements of the environment (other data, agents, governance processes and infrastructure) and ultimately disrupt the ability of a user or intruder to re-identify data.

C.3 Data-focused solutions

Data-focused de-identification solutions require that the data to be released or shared is altered in some way. Usually key variables are removed, replaced or aggregated. Sometimes the same thing is done to those target variables likely to tempt an intruder in order to reduce their sensitivity. We divide these solutions into two types of control: data reduction (sometimes called non-perturbative methods, metadata-level controls or non-perturbative masking) where the overall structure of the data is changed and data modification (sometimes called perturbative methods or data-distortion controls) where the data are changed at the level of individual values for individual cases. We will discuss each in turn.

[28] http://www.ausstats.abs.gov.au/ausstats/subscriber.nsf/0/1DB56E8B5D2733C2CA257FE20017DF19/$File/1406055003_Individual_Undertaking_UC7I.pdf

C.3.1 Data reduction

Data reduction controls work with the overall structure of the data. The key components of such controls are the sampling fraction, choice of variables, and the level of detail of those variables. In many ways these are the key tools for carrying out practical de-identification; they are simple to understand and use, do not distort the data and are transparent in their effects.

Sampling

For surveys, the sampling fraction is specified by the study design and so its choice often rests outside the disclosure control process. However for other forms of data there is some value in considering sampling. It cuts down the risk associated with response knowledge by creating uncertainty that a particular population unit is actually represented in the data, and thus decreases the likelihood that a match is interpreted as a true re-identification. Even a 95% random sample creates uncertainty and for most purposes will not unacceptably reduce the analytical power of the data.[29]

Impact on risk: Sampling is one of the most powerful tools in the toolbox. The key point is that it creates uncertainty that any given population unit is even in the data at all.

Impact on utility: The impact of random sampling is modest; essentially it will increase the variances of any estimates and reduce statistical power. However, if a user wants to analyse small sub-populations the sampling may reduce their capacity to do this.

Choice of variables

An obvious mechanism of disclosure control is excluding certain variables from the released dataset. The data custodian can

- reduce the number of key variables to which a plausible data intruder is likely to have access, and/or
- reduce the number of target variables.

These choices flow naturally from the scenario analyses described in Appendix B.5. With microdata, the choice is whether a variable appears in a dataset or not. With aggregate data, the choices are about which variables will be included in each table. For point-to-point data shares the variable selection will be driven by the requirements of the user although in practice these may be more negotiable than might initially be apparent.

Impact on risk: The impact of variable selection on risk very much depends on the variables selected. If key variables are removed then re-identification risk will be reduced. The effect here is to reduce what Elliot and Dale (1999) call key power; the capacity of a set of key variables to discriminate between records and produce both sample and population uniques. If target variables are removed then the sensitivity of the data is lessened and the potential impact of any breach is reduced.

[29] You might wonder what level of sampling fraction is sufficient to impact effectively on response knowledge. There is no absolute firm line, because it will partly depend on other elements in the data situation. However, we have never encountered sampling fractions over 95% and in some (more open) data situations the sampling faction would probably need to be under 50% in order to be effective.

Impact on utility: If a variable is critical to a user's analytical requirements then removing that variable will obviously disable the analysis. With data releases, you should consider how widespread the use is likely to be and whether the goals of release can be met through a more modest variable selection.

Level of detail

Decisions over level of detail complement those over choice of variables. Here you should consider categories with small counts and determine whether merging them with other categories would significantly lower disclosure risk with minimal impact on the informational value of the data. Not surprisingly, many data users would like the maximum level of detail possible on every dataset. But some variables, especially geography and time, can be particularly disclosive. Area of residence is a highly visible component of an individual's identity, and to reduce risk geographical detail is often constrained and data are released at coarser detail than users would like. Similarly, time-based variables, such as exact date of birth, can be straightforwardly identifying when combined with other variables.

Impact on risk: The effect of changing the detail on variables is similar to that of removing variables. It is mainly a mechanism for reducing key power. If a variable has some categories that might be considered sensitive then sensitivity can be reduced by merging these with other categories.

Impact on utility: The impact on utility is similar but more subtle than the impact of removing whole variables. Some variables can be more important than others. Purdam and Elliot (2007) carried out a survey of users to establish the impacts on their analyses of such measures. On most obvious aggregations there was some loss of utility, with users reporting that analyses that they had previously carried out on the data would no longer be possible.

Data modification

The main alternative to data reduction are various forms of data modification. These techniques manipulate the data itself in order to foil re-identification/subtraction strategies by ensuring that an intruder cannot be certain that any match in a re-identification attack is correct or that any zero recovered through subtraction attack is a real zero. In this section, we will look at methods of data modification that are commonly used for disclosure control.

Data swapping

Data swapping involves moving data between records in a microdata set. A particular form of this, often called 'record swapping', involves swapping the geographical codes of two records.

Impact on risk: Data swapping, like most data-focused controls, increases uncertainty. However, as Elliot (2000) showed, the impact on general risk measures is quite modest. It comes into its own in situations where multiple data products are being released from a single data source. For example, a sample of microdata with coarse geography and aggregate population tables of counts for fine geography is a common set of census outputs. Modest data-swapping amongst the fine geographical areas within the coarser areas means the microdata itself is unchanged. However, the modification in the aggregate data will reduce the risk of subtraction attacks including foiling any attempt to link on the fine geography.

Impact on utility: Even done well, the impact on data utility can be significant and it will often affect relationships between variables in an arbitrary and unpredictable manner. For this reason, it is not used routinely in data situations where a single data product is involved.

Perturbation

The idea of perturbation is that data values are changed slightly, and the changed values are released in place of the exact values. Such data distortion is designed to increase the intruder's uncertainty about any match, and so reduce the risk of re-identification. Methods for perturbation include the use of additive or multiplicative ``noise'' from a known distribution, or adding small amounts to count data according to a randomly generated seed.

Imputation

Imputation involves replacing real values with ones that have been generated through a model. In order for this to work without badly distorting the data, it may be necessary to allow the original values to be modelled back in. A critical decision when imputing will be what you tell the user. There are numerous options. You can choose whether to tell them that the data has been imputed, and if you do then you can also choose whether or not to tell them how many values have been imputed, the model that has been used to do that imputation or even the actual values that have been imputed.

Imputation can be attractive if you are already using imputation to deal with missing values.

Impact on risk: It is difficult to generalise about the risk impact of imputation as it depends on the mechanism that is used to decide on the new value, how transparent you are about what you have done and how much imputation you have done.

Impact on utility: This really depends on how good a model you have used to produce the imputed values.

Rounding

Rounding is a technique most commonly used with tables of counts. In the simplest form all the counts are rounded to the nearest multiple of a base (often three, five, or ten). Counts which are a multiple of the base number remain unchanged. Normally, the margins are rounded according to the same method of the internal cells. Therefore, in many cases this method does not yield an additive table.[30]

One method of de facto rounding which also has some presentational advantages is to release tables of percentages rather than actual counts. Take for example Table 7. Looking at this table, we immediately know that any Dietitian living in Anystate earns less than $50 per hour.

[30] An additive table is one where the row, column and grand totals are correct. When one rounds the values in a table that may well cease to be true.

	Pay per hour ($A)					
Occupation	**<20.00**	**20.00 – 24.99**	**25.00 – 39.99**	**40.00 – 50.00**	**>50.00**	**Total**
Physiologists	838	1923	928	710	51	4450
Dietitians	469	503	341	102	0	1415
Nurse Managers	350	1198	1141	2082	481	5252
Nurses	14376	20361	7986	4571	77	47371
Psychiatrists	40	59	109	198	970	1376
Total	16073	24044	10505	7663	1579	59864

Table 7 A fictitious table of counts showing the pay per hour for residents of Anystate broken down by Occupation

Compare this with Table 8 which presents the same information expressed in terms of row percentages. There are two points here. First, we can no longer tell that the number of dietitians earning >$50 is zero. In fact the range of possible values here is anywhere up to 8. Second the impact of presenting the table this way is minimal, in terms of what might be considered the underlying message of the data about the wage differential.

	Pay per hour ($A)					
Occupation	**<20.00**	**20.00 – 24.99**	**25.00 – 39.99**	**40.00 – 50.00**	**>50.00**	**Total**
Physiologists	19%	43%	21%	16%	1%	7%
Dietitians	33%	36%	24%	7%	0%	2%
Nurse Managers	7%	23%	22%	40%	9%	9%
Nurses	30%	43%	17%	10%	0%	79%
Psychiatrists	3%	4%	8%	14%	70%	2%
Total	100%	100%	100%	100%	100%	100%

Table 8 A fictitious table of counts showing the banded pay per hour for residents of Anystate with selected occupations expressed as percentage of the total number of residents

Impact on risk: Rounding can be very effective in reducing risks when considering individual tables of counts. Smith and Elliot (2008) demonstrate this with data from the UK neighbourhood statistics. Care must be taken to consider the interactions between multiple outputs and particularly what you are doing about the issue of additivity and consistency between marginal totals in different tables.

Impact on utility: For many purposes rounded frequencies are sufficient and using percentages as a form of rounding can be an even more digestible way of presenting information.

Cell suppression

Cell suppression is a statistical disclosure control technique that can be implemented in various forms, whereby the data are only partially released. In one sense, releases of aggregate data are themselves primary examples of suppression, since they are partial releases of the underlying microdata (or underlying high-dimensional table, sometimes called ‘the full table’). If I release two

one-way frequency tables, but not the combined table then I am, in effect, suppressing the cross-classification of those two variables. Cell suppression is effectively a more targeted form of this.

Consider Table 7 again. One alternative is to release Table 9 (where the symbol X denotes a suppressed cell).

	Pay per hour ($A)					
Occupation	**<20.00**	**20.00 – 24.99**	**25.00 – 39.99**	**40.00 – 50.00**	**>50.00**	**Total**
Physiologists	838	1923	928	X	X	4450
Dietitians	469	503	341	X	X	1415
Nurse Managers	350	1198	1141	2082	481	5252
Nurses	14376	20361	7986	4571	77	47371
Psychiatrists	40	59	109	198	970	1376
Total	16073	24044	10505	7663	1579	59864

Table 9 A fictitious table of counts showing the pay per hour for adult residents of Anytown broken down by occupation with cells suppressed in order to reduce disclosure risk

Note that we cannot simply suppress the disclosive cell (dietitians, age >50) as simple arithmetic would allow an intruder to recover it so we must also make what are called complementary suppressions. An alternative possible suppression pattern is shown in Table 10.

	Pay per hour ($A)					
Occupation	**<20.00**	**20.00 – 24.99**	**25.00 – 39.99**	**40.00 – 50.00**	**>50.00**	**Total**
Physiologists	838	1923	928	710	51	4450
Dietitians	469	503	341	102	X	X
Nurse Managers	350	1198	1141	2082	481	5252
Nurses	14376	20361	7986	4571	77	47371
Psychiatrists	40	59	109	198	970	1376
Total	16073	24044	10505	7663	X	X

Table 10 A fictitious table of counts showing the pay per hour for adult residents of Anytown broken down by occupation with cells suppressed in order to reduce disclosure risk

A key point here is that in Table 9 both the users and the intruder can still make inferences about the contents of the suppressed cells. This is not the case in Table 10 because the grand total is suppressed. On the other hand the grand total may well be a piece of information that is published elsewhere and if so it would be simple to unpick the suppressions. The only option for preventing that would be to ensure that the grand total is never published anywhere which may be both impractical and undesirable. For that reason, the pattern in Table 9 will generally be preferable.

Why do we say that we can still make inferences about the suppressed cells in Table 9? Well, for each of the suppressed cells the value is bounded by the other information in the table. Put

simply, for each cell there is a limited range of possible values – referred to as bounds. The bounds for Table 9 can be seen in Table 11.[31]

	Pay per hour ($A)					
Occupation	**<20.00**	**20.00 – 24.99**	**25.00 – 39.99**	**40.00 – 50.00**	**>50.00**	**Total**
Physiologists	170	390	168	710-761	0-51	4450
Dietitians	95	102	59	102-153	0-51	1415
Nurse Managers	71	243	201	402	138	1055
Nurses	5832	8260	2834	1753	224	18903
Psychiatrists	8	12	12	30	217	279
Total	6176	9007	3274	2350	597	21404

Table 11 A fictitious table of counts showing the pay per hour for adult residents of Anytown broken down by occupation showing the bounds for the cells suppressed in Table 9

Impact on risk: Suppression can be effective in hiding disclosive cells. However you should be aware of the actual intervals that are being implicitly published. As with rounding, care also needs to be taken when releasing multiple tables as it may be possible to unpick the suppressions even if that is not possible when considering each table on its own.

Impact on utility: Users tend to strongly dislike cell suppression. Working with tables with suppressed cells is harder work than say the same tables with rounded values.

Value Suppression

Suppression can also be used for microdata where particular variables can be suppressed for particular cases. For example if you had a 16 year old widower with a child on your dataset you might suppress the age on that case – mark it as missing data in effect. This is an alternative to, and arguably more transparent than, imputation.

k-anonymisation

A dataset is regarded as k-anonymised if – on all sets of key variables – each combination of possible values of those variables has at least k records that have that combination of values. In essence, this gives a standard for data to be considered safe.

This is relatively easy to understand and to implement. There are available open software tools that can semi automate the process.[32] However, its simplicity can be beguiling and the user should be aware that there is no method inherent to the k- anonymity model for identifying either the 'correct' level of k or the combinations of the variables that should be considered. Both of these require an understanding of the data environment. Without such understanding, the context is not represented, and the sufficiency of the de-identification can only be estimated from the

[31] The example shown here is relatively straightforward. However, precise bounds calculations can be quite complicated. See Dobra and Fienberg (2000, 2001) and Smith and Elliot (2008) for a discussion of the methods required do this.

[32] See for example ARX http://arx.deidentifier.org/downloads/ (accessed 19/3/2016) or µ-ARGUS http://neon.vb.cbs.nl/casc/..%5Ccasc%5Cmu.htm (accessed 19/3/2016)

properties of the data itself, which as we have argued misses the point. It also implicitly assumes that either you have a population file or that the intruder has response knowledge (otherwise the k is simply a sample k which could be very misleading and may lead to over-aggregation[33]) and there is no easy way of adjusting the method to deviate from those assumptions.

Another issue with k-anonymity is that it does not protect against attribute disclosure. If a record shares key attributes with k-1 other data units, that may not help if all k units share a value on some sensitive attribute. So in Table 12 we have k-anonymised the combination of age and sex to k=3 (by in this case merging the 36-64 and 65+ categories). Unfortunately, because all males in the 36+ group have cancer, I can still infer that any 36+ year old male has cancer.

Person number	Age group	Sex	Has cancer
1	0-16	Male	No
2	0-16	Female	No
3	0-16	Female	Yes
4	0-16	Male	No
5	0-16	Female	No
6	0-16	Male	Yes
7	17-35	Male	No
8	17-35	Female	Yes
9	17-35	Male	No
10	17-35	Male	Yes
11	17-35	Female	No
12	17-35	Female	No
13	36+	Male	Yes
14	36+	Female	No
15	36+	Female	No
16	36+	Male	Yes
17	36+	Female	No
18	36+	Female	Yes
19	36+	Female	No
20	36+	Male	Yes

Table 12 Hypothetical population microdata for the people living in Anytown

To deal with this problem the concept of l-diversity was introduced. l-diversity imposes a further constraint whereby each equivalence class (group of data units sharing the same attributes) must have multiple values on any variable that is defined as sensitive (or in our terms a target variable). Unlike k-anonymity there are various different definitions of l-diversity. The simplest is that there has to be at least l different values for each sensitive variable within each equivalence class on the key variables.

[33] The degree of aggregation required to achieve the desired level of k will become more severe as the number of data units decreases. So if one is only focused on the data (and not the underlying population) then a sample dataset would be more heavily aggregated than the equivalent population dataset which is clearly counterintuitive unless you are assuming response knowledge.

But l-diversity too can lead to counterintuitive outcomes. In Table 13 we have achieved l-diversity with l = 2 but arguably this table is more disclosive rather than less, for two reasons. First, we now know more precisely the type of cancer that the 36+ year old men have, either Liver or Prostate cancer. Second, if we happen to know someone who is in the dataset and has cancer, but do not know which type, then Table 13 reveals the type is either Bone Marrow, Leukaemia, Liver or Prostate.

Person number	Age group	Sex	Cancer type
1	0-16	Male	N/A
2	0-16	Female	N/A
3	0-16	Female	Leukaemia
4	0-16	Male	N/A
5	0-16	Female	N/A
6	0-16	Male	Bone Marrow
7	17-35	Male	N/A
8	17-35	Female	Breast
9	17-35	Male	N/A
10	17-35	Male	Leukaemia
11	17-35	Female	N/A
12	17-35	Female	N/A
13	36+	Male	Liver
14	36+	Female	N/A
15	36+	Female	N/A
16	36+	Male	Prostate
17	36+	Female	N/A
18	36+	Female	Breast
19	36+	Female	N/A
20	36+	Male	Prostate

Table 13 Hypothetical population microdata for the people living in Anytown

To deal with this and other problems with l-diversity a third notion, t-closeness, has been introduced. To satisfy the definition of t-closeness, the value of each distribution-sensitive variable within each equivalence class should be no further than the threshold t from the value associated with the variable's distribution across the whole dataset.

It would be reasonable to say at this stage that we have moved some distance away from the neat and simple idea of k-anonymity. Even if you are using a software package to do the heavy lifting for you, you are still going to need to understand what k, l, and t actually mean for your data and how this relates to what the intruder might be able to do. The risk here is that you make arbitrary decisions led by the privacy model rather than the data situation. We are not averse to the use of privacy models. If used carefully with full awareness of the meaning of the data, k- anonymity and its companion concepts can be useful tools in some data situations. However, they are not magic bullets, being neither necessary nor sufficient.

Differential Privacy

Amongst several attempts to provide a more robust de-identification standard, differential privacy has emerged over recent years as a strong contender.

Informally, the property of differential privacy for an analysis essentially guarantees that adding the data of a single individual to a dataset is unlikely to change the output of that analysis by very much, and hence it should be unlikely that much can be learned about the added individual from the analysis output. Differential privacy is an attractive standard because of its clear standard of privacy and the strong guarantees that it promises.

While differential privacy has had a marked impact on theory and literature in computer science (see Dwork and Roth 2014), it has had far less impact in the statistical literature and statistical practice. The main concern seems to be that the definition of differential privacy does not mention statistical usefulness at all, so that in some cases the guarantees may be so strong that analysis outputs are altered to the point where they no longer provide sensible inferences, see for example (Fienberg, Rinaldo and Yang2010.

Part II Standard Key Variables

Standard keys are generated by organisations carrying out ongoing data environment analysis (scanning the data environment for new data sources). You should be aware that standard keys are generic and are set up primarily for use with licence-based dissemination of official statistics and will not be relevant to every data situation. If you are using a highly controlled access environment, or at the other end of the scale open data, or if you have data that is unusual in any way, this may not be the method to use.

However, the standard keys can be useful because if your data are not safe relative to these standards then in itself that indicates that you may have a problem, even before you consider non-standard keys.

The sets of keys presented here are adaptations to the Australian context of subsets of those generated by the Data Environment Analysis Service at the University of Manchester using the methodology reported in Elliot et al (2011). They are focused on demographics and socio-economic variables. It should be stressed that these lists are time-dependent and are very much subject to change as the data environment changes. **It should also be stressed that these have not been specifically developed for the Australian context.** However, they will serve as a good starting point for considering your own data situation and its key variables.

Within each scenario set, similar scenarios are grouped into scenario subsets.

Appendix D Restricted access database linkage

D.1 Restricted access database cross match (general)

D.1.1 Scenario: Restricted access database cross match (general)

This scenario is based upon an analysis of the information commonly available in restricted access databases.

- Home address
- Age
- Sex
- Marital status
- Number of dependent children
- Distance of journey to work
- Employment status
- ANZSCO (Australian and New Zealand Standard Classification of Occupations)

Attacker Profile: Person with access to restricted access dataset or hacker able to obtain such access.

D.1.2 Scenario: Restricted access database cross match (general, extended)

This scenario is based upon an analysis of the information commonly available in restricted access databases, a slightly extended version of Scenario A.1.1 with additional, less common variables. Typical variables are:

- Age
- Sex
- Marital status
- Number of dependent children
- Workplace (typically a geographical identifier)
- Distance of journey to work
- Household tenure type
- Number of cars
- ANZSCO
- Employment status
- Income

Attacker Profile: Person with access to restricted access dataset or hacker able to obtain such access.

D.2 Restricted access database cross match (health)

D.2.1 Scenario: Restricted access database cross match (health)

This represents an attack from a restricted access dataset which also contains health information. Such datasets are becoming more common. Typical core variables are:

- Home address
- Age
- Sex
- Marital status
- Employment status
- Cultural and ethnic group
- Alcohol consumption
- Smoker/non-smoker
- Long term illness

- Type of primary long term illness (possibly match against multiple variables)

Attacker profile: Individual with access to restricted access dataset.

D.2.2 Scenario: Restricted access database cross match (health, extended)

This represents an attack from an extended restricted access dataset which also contains health information. Such datasets are becoming more common. Typical core variables are:

- Home address
- Age
- Sex
- Marital status
- Employment status
- Cultural and ethnic group
- Alcohol consumption
- Smoker/non-smoker
- Long term illness
- Type of primary long term illness (possibly match against multiple variables)
- Number of dependent children
- Workplace (typically a geographical identifier)
- Distance of journey to work
- Household tenure type
- Number of cars
- ANZSCO

Attacker profile: Individual with access to restricted access dataset.

D.3 Restricted database cross match (personnel)

D.3.1 Scenario: Restricted database cross match (personnel)

This scenario is based on information commonly held in personnel databases. Typically this includes considerable detail on economic characteristics such as occupation, industry, economic status, basic physical characteristics (such as age, sex and cultural and ethnic group) and some information on personal circumstances (area of residence, long term illnesses, marital status and number of children).

- Home address
- Age
- Sex
- Marital status
- Employment status (filter)
- Occupation
- Industry
- Hours of work
- Migration in the last year
- Cultural and ethnic group
- long term illness
- Number of children.

Attacker Profile: Person working in personnel office of large organisation.

Appendix E Publicly available information based attacks

E.1 Commercial database cross match

E.1.1 Scenario: Commercial database cross match (common)

This scenario is based upon an analysis of the information commonly available in commercial databases. Typical variables are:

- Home address
- Age
- Sex
- Marital status
- Number of cars
- Number of dependent children
- Household tenure type
- Employment status
- Socio-economic index
- Household composition

Attacker Profile: Person or organisation with sufficient resources to purchase lifestyle database type information.

E.1.2 Scenario: Commercial database cross match (superset, resource cost high)

This scenario is based upon an analysis of the information available in commercial databases. This is effectively a superset of available variables which could be exploited by a well-resourced attacker who links multiple data sources together.

- Home address
- Age
- Sex
- Marital status
- Number of cars
- Number of dependent children
- Household tenure type
- Accommodation type
- Employment status
- Socio-economic status
- Household composition
- Religion
- Number of rooms
- Income
- Transport to work
- Highest qualification
- Long term limiting illness
- Workplace

Attacker Profile: Person or organisation with sufficient resources to purchase multiple lifestyle databases.

E.2 Local search

E.2.1 Scenario: Local search

This scenario corresponds to what might be obtained through estate agent details combined with the electoral register. The variable age and cultural and ethnic group from the electoral register that could be used in a crude form are not included in this variant. Typical variables are:

- Home address
- Accommodation type
- Sex
- Number of rooms
- Number of bathrooms
- Presence of central heating/cooling

Attacker Profile: Anyone.

E.3 Extended local search

E.3.1 Scenario: Extended local search

This scenario corresponds to what might be obtained through estate agent details combined with the electoral register. The variables (new voter/adult) and cultural and ethnic group that could be used in a crude form from the electoral register are included in this variant. Typical variables are:

- Home address
- Accommodation type
- Sex
- Number of rooms
- Number of bathrooms
- Presence of central heating/cooling
- Cultural and ethnic group
- Age group (new voter/adult) Attacker Profile: Anyone.

E.4 Public Information

E.4.1 Scenario: Public information (low resources, subgroup)

This scenario imagines an intruder who is drawing on publicly available data sources focusing on a particular subgroup or groups, and who is constrained in his/her use of resources.

- Home address
- Cultural and ethnic group (crude)
- Age
- Sex
- Qualifications
- Occupation
- Workplace

E.4.2 Scenario: Public information (high resources, subgroup)

This scenario imagines an intruder who is drawing on publicly available data sources focusing on a particular subgroup or groups, without effective resource constraints.

- Home address
- Cultural and ethnic group (crude)
- Age
- Sex

- Qualifications
- Occupation
- Workplace
- Household tenure type
- Accommodation type

E.4.3 Scenario: Public information (high resources, opportunistic targeting attack)

This scenario imagines an intruder who is drawing on publicly available data sources, targeting a small number of individuals, who have visibility perhaps because of media coverage, without any resource constraints.

- Home address
- Cultural and ethnic group
- Age
- Sex
- Qualifications
- Occupation
- Workplace
- Household tenure type
- Accommodation type
- Marital status
- Country of birth
- Religion
- Ancestry

E.5 Online data sweep

E.5.1 Scenario: Online data sweep (low resources, opportunistic targeting attack)

This scenario envisages somebody trawling the net for available sources of information. The status of such information is questionable since much of it is deliberately self-published. For specific individuals the list of variables may be much longer than this. However, these will be commonly obtainable from online CVs and sites such as dating sites:

- Home address
- Cultural and ethnic group
- Age
- Sex
- Qualifications
- Occupation
- Workplace
- Marital status
- Dependents (y/n)
- Religion
- Income
- Language

E.6 Information about acquaintances

E.6.1 Scenario: Worker using information about colleagues

This scenario is based upon a study of what people commonly know about people with whom they work. Typically this includes considerable detail on economic characteristics, basic physical characteristics and some very crude information about personal circumstances. Typical variables are:

- Age
- Sex
- Cultural and ethnic group
- Occupation
- Workplace
- Distance of journey to work
- Industry
- Hours
- Economic status
- Long Term illness
- Number of children

Attacker profile: Anyone working in a large organisation.

E.6.2 Scenario: Nosy Neighbour

This scenario encompasses information that would be relatively easy to obtain by observation of one's neighbours. Obviously this does not entail either a standard match or fishing type attack. In effect one would be fishing for one's neighbours in the dataset. However if one found a match one could use information in the dataset to determine whether it is rare or not. Typical variables are:

- Home address
- Age
- Sex
- Marital status
- Number of cars
- Number of dependent children
- Number of elderly persons
- Density (persons/rooms)
- Cultural and ethnic group
- Family type
- Accommodation type
- Multiethnic household
- Number of residents
- Number of rooms

E.7 Combined sources

E.7.1 Scenario: Combined public and visible sources

This is essentially the combination of nosy neighbour with publicly available information scenarios. This is quite a resource intensive attack because it involves hunting for information on a small group of people in public records. It is not likely to yield the information below on all neighbours.

- Home address
- Age
- Sex
- Marital status
- Number of cars
- Number of dependent children
- Number of elderly persons
- Density (persons/rooms)
- Cultural and ethnic group
- Family type
- Accommodation type
- Multi-ethnic household
- Number of residents
- Number of rooms
- Presence of central heating/cooling
- Qualifications
- Occupation
- Workplace
- Household tenure type
- Country of birth
- Religion
- Ancestry

E.7.2 Scenario: Combined public, visible and commercial sources.

This is essentially the combination of nosy neighbour with publicly available information together with a superset of commercially available data. This implies a very well-resourced attacker who is carrying out a deep information gathering exercise on a small targeted population. Note the list of variables is more extensive than might be obtained on any restricted access database.

- Home address
- Age
- Sex
- Marital status
- Number of cars
- Number of dependent children

- Number of elderly persons
- Density (persons/rooms)
- Cultural and ethnic group
- Family type
- Accommodation type
- Multi-ethnic household
- Number of residents
- Number of rooms
- Presence of central heating/cooling
- Occupation
- Workplace
- Household tenure type
- Country of birth
- Religion
- Ancestry
- Number of cars
- Accommodation type
- Employment status
- Socio-economic index
- Household composition
- Income
- Transport to work
- Highest qualification
- Long term limiting illness

Appendix F Collusive attacks

Collusive attacks are ones where the data subjects collude in providing information about themselves. These do not intrinsically constitute a set against which a data custodian is legally bound to protect. However, a successful collusive attack could still carry some risk, for example in terms of reputational damage.

F.1 Demonstrative political attack

F.1.1 Scenario: Demonstrative political attack: restricted set

The assumption underlying this scenario is that a political group, such as an anti- government group, acts in collusion with a data subject for the purpose of embarrassing the Government by undermining its data collection/release activities. Imagine that the data subject provides the group with copies of the information they gave to the interviewers. This scenario could happen in a census, which is a major public investment. Here the data collection process is familiar to everyone, and colluding respondents could be prepared in advance, and be absolute to be in the collected data (and also in the outputs with a relatively high probability). In principle, a larger number of variables could be used, but in the restricted variant, we have avoided those that are difficult to code (such as occupation), on the assumption that the political organisation will attempt to minimise divergence to prevent the demonstration backfiring. We have also avoided those that give information about other individuals apart from the colluding agent, on the assumption that the use of such variables would go against the underlying rationale for the attack.

- Home address
- Age
- Sex
- Education
- Marital status
- Employment status
- Cultural and ethnic group
- Religion
- Country of birth
- Migration in the last year
- Household tenure type
- Long term limiting illness
- Self-reported health
- Income

Attacker Profile: Person or organisation with specific desire to cause political impact on the government.

F.1.2 Scenario: Demonstrative political attack: extended set

- Home address
- Age
- Sex
- Marital status
- Employment status
- Cultural and ethnic group
- Religion
- Country of birth
- Migration in the last year
- Long term limiting illness
- Self-reported health
- Income

- Number of rooms
- Housing type
- Number of children
- Household tenure type
- Number of residents

Attacker Profile: Person or organisation with specific desire to cause political impact on the government.

Part III Information useful for implementation

Appendix G Instructions for calculating the number of uniques in a file

These instructions assume that you have downloaded the relevant data from the UKAN website (either Basetton.xlsx for Excel or Basetton.sav for SPSS) and have it open in the relevant software. They also assume that you have a basic familiarity with the software package. The file is synthetic data but the data structure is that which might typically be found in a UK census, survey or administrative file.

In both cases we are using an eight variable key which represents information that somebody might plausibly know about a neighbour. You can play about with different variable combinations to see the impact on the number of uniques. Nothing should be read into the specific details of the results (the data is not real) – the exercises simply serve to demonstrate the technique which you can then use with your own data.

G.1 B.1 Instructions for Excel

4. Sort the file by the following columns (checking the 'my data has headers' box is checked): sex, age, ethnic, accomtype, tenure, marstatus, ncars, cenheat. For each column, sort from smallest to largest.
5. Enter the word 'ccount' into cell N1
6. Enter 1 in cell N2
7. Enter the following formula into cell N3

   =IF(AND(A3=A2,B3=B2,C3=C2,D3=D2,E3=E2,F3=F2,J3=J2,M3=M2),N2+1,1)
8. Fill down from N3 to N210745
9. Select and copy column N
10. Right click 'Paste' and pick the values option (ensuring the values are associated with the correct row as you carry out further sorting and calculations)
11. Repeat the sort you did at stage 1, but adding in ccount to the end of the list sorted from largest to smallest.
12. Enter the word 'csize' into cell O1
13. Enter the following formula into cell O2:

    =N2
14. Enter the following formula into cell O3:

    =IF (N3<N2,O2,N3)
15. Fill down from O3 to O210745

16. Switch to the output page tab
17. In cell B2 type the formula

    =COUNTIF (Barsetton!O:O,1)

G.2 Syntax for SPSS

```
SORT CASES BY sex(A) age(A) ethnic(A) accomtype(A) tenure(A) marstatus(A) ncars(A) cenheat(A).

COMPUTE eccount=1.
IF (sex=lag(sex) & age=lag(age) & ethnic=lag(ethnic) & accomtype=lag(accomtype) &
tenure=lag(tenure) & marstatus=lag(marstatus) & ncars=lag(ncars) & cenheat=lag(cenheat))
eccount=lag(eccount)+1.
EXECUTE.

SORT CASES BY sex(D) age(D) ethnic(D) accomtype(D) tenure(D) marstatus(D) ncars(D) cenheat(D)
eccount(D).
COMPUTE ecsize=eccount.
IF (eccount<lag(eccount)) ecsize=lag(ecsize). EXECUTE.

COMPUTE unique=0.
VARIABLE LABELS unique 'Is the case unique?'. VALUE LABELS unique 0 'No' 1 'Yes'.
IF (ecsize=1) unique=1. EXECUTE.

FREQUENCIES VARIABLES=unique
/ORDER=ANALYSIS.
```

Appendix H A Description of the Data Intrusion Simulation (DIS) Method

H.1 Introduction

The concept behind the DIS method derived from concerns expressed by Elliot (1996) regarding the need to examine statistical disclosure risk from the viewpoint of the data intruder (intruder-centrically) rather than from that of the data itself (data-centrically). A rational intruder would be indifferent to questions such as, for example, whether a record was sample or population unique, because they will know such attributions of status are unreliable and more importantly because they will have more pragmatic concerns, such as whether any matches are correct. The DIS method simulates the intruder perspective by focusing on the probability of a unique match being correct. The basic assumption is that the intruder has some information about a population unit and uses that information to attempt to find the record for that individual in a microdata file (which is a sample of the relevant population). If there is only one record in the dataset which corresponds to the information that the intruder has that is called a unique match. If that record is the correct record for that population unit that is called a correct match. These basic elements form the headline statistic of a DIS analysis; the probability of a correct match given a unique match, or Pr(cm|um).

The basic principle of the DIS method is to remove records from the target microdata file and then re-sample them according to the original sampling fraction (the proportion of the population that are in the sample). This creates two files, a new, slightly truncated, target file and a file of the removed records which can then be matched against the target file. The method has two computational forms, the special form, where the sampling is actually done, and the general form, where the sampling is not actually performed, but its effect is derived using the equivalence class structure and sampling fraction.

H.2 The special method

The special DIS method uses a similar technique to Briggs (1992).

18. Set counters U and C to zero.
19. Take a sample microdata file (A) with sampling fraction S.
20. Remove a random record (R) from A, to make a new file (A').
21. Generate a random number (N) between 0 and 1. If N<=S then copy back R into A' with each record having a probability of being copied back equal to S.
22. The result of this procedure is that B will now represent an arbitrary population unit whose probability of being in A’ is equal to the original sampling fraction.

23. Match fragment against A'. If R matches a single record in S' then add record 1 to U if the match is correct add 1 to C.

24. Iterate through stages ii-v until C/U stabilises.

H.3 The general method

A more general method can be derived from the above procedure. Imagine that the removed fragment (B) is just a single record. Clearly there are six possible outcomes depending on whether the record is resampled or not and whether it was a unique, in a pair, or in a larger equivalence class.

Table 1: Possible per record outcomes from the DIS general method

Record is:	*Copied back*	*Not copied back*
Sample unique	**Correct unique match**	Non-match
One of a sample pair	Multiple match including correct	**False unique match**
One of a larger equivalence class	Multiple match including correct	False multiple match

Given this, one can derive the estimated probability of a correct match given a unique match from:

$$\frac{U\ x\ \Pi}{(U\ x\ \Pi) + P(1-\Pi)}$$

where *U* is the number of sample uniques, *P* is the number of records in pairs and *Π* is the sampling fraction.

For full statistical proof of the above theory see Skinner and Elliot (2002). For a description of an empirical study that demonstrates that the method works see Elliot (2000). For an elaboration using the special method for post-perturbation disclosure risk assessment see Elliot (2001). For an extension which takes account of general misclassification errors see Elamir and Skinner (2006)

Appendix I Instructions for calculating the DIS score

These instructions assume that you have downloaded the relevant data from the UKAN website (either Basetton sample.xlsx for Excel or Basetton sample.sav for SPSS) and have it open in the relevant software. The file is synthetic data but the data structure is that which might typically be found in a UK census, survey or administrative file.

In both cases we are using an eight variable key which represents information that somebody might plausibly know about a neighbour. Nothing should be read into the details of the results (the data is not real) – the exercises simply serve to demonstrate the technique which you can then use with your own data.

In both cases we are using a file where the sampling faction is 10%.

I.1 Instructions for Excel

1. Sort the file by the following columns (checking the 'my data has headers' box is checked): sex, age, ethnic, accomtype, tenure, marstatus, ncars, cenheat. For each column, sort from smallest to largest.
2. Enter the word 'ccount' into cell N1
3. Enter 1 in cell N2
4. Enter the following formula into cell N3

   =IF(AND(A3=A2,B3=B2,C3=C2,D3=D2,E3=E2,F3=F2,J3=J2,M3=M2),N2+1,1)
5. Fill down from N3 to N210745
6. Select and copy column N
7. Right click 'Paste' and pick the values option (ensuring the values are associated with the correct row as you carry out further sorting and calculations)
8. Repeat the sort you did at stage 1, but adding ccount to the end of the list sorted from largest to smallest.
9. Enter the word 'csize' into cell O1
10. Enter the following formula into cell O2:

    =N2
11. Enter the following formula into cell O3:

    =IF(N3<N2,O2,N3)
12. Fill down from O3 to O210745
13. Switch to the output page tab
14. In cell B2 type the formula

=COUNTIF(BarsettonSample!O:O,1)

15. In cell B3 type the formula

=COUNTIF(BarsettonSample!O:O,2)

16. Enter the sampling fraction 0.1 into cell B4
17. Enter the following formula into Cell B5

=B2*B4/(B2*B4+B3*(1-B4))

I.2 Instructions for SPSS

The syntax to use is shown below. When you have run it you will have a frequency table which will give you counts for the number of unique records and the number which are members of identical pairs. You simply need to insert those numbers into the standard DIS formula:

$$\Pr(cm|um) = \frac{U\ x\ \Pi}{(U\ x\ \Pi) + P(1-\Pi)}$$

where U is the number of sample uniques, P is the number of records in pairs and Π is the sampling fraction, in this case 0.1.

SPSS syntax

```
SORT CASES BY sex(A) age(A) ethnic(A) accomtype(A) tenure(A) marstatus(A) ncars(A) cenheat(A).

COMPUTE eccount=1.
IF (sex=lag(sex) & age=lag(age) & ethnic=lag(ethnic) & accomtype=lag(accomtype) &
tenure=lag(tenure) & marstatus=lag(marstatus) & ncars=lag(ncars) & cenheat=lag(cenheat))
eccount=lag(eccount)+1.
EXECUTE.

SORT CASES BY sex(D) age(D) ethnic(D) accomtype(D) tenure(D) marstatus(D) ncars(D) cenheat(D)
eccount(D).

COMPUTE ecsize=eccount.
IF (eccount<lag(eccount)) ecsize=lag(ecsize). EXECUTE.

COMPUTE uniquepair=0.
VARIABLE LABELS uniquepair 'Is the case unique or a one of a pair?'. VALUE LABELS uniquepair 0
'Not unique or pair' 1 'Unique' 2 'One of a pair'. IF (ecsize=1) uniquepair=1.
IF (ecsize=2) uniquepair=2. EXECUTE.

FREQUENCIES VARIABLES=uniquepair.
/ORDER=ANALYSIS.
```

Appendix J Data Features Template

Feature type	Question	Answer/Actions
Data subjects	Who are they?	
	What is their relationship with the data?	
Data type	Microdata, aggregate data, or something else?	
Variable types	What common key variables are there?	
	Which variables are sensitive?	
Data properties	Is the data accurate?	
	How old is the data?	
	Is the data hierarchical or flat?	
	Is the data longitudinal or cross-sectional?	
	Is it population data, or a sample?	
	If a sample, what is the sampling fraction?	
Is there anything else of note?		

References

DOBRA, A. & FIENBERG, S. E. (2000) Bounds for cell entries in contingency tables given marginal totals and decomposable graphs; in Proceedings of the National Academy of Sciences, 97(22): 11885-11892, available at: http://tinyurl.com/BNDS-DECOM [accessed 30/5/16].

DOBRA, A. & FIENBERG, S. E. (2001) Bounds for cell entries in contingency tables induced by fixed marginal totals; Statistical Journal of the United Nations Economic Commission for Europe, 18(4): 363-371, available at: http://tinyurl.com/BNDS-MARGINAL [accessed 30/5/16].

DOMINGO-FERRER, J. & TORRA, V. (2008) A critique of k-anonymity and some of its enhancements; In 3rd Intl. Conference on Availability, Reliability and Security (ARES 2008), Los Alamitos CA: IEEE Computer Society, 2008: 990-993, DOI: 10.1109/ARES.2008.97.

DOMINGO-FERRER, J, SÁNCHEZ, D., & SORIA-COMAS, J. (2016) Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter- model Connections; Synthesis Lectures on Information Security, Privacy, & Trust 15: Morgan & Claypool, DOI: 10.2200/S00690ED1V01Y201512SPT015.

DUNCAN, G. T., ELLIOT, M. J., & SALAZAR-GONZÁLEZ, J. J. (2011) Statistical Confidentiality. New York: Springer.

DWORK, C. & ROTH, A. (2014) The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science 9(3-4), 211-407.

ELAMIR, E.A. & SKINNER, C. (2006) Record level measures of disclosure risk for survey microdata; Journal of Official Statistics, 22(3): 525, available at: http://tinyurl.com/REC-RISK [accessed 30/5/16].

ELLIOT, M. J. (1996) Attacks on Confidentiality Using the Samples of Anonymised Records; In Proceedings of the Third International Seminar on Statistical Confidentiality. Bled, Slovenia, October 1996. Ljubljana: Statistics Slovenia- Eurostat.

ELLIOT, M. J. (2000) DIS: A new approach to the measurement of statistical disclosure risk; Risk Management, 2: 39-48, DOI:10.1057/palgrave.rm.8240067.

ELLIOT, M. J. (2001) Advances in data intrusion simulation: A vision for the future of data release; Statistical Journal of the United Nations Economic Commission for Europe, 18(4): 383-391.

ELLIOT, M. J. & DALE A. (1998) Disclosure risk for microdata; report to the European Union ESP/204 62 (1998): 361-372.

ELLIOT, M. J. & DALE, A. (1999) Scenarios of Attack: The Data Intruder's Perspective on Statistical Disclosure Risk; Netherlands Official Statistics, Spring 1999: 6-10, available at: http://tinyurl.com/ATTACK-SCENARIO [accessed 30/5/16].

ELLIOT, M. J., DIBBEN, C., GOWANS, H., MACKEY, E., LIGHTFOOT, D., O'HARA, K., & PURDAM, K. (2015) Functional Anonymisation: The crucial role of the data environment in determining the classification of data as (non-) personal; CMIST work paper 2015-2 available at http://tinyurl.com/FUNC- ANON [accessed 27/5/2016].

ELLIOT, M. J., MACKEY, E., O'HARA, K. & TUDOR, C. (2016a) The Anonymisation Decision-Making Framework. UKAN Publications.

ELLIOT, M., J., MACKEY, E. & PURDAM, K. (2011) Formalizing the Selection of Key Variables in Disclosure Risk Assessment, 58th Congress of the International Statistical Institute, Aug 2011, Dublin, Ireland.

FIENBERG, S.E., RINALDO, A., YANG, X. (2010) Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables. In: International Conference on Privacy in Statistical Databases, pp. 187{199. Springer.

HUNDEPOOL, A., DOMINGO-FERRER, J., FRANCONI, L., GIESSING, S., NORDHOLT, E. S., SPICER, K. & DE WOLF, P. P. (2012) Statistical Disclosure Control. London: John Wiley & Sons.

MACKEY, E. (2009) A framework for Understanding Statistical Disclosure Control processes; PhD Thesis, The University of Manchester. Manchester: University of Manchester.

MACKEY, E. & ELLIOT, M. J. (2013) Understanding the Data Environment; XRDS: Crossroads, 20 (1): 37-39.

OFFICE OF THE AUSTRALIAN INFORMATION COMMISSIONER (OAIC 2014a) Data breach notification – A guide to handling personal information security breaches, available at www.oaic.gov.au.

OHM, P. (2010) Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization; UCLA Law Review, 57: 1701-1777, available at: http://www.uclalawreview.org/pdf/57-6-3.pdf [accessed 30/5/2016].

O'KEEFE, C.M., GOULD P. & CHURCHES, T. (2014) Comparison of two remote access systems recently developed and implemented in Australia; In Domingo-Ferrer J. (Ed.), Privacy in Statistical Databases 2014, LNCS 8744, pp 299-311.

O'KEEFE, C.M., WESTCOTT, M., O'SULLIVAN, M., ICKOWICZ, I., & CHURCHES, T. (2016) Anonymization for outputs of population health and health services research conducted via an online data center, Journal of the American Medical Informatics Association 2016; doi: 10.1093/jamia/ocw152.

PURDAM, K. & ELLIOT, M. J. (2007) A case study of the impact of statistical disclosure control on data quality in the individual UK Samples of Anonymised Records; Environment and Planning A, 39(5): 1101-1118, DOI: 10.1068/a38335.

RITCHIE, F. Access to Sensitive Data: Satisfying Objectives Rather than Constraints, Journal of Official Statistics, Vol. 30, No. 3, 2014, pp. 533–545, http://dx.doi.org/10.2478/JOS-2014-0033

SAMARATI, P. (2001) Protecting respondents' identities in microdata release; IEEE Transactions on Knowledge and Data Engineering, 13(6): 1010–1027.

SAMARATI, P. & SWEENEY, L. (1998) Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Washington: SRI International, available at: http://tinyurl.com/sam-swe-kanon [accessed 28/5/16].

SKINNER, C. J. & ELLIOT, M. J. (2002) A measure of disclosure risk for microdata; Journal of the Royal Statistical Society: series B (statistical methodology), 64(4): 855- 867, DOI: 10.1111/1467-9868.00365.

SMITH, D. & ELLIOT, M. (2008) A Measure of Disclosure Risk for Tables of Counts; Transactions on Data Privacy, 1(1): 34-52, available at: http://www.tdp.cat/issues/tdp.a003a08.pdf [accessed 30/5/16].

THOMPSON, G., BROADFOOT, S. & ELAZAR, D. (2013) Methodology for the Automatic Confidentialisation of Statistical Outputs from Remote Servers at the Australian Bureau of Statistics; In Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Ottawa, Canada, 28–30 October 2013, available at: bit.ly/1PTippv [accessed 17/2/16].

VAN DEN HOVEN, J., HELBING, D., PEDRESCHI, D., DOMINGO-FERRER, J.,

CONTACT US

t 1300 363 400
+61 3 9545 2176
e csiroenquiries@csiro.au
w www.data61.csiro.au

FOR FURTHER INFORMATION

Dr Christine M O'Keefe PhD MBA
t +61 2 6216 7021
e Christine.O'Keefe@csiro.au
w www.data61.csiro.au

AT CSIRO WE SHAPE THE FUTURE

We do this by using science and technology to solve real issues. Our research makes a difference to industry, people and the planet.