ADF COMPANION DOCUMENT F: Building disclosure scenarios

A key component of a well-formed SDC exercise is the development of disclosure scenarios to ground risk analysis, specifying the risks semi-formally. 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.

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 do not suppose we will get many naysayers if we assert a belief 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 widows 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 the UK, or at least in my area?

Now suppose that I am using a de-identified dataset 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 then 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, overconfidence effects and other forms of cognitive bias. So claims to have found someone can easily be misjudgements. Let us look at the example a little more objectively. In the 2011 census there were seven 16-year-old widowers in the UK. 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 any sort of geographical indicator then the probability of the data actually singling out my neighbour would be quite high. So, theoretically, the risk of inadvertent and accurate recognition is non-zero.

However, bear in mind here that the presumption of this scenario is that the recognition is inadvertent, and the lack of any prior motivation substantially reduces the privacy risk for two reasons.

Firstly, the example is not so much of me finding a needle in a haystack but just happening to sit on one. The dataset has to be configured in such a way that the unusual combination of characteristics that my neighbour has appears simultaneously in my software window as I am browsing the data. For a large dataset the likelihood of such an event will be pretty low.

Secondly, having recognised my neighbour, what am I going to do? If I decide to act on my discovery then this is no longer simply a case of spontaneous recognition but a particular type of deliberate attack called fishing. If on the other hand I do nothing then this might be a 'so what?' situation, in which no harm befalls my neighbour, with minimal privacy impact. The meaning of the recognition will partly depend on what the dataset is about; if it is a dataset of criminals or sufferers from sexually-transmitted diseases then simply being in the data is sensitive and me finding my neighbour in there might matter a lot. On the other hand, if it is a random sample of some country's population then maybe spontaneous recognition matters less.

Other factors which will indicate whether one need be concerned with spontaneous recognition are the size of the dataset, whether the user has response knowledge and who the users are.

Dataset size can have a counterintuitive effect. A smaller dataset effectively decreases the size of the haystack so it increases the likelihood of coming across someone (if they are in there).

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 more likely to assume that it is you if I do so.

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 on student data) 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.

Deliberate attacks and the data intruder

In SDC, the agent who attacks the data is usually referred to as the *data intruder*.[1] 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. We have further developed this system 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/mismatches against reality. How will that affect the attack?

[1] Other terms that are used are 'the attacker', 'the data snooper' and 'the adversary'. These are synonymous.

INTERMEDIATE OUTPUTS (to be used in the risk analysis)

- ***Attack Type:*** What is the technical aspect of 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:***[2] 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 essential for the intruder to achieve re-identification and allow 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 and which provide a 'hook' into the target dataset, allowing individuals to be matched. See Figure 2.1 for a schematic view of how this works. Ideally, from the intruder's point of view, the coding method of a key variable must be the same on both the attack and target datasets, or the two must at least be harmonisable.

Essentially, there are four sources of auxiliary information: (i) datasets containing the same information for the same (or sufficiently similar) population, (ii) information that is publicly available (e.g. in public registers or on social media), (iii) information obtained from local knowledge (e.g. house details obtained via an estate agent or by

[2] Recall that a data situation concerns the relationship between some data and their environment. We discuss this in more detail below.

physical observation), and (iv) information obtained through personal knowledge (e.g. things I know about my neighbours or work colleagues).

There is obviously a terminological overlap between the notion of a key variable and that of an 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 critical – 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.[3]

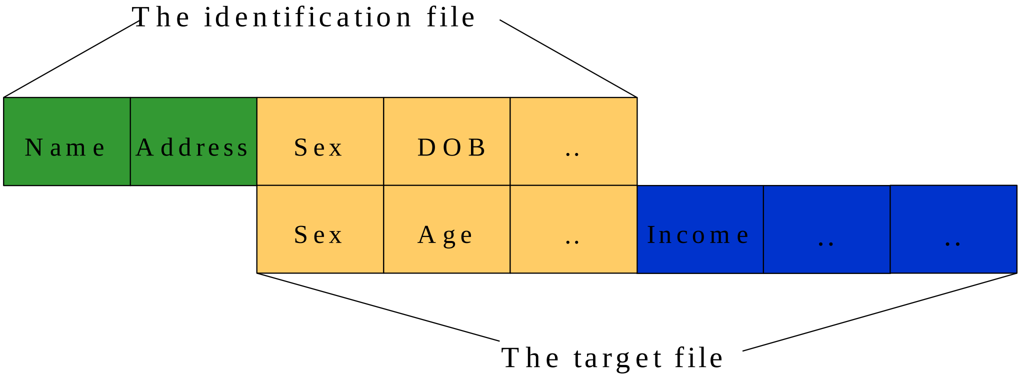

Figure 2.1: An illustration of the key variable matching process leading to disclosure. From Duncan et al (2011).

Data divergence

Another critical 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 disseminated and characteristics will have changed since the data were generated. This is true of the target dataset as well as the auxiliary information

[3] We note that the k-anonymity literature uses the term *quasi-identifiers* to refer to both key variables and indirect identifiers which in our experience does sometimes lead to some confused thinking by practitioners; so the terminological separation is not just a matter of semantics.

held by an intruder. The combination of these will introduce the possibility of error into any linkage.

Collectively, we refer to these sources of ‘noise’ in the data as *data divergence*. The term refers to two situation types (i) *data-data divergence*, or differences between datasets, and (ii) *data-world divergence*, differences between datasets and the world. In general, both types 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, 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 the 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 intruder tests as well as data analytical risk assessments.

Notwithstanding the above remarks, a paradox of data divergence is that it is not a reliable protection of confidentiality. Firstly, for any particular individual-record pair there may be no divergence at all. Secondly, analysts are getting increasingly sophisticated in dealing with linkage in the face of ‘fuzziness’ – when we talk through the process of doing a penetration test in chapter 3 you will see that we 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

At its simplest level the issue of response knowledge can be captured by a single question: ‘Do you know I am in the data?’ If the answer to that is ‘yes’ then you are said to have response knowledge of me in respect of those data.[4] In that case, one key

[4] Of course you might be wrong – perhaps the information which tells you I am in the data is out of data 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

element of uncertainty, whether the person is 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).[5]

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 anonymised dataset of the members of the Bognor Regis Bicycle Club. Straightforwardly, I could know that my target is in the club and therefore in the dataset. That is clear cut response knowledge but I could have other information about you which falls short of full response knowledge but is nevertheless informative. I could know that you live or work in Bognor Regis 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 Bognor Regis Bicycle Club population and that in turn increases the effective sample fraction.[6] As we will see in chapter 3 the sample fraction is an important element of the risk.

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.

[5] 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 an anonymised 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 anonymisation problem.

[6] 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.