Chapter 2
Concepts of Statistical Disclosure Limitation

The SDL literature has its own terminology. Understanding this terminology and, more importantly, the concepts underlying the terminology is essential to learning how statistical confidentiality can be best employed. In this chapter we look at the structure of disclosure risk, its assessment, and its limitation. Complicating our task is that many terms, such as "protecting data" or "sensitive data," have no universally accepted meaning. Driven by variations in their historical and legal environment, DSOs exhibit differences in how they use terminology. This can lead to confusion in discussions among DSOs and indeed within the SDL research community as well. In this chapter we lay out widely accepted concepts and a terminology that provides a common framework intended to minimize confusion. These concepts and terminology are used consistently throughout the rest of this book.

2.1 Conceptual Models of Disclosure Risk

DSOs develop confidentiality policies and procedures in order to address the concern that the data products they disseminate may disclose the identity of a respondent or information about the characteristics of a particular respondent. For a DSO to assess disclosure risk, it must understand what is being risked and what are the consequences of a disclosure.

DSOs need to preserve the trust accorded to them by respondents. Assurances of confidentiality appear at the top of any responsible survey questionnaire or are provided at the start of an interview. Because of these confidentiality statements, the need to assure data protection becomes explicitly part of the data processes from collection through to dissemination. There is a social contract between the DSO and the respondent that the respondent's information will be looked after and any data dissemination arising from those responses will be carried out responsibly.

DSOs often treat disclosure as if it would be a catastrophe, doing major damage to their mission. Many DSOs, especially National Statistical Offices, believe that if a disclosure happens it will have consequences far beyond the informational impact of the disclosure for the individual for whom disclosive attributions have been made. They believe that a single publicized disclosure could lead to such a breakdown in trust that there will be a significant reduction in respondent cooperation. A

G.T. Duncan et al., *Statistical Confidentiality*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-1-4419-7802-8_2,

breakdown of trust leading to non-cooperation is a sociological construct that is supported by historical evidence. In a review of the census experience of 11 western European countries, McDonald (1984) found significant concerns for privacy and confidentiality in England, the Federal Republic of Germany, Italy, the Netherlands, Norway, and Sweden. In particular, the German census originally scheduled for 1981 was postponed for budgetary reasons, but then in 1983 there were large-scale public protests and over a thousand lawsuits were filed against the proposed census. Many reasons could be advanced for this public agitation including worries about increased use of microdata, such as the specter of advanced computer technology, and lack of support for the value of the census by the media and government officials. In addition to these possibilities, Butz (1985) posited the reason of fears that government agencies would use census data against individuals. Presumably, these fears were particularly salient in Germany given the abuses of census data under the Nazi regime. In March 1983, the Constitutional Court suspended the census. The Court argued that the planned dissemination procedures lacked adequate confidentiality protection (Butz and Scarr, 1987). The census was not held until 1987.

However, the first step in the sociological reasoning that a single publicized breach would lead to a reduction of trust is less justified. It is by no means clear that a reduction in trust necessarily follows from a single breach nor that it is unmediated by the DSO's reaction to that breach (Mackey and Elliot, 2009). The paradox is that there are only isolated cases publicly known of a statistical disclosure event having happened. Therefore, we have no real evidence either for or against the proposition. In the face of such lack of evidence, DSOs feel obliged to maintain the view that such events would be near catastrophic and therefore view only minimal disclosure risk as acceptable.

To further complicate this already fuzzy picture, it is clear that zero (or even effectively zero) disclosure risk is not compatible with a DSO meeting its obligation to disseminate data that have high utility. In their actual behavior DSOs must therefore be more pragmatic. In the United Kingdom, for example, the policy of the Office for National Statistics (ONS) is that no disclosure takes place if more than a "reasonable amount of effort" on the part of a data snooper would be required to break confidentiality. Underlying this pragmatism is the interplay between perceived and objective risk. If the risk is perceived as negligible then would-be snoopers are likely to look for other ways to achieve their goals. Importantly therefore a reduction in a data snooper's perception of disclosure risk is *de facto* a reduction in the objective risk.[1]

Marsh et al. (1991) used conditional probabilities to formalize this interplay as follows:

$$P(\text{identification}) = P(\text{identification}|\text{attempt}) \cdot P(\text{attempt})$$
$$P(\text{identification}|\text{no attempt}) = 0$$

[1] This is analogous to a fake burglar alarm box.

We can limit the actual risk by controlling either or both the probability of an attempt being made as well as the probability that it is successful given that it has been made. Undoubtedly, DSOs engage in both forms of control, sometimes simultaneously. For example, the United Kingdom's ONS used a record swapping algorithm for its 2001 census outputs. A proportion of records between 0 and 5% had their local geographical codes swapped. ONS announced this range for the swap rate, thus having some impact on the disclosure risk. Within the announced range the swap rate could be zero, so the impact on objective risk would also be zero. But whatever the actual swap rate—provided it is keep secret, a reduction in risk is achieved beyond the objective impact of the actual perturbation. The reason for this is the increase in the uncertainty for any inferences that a data snooper might make. This shows that effective public communication is as important in disclosure risk management as is manipulation of data before dissemination.

2.1.1 Elements of the Disclosure Risk Problem

The usual reasoning about disclosure risk is this: a data snooper who is in possession of some information about one or more identified population units wants to infer further information about those population units. The data snooper seeks to do this by linking that information to items in the target data set. The main alternative to this line of reasoning is the spontaneous recognition of a data element by a bona fide user of the data.

This basic formulation has different meanings depending on whether we are considering aggregate data or microdata. Chapter 4 provides a detailed development for aggregate data; Chapter 5 does the same for microdata. Here we introduce the basic concepts for each.

2.1.1.1 Microdata

Microdata are the raw material at DSOs. They are the direct information collected from the respondents. Traditionally microdata are organized in a database, where each record is an encoding of the answers of each contributor (person, household, organization, etc.) to a survey or census. In principle, though, microdata could be collected in a variety of ways. For example, economic microdata at a country-level could be generated from a study of national accounts; hospital in-patient data could be collected from hospital administration records; and data on how the response rate and accuracy of a census might change with alterations in the number of questions presented could be collected through an experiment. The crucial point about microdata is that for each population unit represented in the data set there is a record of the values of multiple attributes.

A microdata set that contains all the variables collected for the whole population is called the *full table*, acknowledging that microdata are transformable into frequency tables. Invariably, the DSO stops short of releasing the full table. Instead, it

may release subsets that are marginal tables of the full table or possibly samples of microdata (or indeed it may have only collected a sample as in a social survey). If it does release a sample of microdata then it will usually reduce the number of variables that it releases as well as the level of detail on those variables it does release (particularly temporal and geographical detail). Limiting the data to be released is typically the initial phase of statistical disclosure limitation. Before looking at SDL, we examine the event we are trying to limit—"disclosure."

In the case of microdata there are two different concepts related to a disclosure: *identification* and *attribution*. Identification is the association of a known population unit with a particular microdata record. Attribution is the association of information in a microdata set with a particular population unit (see Lambert, 1993; Reiter, 2005b).

Operationally, the data snooper must usually identify a population unit in order to make attributions about it. It is the attribution itself which forms the disclosure. However, there are cases where identification and attribution occur independently. Identification may occur without attribution if the data snooper already knows all the information contained within the microdata set about the population unit in question, so no new information was attributed to the population unit from the linkage (an obvious example of this being self-identification). Conversely, where two or more population units with a given set of variable values necessarily share another attribute, the attribution of that additional information to each population unit for which the key information is known can take place even without direct identification. This could arise, for example, where a data set contained occupational and income information and it was known that all individuals with a given occupation had a particular income. Notwithstanding these two cases, the canonical process is identification leading automatically to attribution.

2.1.1.2 Deliberate Linkage

Deliberate record linkage (or *matching)* is the typical mode of disclosure. The presupposition is that a data snooper has access to a data set which contains direct identifiers for population units (name, address, etc.) and a set of *key variables* which are also present in the target data set. The key variables are then used to link the identifiers to the data snooper's target, as shown in Fig. 2.1.

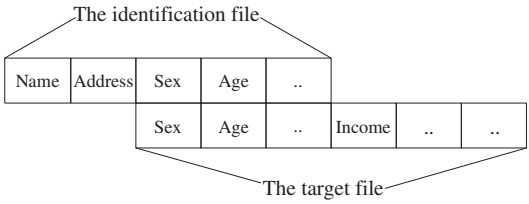

Fig. 2.1 An illustration of key variable matching leading to disclosure (after Elliot, 2000)

2.1.1.3 Aggregate Data

Aggregate data are most commonly published as tables, often tables of counts (frequency tables), but also cross-classified tables of statistics such as means and sums. Tables of counts are aggregate data obtained from the microdata. Characteristic of a statistical table is that it also contains marginal sums, so that in addition to the data values, we also have relations linking these values, as for a sex variable a marginal total may be the sum of the number of males and the number of females. These links between values add complexity to the problem of protecting tables which does not exist when protecting microdata.

With tabular data there are two conceptual processes underpinning the notion of disclosure—*subtraction* and *attribution.* Aggregate data are often data for the whole of a population (where this is not the case the aggregate data are effectively sample microdata for our purposes). Attribution with aggregate data is similar conceptually to that with microdata; that is, it is the association of information in an aggregate data release with a particular population unit. Importantly however, whereas with microdata, attribution follows automatically from correct identification, with aggregate data, attribution follows contingently from the presence of non-structural zeroes anywhere in the aggregate data set.

2.1.1.4 Attribution and Subtractive Attack

Subtraction is the removal from an aggregate data set (which could be a single table or set of tables) of particular individuals for which the values or all the variables within the aggregate data set are known to the data snooper. To clarify attribution and subtractive attack for tabular data consider Table 2.1. Take the population represented in this table to be everyone at a workshop you are attending. Over coffee, you overhear someone saying that they earned over 1 million pounds in the last quarter. Given that you know the data in Table 2.1, you can infer that person is a lawyer. This is positive attribution—the association of a particular value with a particular population unit. Conversely, if you talk to some people and find out that they are academics, you 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.

Table 2.1 Table of counts of income levels for two occupations from hypothetical population

	High	Medium	Low	Total
Academics	0	100	50	150
Lawyers	100	50	5	155
Total	100	150	55	305

Note that association and disassociation are different forms of the same process, attribution arising from zeroes in the data set. The presence of a non-structural zero in the internal cells of a table is potentially disclosive. In effect, the attributive process is disassociative.

Now consider Table 2.2. The population in this table differs in one respect—there is one highly paid academic. Give this table, we can no longer make the inferences that we could from Table 2.1 (at least not with certainty). However, what about myself? I am a member of the population represented and I know my own occupation and income. Suppose that I am a highly paid academic. Given this extra-table information, I can subtract 1 from the high-income cross academic cell in the table, which then reverts to Table 2.1. I am then back to the situation where I can make disclosive inferences from overhearing partial information about particular individuals.

Table 2.2 Table of counts of income levels for two occupations from hypothetical population

	High	Medium	Low	Total
Academics	1	100	50	151
Lawyers	100	50	5	155
Total	101	150	55	305

We can extrapolate this further. Consider a situation where you have complete information (for the two variables contained in Table 2.2) about multiple individuals. In effect such information represents a table of counts of the subpopulation of the individuals for whom I have complete knowledge. On the assumption that identification is available for both that subpopulation and for any additional information you gain through overheard conversations (or other sources), you can subtract the whole of that table from the population table before proceeding. In principle, this could lead to more zeroes appearing in the residual table. This is illustrated in Tables 2.3 and 2.4. 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 high disclosure risk, in that,

Table 2.3 Table of counts of income levels for two occupations for a subpopulation of Table 2.2 about which a hypothetical snooper has complete knowledge

	High	Medium	Low	Total
Academics	0	80	25	105
Lawyers	70	20	5	95
Total	70	100	30	200

Table 2.4 Residual table of counts resulting from subtracting Table 2.3 from Table 2.4

	High	Medium	Low	Total
Academics	1	20	25	46
Lawyers	30	30	0	65
Total	31	50	30	111

with low cell counts, it will be easier to obtain sufficient information external to the aggregate table to enable subtraction to zero than with high cell counts.

Beyond the subtraction to zeroes there is another sense in which low cell counts constitute a risk. Consider again Table 2.2. Recall that without external information about the population represented in the table it is impossible to make inferences *with certainty* about an individual given partial information about that individual. However, imagine again that you overhear someone at the workshop boasting about their high income. Although not sure that this individual is a lawyer, you can assert it so with a high degree of confidence. From the table, the conditional probability that a randomly selected individual is a lawyer, given that they are higher earner, is greater than 0.99. Depending on my purposes in disclosing this information, this may be good enough to do so.

The principle of subtraction is also central to understanding attacks on magnitude tables. Consider Table 2.5 with hypothetical total consultancy earnings by statisticians in the University sector in a certain year.

Table 2.5 Hypothetical magnitude table

	Consultancy sales (Euros)
United Kingdom	700,000
Spain	1,000,000
France	2,000,000
Germany	380,000
Italy	700,000
Netherlands	460,000

Let us suppose that I am in charge of Consultancy sales at Camford University in the United Kingdom and know thereby that we earned a total of 550,000 Euros last year. I therefore know, by subtraction, that all of the other UK universities combined have earned 150,000 Euros. This illustrates the principle of dominance. As my university dominates the cell for the United Kingdom, I (or anyone else who has this information about my university) am able to make some limiting inferences about other contributants to the cell.

2.1.1.5 Linking Tables

A further problem that arises with aggregate data is that of table linkage. This occurs when two or more tables in a release of aggregate data have variables in common (they are said to overlap). In effect, all aggregate tables of counts are margins of the *full table* (the underlying microdata), and from which, given enough external information, a user of the aggregate data could reconstruct that full table. Even though this is unlikely, a more pragmatic concern is that a data snooper might combine tables by linking subtables to recover larger tables and these larger tables themselves may be disclosive, or more vulnerable to a subtractive attack. For example, consider Tables 2.6–2.8, all drawn from the same hypothetical population.

Tables 2.6–2.9 An illustration of how groups of tables might be disclosive

Table 2.6

	Var 1	
Var 2	A	B
C	3	9
D	2	2

Table 2.7

	Var 1	
	A	B
E	1	10
F	4	1

Table 2.8

	Var 2	
Var 3	C	D
E	8	3
F	4	1

Table 2.9

	Var 1 and 2			
Var 3	A,C	A,D	B,C	B,D
E	0	1	8	2
F	3	1	1	0

None of these three tables is themselves disclosive, although we note some low counts and therefore we consider the cells to be risky.

However, because these three tables overlap and form the margins of a three-way table, we can, through linear programming or other means, identify all of the possible three-way tables that correspond to the two-way margins. In this case there exists only a single feasible three-way table (Table 2.9) that has the two-way margins shown in Tables 2.6–2.8, and in that three-way table there are two cells with zero counts. Since this table contains zeroes, it is potentially disclosive and therefore so is the release of Tables 2.6–2.8, even though those tables contain no zeroes. For extensions of these ideas see Chowdhury et al. (1999).

The situation is complicated further when we consider table linkage in interaction with the subtraction problem discussed above. Even if a set of tables are not linkable to form a single feasible larger table, subtraction of individual population units for which complete information for the larger table is known might lead to a single feasible table. For example, adding an individual with characteristics AED to the data underlying Tables 2.6–2.8 leads to more than one feasible table, so removing that individual leads to the reduction of the number of feasible three-way tables to 1.

Clearly, risk assessment needs to take account of all of the above possibilities. In Chapters 3 and 4 we will discuss methods for doing so.

2.1.1.6 Hierarchical Tables

A table is said to be a "hierarchical table" when one of the variables is a hierarchical variable. This means that there exists a variable which assumes several values, each one decomposable into other values, and so on. An example of a hierarchical variable is "geographical location," which may represent countries, regions, counties,

local authorities, and so on. The existence of hierarchical variables in a table creates additional links (aggregations) that complicate protecting confidential data.

2.1.1.7 Linking Anonymized Data Sets

As well as the internal linkages that we have discussed, data from different data sets might be linked. In fact there is now a well-established field of research concerned with how to link anonymized data sets legitimately, especially to improve data quality or the ability to improve statistical inferences, and so on (see for example Tranmer et al., 2005). However, this kind of linkage also has an impact on the risk of disclosure; Smith and Elliot (2004, 2005) show that it is possible to improve the ability to link two distinct samples of microdata if an aggregate table of population counts for a subset of the microdata is introduced into the mix. So having multiple data sets available in the same data environment adds an extra layer of complexity onto assessing disclosure risk.

2.1.1.8 Spontaneous Recognition

The notion of spontaneous recognition is simple. You know of person X who has an unusual combination of attribute values. You are working on a data set and observe that a record within that data set also has those same attribute values. You infer that the record must be that of person X. In order to be truly spontaneous you must have no intent to identify. Otherwise this is just a specific form of deliberate linkage. Note that you could, of course, be wrong in this inference. Even if the record was unique, the data set may only be a subset (as in a sample survey) of the population and the record may not be unique in the population (Skinner and Holmes, 1992; Bethlehem et al., 1990).

2.1.2 Perceived and Actual Risk

So far we have focused on the "objective" components of the risk based on the data to be released and its relation to what a data snooper may or may not know. However, intrusion also has a subjective component which can have an impact on overall risk. Here are examples:

(1) The perceived sensitivity of the target data not only affects the motivation of a data snooper but also affects the likelihood of non-response.
(2) The perceived likelihood of success of an attack affects the motivation of data snoopers.
(3) The presence of low counts in aggregate data may make respondents believe that the data are risky, independently of whether they are or not.
(4) Common sense demographic knowledge may frame individual's perception of unusual combinations of characteristics in microdata (this is the essence of the spontaneous recognition situation).

The exact relationship between perceived and actual risk is complex. Clearly, with Example 2 above there is a feedback loop whereby the subjective and objective elements of risk are causally related. This is another reason why measuring disclosure risk in an absolute sense is difficult, if not impossible, and why SDL researchers tend to focus on the objective component. Nevertheless, pragmatically speaking, respondent co-operation depends directly on the subjective component and only via this on the objective component. The subjective component of disclosure risk is therefore important and DSOs should pay at least as much attention to managing that perception as to directly controlling the actual risk.

2.1.3 Scenarios of Disclosure

Once we understand that disclosure risk management is as much about psychology as about objective characteristics of data, we naturally turn to thinking about how a disclosure event might occur. Who are data snoopers? What are they trying to achieve by their snooping? Answering such questions is an important first step in risk management. Elliot (2000, 2005) and Elliot and Dale (1999) have produced a classification for the analysis of snooping attempts that enables the generation of *key variables* available to a data snooper and therefore should be considered in any risk assessment:

- Motivation
- Means
- Opportunity
- Attack types
- Key/matching variables
- Target variables
- Effect of data divergence
- Likelihood of Success
- Goals achievable by other means?
- Consequences of attempt
- Likelihood of attempt
- Effect of variations in database structure

Let us consider each of these items in turn.

2.1.3.1 Motivation

The motivation for a disclosure attempt comprises two elements:

Rationale: A description of the motivations of the data snooper (for example, to discredit the DSO).

Goal: A specification of the state of the world that the data snooper wishes to achieve; that is, an operational definition of what would be achieved by the disclosure attempt (for example, the release of a match into the public domain).

2.1.3.2 Means

There are three elements in how an attack would be made:

Skills: The statistical and computational skills that would be needed to select and apply an adequate matching technique and to interpret the results.

Knowledge: Any factual knowledge that assists the data snooper in their attempt. As well as information stored in databases, available knowledge might include knowledge about a particular locality (such as the prevailing housing type) and direct knowledge (e.g., by visiting an address you can establish the housing type or by talking to an individual you can establish their occupation).

Computational power: Enough computing power to perform the analyses required to achieve the goals of the data snooper. Because of readily available computer hardware and inexpensive mathematical software it makes sense to assume that all data snoopers can compute (for example) the maximum and the minimum values for any missing data in a table. Today a prudent DSO takes a worst-case attack using high-performance computing as the expected attack.

2.1.3.3 Opportunity

Without access to the data set the snooper lacks the opportunity to break confidentiality. Access may be limited because the target data sets might be distributed only to those who have signed licensing agreements which are legally binding and which prohibit the licensee from: (a) attempting to identify an individual or household in the data files and (b) passing the data to an unlicensed individual.

Snooper access to the target data set could come through several routes:

- Through an authorized data set user.
- In collusion with an authorized user. The collusion could be voluntary (either because the colluder shares the data snooper's goals or has some goal/reward, for example financial payment) or involuntary (if the colluder is threatened or blackmailed).
- Security of a computer containing a copy of the microdata is breached, either by the data snooper or by a hired hacker.
- A copy of the data set is stolen.

In many cases because of legal restrictions, all these access routes involve illegal activity. Therefore, in pursuit of their goals data snoopers must either accept the consequences of their illegal activity; or not release information into the public domain; or release information in a way that conceals the data snooper's identity.

Calculating the probability of such access is prohibitively daunting. However, given the large number of data users, the DSO prudently assumes that the probability of an opportunity is nearly 1. That is, if any individual or organization has decided to attack an anonymized data set, then they will be able to gain access

to it. Nevertheless, access to the data for unauthorized usage has potential legal consequences which are likely to lower the probability of an attempt actually being made.

2.1.3.4 Types of Attacks

An attack type is a method for achieving a class of data snooper goals. Five different types of attacks for microdata are identified below plus one for aggregate data.

- *Data set cross-match*. An outside data set with several fields which are identical to or recodable to target microdata record fields (key matching variables) is cross-matched with the target microdata set. The most likely goal for this type of attack would be the enhancement of an outside data set.
- *Match for a single specific individual.* The intention behind this type of attack is to enhance or verify information available about a target individual. Matching information could be from an outside data set—as a hypothetical example, the Inland Revenue in the United Kingdom may want to search the target microdata for income-related information of an individual suspected of tax evasion.
- *Match for single arbitrary individual.* Here the data snooper is not interested in information gathering but instead with being able to claim that identification has been achieved and information could be disclosed. The data snooper is not interested in the actual identity of the individual who has been identified. A journalist in search of a good story might follow this route.
- *Match to a specific group of individuals.* This type of attack is an alternative to Type 3 for certain scenarios and would have the same goal. A set of individuals are selected either because they are distinctive (e.g., they come from a minority ethnic group) or because matching information is available on them (e.g., they belong to an occupation with a register of all members).
- *Fishing.* Strictly, this is not a separate type of attack but a variation on any of the other types. Rather than starting with an individual or a set of individuals in the outside world and attempting to identify them in the target data set, the data snooper starts with the target data set and locates one or more individuals with distinctive characteristics and attempts to find them in the world. Fishing thus starts with a search through the records in a target microdata file. Typically the search seeks to identify records that have unusual combinations of characteristics. The concern is that through demographic knowledge or the analytical properties of the data set, a snooper might be able to identify which records are unique within the population. Having identified such records, the snooper then searches within the population for an individual with those characteristics.
- *Subtraction.* This is mainly a concern for tables of population counts. The principle is that a snooper has knowledge about some individuals within the population which is represented in a given population table. In the simplest case that knowledge comprises all the variables represented in the table. The snooper can then remove the individuals from the table, possibly leading to a table from which disclosive attributions can be made.

2.1.3.5 Key Variables

For all disclosure attempts, key variables are essential to identification. Key variables are those which are available to the data snooper and which are also in the target data set and can therefore allow individuals to be matched. Ideally, the coding of a key variable is identical on both the attack and target data sets (or a harmonized coding can be produced).

There are two sources for key variables: (i) formal data sets containing the same information for the same population and (ii) informal information obtained from local knowledge (e.g., house details obtained via a real estate agent) or personal knowledge (e.g., one's neighbors or fellow workers).

2.1.3.6 Target Variables

The data snooper wants the values of target variables. If the data snooper seeks to gain information, the contents of the target variables are directly relevant. In situations where the data snooper is motivated by the secondary consequences of an attack, identification may be sufficient and therefore the information content of target variables is of little importance. However, in many situations the perceived sensitivity of the information contained in the variable is crucial to the impact of the disclosure on the DSO. Given this, target variables have two central properties, *usefulness* and *sensitivity.*

> *Usefulness*—for a variable to be a target it must contain information which improves upon or verifies data already available to the data snooper.
>
> *Sensitivity*—the sensitivity of a target variable is governed by the perceived importance of the information disclosed for the individual concerned. One might generalize sensitivity as the prevailing perception in the population of the sensitivity of the disclosed information. However, an individual's context will frame the sensitivity of any piece of information. For example, the number of cars I own might be regarded as of low general sensitivity, but an individual who is trying to keep his or her wealth hidden, might regard this information to be of high sensitivity.[2]

2.1.3.7 Effect of Data Divergence

All data sets contain errors and inaccuracies. Respondents do not always supply correct data. There may also be errors in recording. Interviewers make mistakes in recording.

Coders transcribe incorrectly. Oftentimes data items are missing. Missing or inconsistent values may be imputed using methods with no guarantee of accuracy.

---

[2]There are several likely inputs to both personal and public sensitivity, such as: *perceived deviance* (for example, knowledge that someone does not have an inside toilet could be considered more sensitive than knowledge that they do, because being told of a household that it has an inside toilet does not yield much information against expectations), *social acceptability*, and *information dispersal* (how widely known such information is about the specific individual and about a hypothetical average individual).

Data available from censuses and surveys will be months and possibly years old by the time they are available for analysis outside the DSO. This means that individual and household characteristics will have changed since the date of data collection. An equivalent set of issues will hold for information held by a snooper. The combination of these will create errors in 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*—differences between data sets and (ii) *data–world divergence*—differences between data sets and the world. In general both types can be assumed to reduce the success rate of matching attempts. However, where two data sets diverge from the world in the same way, referred to as *parallel divergence*, then the probability of matching is unaffected. This would be the case, for example, if a respondent has lied consistently or when two data sets both have out-of-date data, but yet have identical values.

2.1.3.8 Likelihood of Success

Likelihood of success is *not* the same as the likelihood of achieving identification given an attack. Rather, it refers to the likelihood of the data snooper achieving their goal, which in some scenarios may not be identification per se (for example, a hypothetical journalist could get a "good story" without a fully verified match).

Goal Achievable by Other Means?

Can the data snooper achieve their goals by other means, which are easier to execute, legal, and/or have an equal or better likelihood of success? This is a crucial factor in determining the likelihood of an attempt being made.

Consequences of Attempt

Each scenario must also consider the likely consequences of an attempt. These consequences will be dependent on the goals of the data snooper and the success or failure of the attempt. Broadly, consequences can be divided into two groups (i) whether or not confidentiality is broken and (ii) the effect on public confidence.

(i) Whether or not confidentiality has been broken

Confidentiality is broken if at least one verified match has taken place, thereby identifying the record in the target data set. A snooper attempt may not break confidentiality, either because a match cannot be achieved with the specified set of variables, or a match cannot be verified. Alternatively, the data snooper may not intend to break confidentiality *per se* but rather to demonstrate that it could be done (as when the individual who is identified has colluded in the matching exercise).

(iia) Knowledge of the attempt to breach confidentiality is released into the public domain.

Whether or not an attempt is successful, information might be released into the public domain regarding the attempt. Release of such information in itself could be dangerous to the DSO because of the effect on public confidence. If the attempt is known to be successful then this impact would, almost certainly, be increased. If the attempt is unsuccessful or demonstrably unverified then the effect is potentially double-edged. The mere knowledge that an attempt has been made might, by bringing the issue to the attention of the public, have a damaging effect on the public's perception of disclosure risk. Also the fact that an attempt has been made indicates that someone believes a breach in confidentiality is possible, which may have an adverse effect on perceived risk. Against this, a DSO could present an unsuccessful attempt as an indication of the security of the confidentiality guarantee. This would depend on the public relations or "fire-fighting" policy of the DSO; see Mackey (2009) for an in-depth discussion of this.

(iib) Knowledge that a breach of confidentiality is possible is released into the public domain.

Here disclosure is demonstrated without a breach in confidentiality taking place (i.e., where the identified individual(s) have colluded in the disclosure exercise). Again the effect on public confidence would depend on the public relations or "fire-fighting" policy of the DSO.

(iic) Details of matched individuals have been released into the public domain.

This is the most damaging consequence in terms of public confidence and future co-operation with the DSO. Where details of matched individuals pass into the public domain, media may run a "personal story" which may magnify (in the public mind) the importance of the confidentiality break. This will be so even if the match is unverified.

Additional damage may be done if the information disclosed is sensitive or personally embarrassing to the matched individual.

Likelihood of Attempt

Given the key variables, likelihood of attempt is a critical output of the scenario analysis. Surely it is difficult in general to put a numerical probability to the likelihood of a data snooper attack. Many of the inputs into our classification are highly contextual. Also, in order to make any assessment at all we must assume some rationality on the part of the would-be snooper. However, it is usually possible to arrive at a conclusion/exclusion decision by making this assumption and then to review the input and process information in each possible scenario.

Effect of Variations in Data Set Structure

Variations in the target data set, for example in content, sample size, or geographical detail, may alter the likelihood or meaning of a particular scenario.

2.1.4 Data Environment Analysis

Scenario analysis as described in Section 2.1.3 is important in understanding disclosure risk. However, the Achilles' heel for DSOs in dealing with the practicalities of risk assessment is their lack of knowledge about what data snoopers know. Do they have access to data that are external to an intended release that could be used in compromising confidentiality? This in particular makes it difficult for the DSO to generate an informed list of key variables. Purdam et al. have developed a method called *data environment analysis* to help rectify this (Purdam and Elliot, 2002; Purdam et al., 2003b, a). In data environment analysis, forms used to collect information about individuals are decoded into metadata. Corresponding to that form, a record is created on the data environment metadata set which records all the information obtained from the form plus, where possible, an estimate of coverage. Coverage information is obtained from common sense estimates of service value, interviews, questionnaires, and public statements of data set holders. This data environment analysis enables input into scenario generation. Elliot et al. (2005) have outlined a design using grid technology to automate Data Environment Analysis as part of a larger system of automated statistical disclosure risk analysis.

2.2 Assessing the Risk

Having outlined the concepts underlying data snooping, we now consider more fully the practice of risk assessment, which is a core concern of a DSO. Much early research was data-centric, that is it focused on defining properties of a data release that were more or less risky. This leads to the concept of uniqueness.

2.2.1 Uniqueness

Underpinning much of the work on disclosure risk analysis, particularly for microdata, is the notion of *uniqueness*. A record is unique on a set of key variables if no other record shares its values for those variables. We need to examine two types of uniqueness: *population uniqueness*—a unit is unique in the population (or within a population data file such as a census) and sample *uniqueness*—a sample unit is unique within the sample file. These two concepts are the basis for many of the disclosure risk assessment methods for microdata. If a unit is population unique

then disclosure will occur if a snooper knows it is unique. Sample uniqueness is a necessary precondition of population uniqueness. Much existing methodology concerns using sample information to make inferences about population uniqueness. These methods are described in more detail in Chapter 3.

2.2.2 Matching/Reidentification Experiments

Using the same methods that a data snooper would, matching or reidentification studies simulate the linking of records from an identification file with those on the target microdata file (Elliot and Dale, 1998; Müller et al., 1995; Winkler, 1995a; Elliot, 2007). Such studies have the advantage that they are generated by empirical data, rather than depending on theoretical values provided by the uniqueness statistics discussed in the last section. However, there are corresponding disadvantages:

(i) We cannot be certain that a particular identification data set will provide a generalizable measure of the level of disclosure risk associated with the target microdata set; the results will be specific to the identification data set and the particular experiments conducted. A different data set with different data divergence from the target microdata set might well produce substantially different results.
(ii) Setting up matching experiments is time consuming, with considerable effort usually required to arrive at a harmonized coding for the two data sets.[3] Complicated procedures are usually necessary to verify accuracy.

2.2.3 Disclosure Risk Assessment for Aggregate Data

Much of the work on disclosure risk assessment has focused on modeling identification risk in microdata. Little has been done on exploring and developing equivalent risk assessment metrics for aggregate data. One reason for this imbalance is that whilst the conceptual structure of attacks on microdata (identification-attribution) is established and basically understood the equivalent conceptual structure for aggregate data (subtraction-attribution) is not.

Generally, risk assessment for tabular data has used ad hoc proxy measures. For frequency data, one commonly used measure is based on the numbers of "small" cell counts. For magnitude tables, a measure which identifies cells with dominant respondents leads to the heuristic *p/q* rule. Smith and Elliot (2008) provide a more theoretically grounded algorithm known as the Subtraction-attribution-probability (SAP), which generates a probability of a snooper being able to recover one or

[3] Although these are the same processes data snoopers would have to go through if they are to attempt a confidentiality breach.

more zeroes in a table, given specified knowledge about the population. One of the advantages of this metric is that it can be applied equally to unperturbed or perturbed tables, to single or sets of tables, and even to unreleased cross-classifications of released margins (effectively dealing with the linked tables problem). The method is described in more detail in Chapter 3.

2.3 Controlling the Risk

Having appropriate methods for measuring disclosure risk, the DSO must consider what to do about the measured level of risk. If the assessed risk is too high, it has two options: restrict access to the data or restrict the data. Often both of these options are pursued. In this section we consider several ways of controlling disclosure risk.

2.3.1 Metadata Level Controls

Controls at the metadata level work with the overall structure of the data release. The key components of such controls are the sampling fraction, choice of variables, and level of detail on those variables.

Sampling Fraction

For surveys, the sampling fraction is specified by the study design and so its choice often rests outside disclosure control. Nevertheless, the sampling fraction is critical in determining disclosure risk for a microdata file.

Choice of Variables

An obvious mechanism of disclosure control is by excluding certain variables from the released data set. The DSO can (i) reduce the number of key variables—those which a plausible snooper is likely to have access to or (ii) reduce the number of target variables. These choices flow naturally from the scenario analyses described in Section 2.1.3. With microdata, the choice is whether a variable appears in a data set or not. With aggregate data, the choices are about which variables will be included in each table.

Level of Detail

Decisions over level of detail mirror those over choice of variables. Here the DSO will look at categories with small counts and determine whether merging them with other categories would significantly lower disclosure risk without losing appreciable information. Not surprisingly, many data users would like the maximum level of detail possible on every data set. But DSOs regard some variables, especially geography and time, as particularly problematic in maintaining confidentiality. Area of

residence is a highly visible component of an individual's identity, therefore geographical detail is often constrained and data are released at coarser detail than data users would like. Similarly, sometime variables, such as exact date of birth, can when combined with other variables be straightforwardly identifying.

2.3.2 Distorting the Data

The main alternative to metadata controls is various forms of data distortion, which we call *perturbation*. These techniques manipulate the data in order to foil any identification/subtraction strategy so that a snooper cannot be certain of any match or recovered zero. In this section we will look at several methods of perturbation that are commonly used for disclosure limitation.

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. Data swapping will be discussed in detail in Chapter 5.

Rounding is a technique 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. This fact has motivated rounding variants that are described in Chapter 4.

For magnitude tables, *controlled tabular adjustment* or *cell perturbation* have been suggested. These techniques modify the original values in the table, but typically only the internal cells. In this way, marginal values are precise whilst the internal cell values are uncertain. Again these techniques are described in Chapter 4.

Cell suppression is an SDL 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 full table). If I release two one-way marginal frequency tables, but not the joint table, I am suppressing the cross-classification. In the cell suppression approach which is described in Chapter 4 individual cells are suppressed according to specified rules. Suppression can also be used in microdata where particular variables can be suppressed for particular cases.

2.3.3 Controlling Access

The DSO can control who may access the data and the manner of that access. In practice, only coarse aggregates are released without any restriction. With microdata, licensing is typically required to get access. Control over who accesses the data, for what purposes, and by which medium is often used in combination

with SDL techniques. For example, the Australian Bureau of Statistics (ABS) has released microdata from its census at three levels of detail. The least detailed is released on CD to users under license, a more detailed version is accessible via the Internet through ABS's remote access data laboratory, and the most detailed version is only accessible to trusted users who must do their work at ABS's offices.

2.4 Data Utility Impact

The final piece of the jigsaw in building an understanding of the risk assessment domain is the notion of *data utility*. The concern is that disclosure limitation will not only stymie a would-be data snooper, but also make them unusable for intended users.

The common approach to data utility is to generate metrics for the *information loss* caused by the disclosure control employed with a given data set. This can be either based on maintaining certain key statistics, for example, means, variances, co-variances, and so on (Cox et al., 2004; Domingo-Ferrer and Mateo-Sanz, 2001) or by maintaining some construct such as Shannon entropy or uncertainty functions (Duncan and Lambert, 1986).

The advantage of these approaches is that they are easily replicable and comparisons can be made between data sets. The major drawback of the information loss approach to data utility impact is that it is difficult to relate it to the actual utility of the data because data users are ultimately interested only in the data's usability for the analyses that they intend to conduct. If the data are fine for that purpose then the user is indifferent to what the DSO has done to them. Conversely, if they are not fit for the data users' purposes, then the DSO has contributed nothing of value, even if according to some measure there has been no substantial information loss.

The impact on data utility of SDL techniques falls into two categories, *reduction of analytical completeness* and *loss of analytical validity* (Purdam and Elliot, 2007). With some SDL methods, typically metadata controls, analyses that could have been conducted cannot be. This is a reduction in analytical completeness. Use, for example, of geographical thresholds in microdata sets leads to smaller administrative units being grouped together, thereby preventing researchers concerned with inferences about the smaller units from using the data set effectively. The loss of analytical validity is harder to define, but in some ways more critical because of its insidious nature. Loss of validity occurs when an SDL method has altered a data set to the point where a user reaches different conclusions from the same analysis. This is a danger with perturbative SDL techniques.

Removing a single variable from a data set may have an impact on the quality of the data as measured in an information-theoretic sense, but might have no effect on analytical completeness, certainly if no user would use that variable. Conversely, a minor perturbation in information-theoretic terms could have a significant effect on analytical validity if the perturbation affects important variables disproportionately.

Given the problems of the information-theoretic approach, Purdam and Elliot (2007) have developed retrospective methods for more directly assessing the impact on data utility. They survey data users (typically authors of studies) to assess the impact on analytical completeness. They also replicate published studies after the application of disclosure control techniques to assess the impact on analytical validity. These allow a more direct analysis of utility but suffer from the same problem of non-generalizability as the direct record linkage studies of risk.

In Chapter 6 we examine data utility in more detail describing notions such as the Risk-Utility Confidentiality Map.

2.5 Summary

As we have described it, the SDL field is comprised of three areas: (i) disclosure risk analysis (the assessment of the risk), (ii) disclosure limitation techniques (to reduce the risk); and (iii) the analysis of the impact on the utility of the data of those limitation techniques. Each of these areas poses interesting and challenging academic questions. However, development of SDL methodology is not solely or even primarily academically driven. One of the features of the SDL field is that as the academic questions are addressed they lead to changes in procedures within DSOs, usually first in the national statistical institutes. Indeed, many of the key researchers in the field are employed by the National Statistical Offices rather than by academic institutes. This combination of theoretical interest with policy and practice relevance makes SDL a compelling field for both researchers and practitioners of statistical confidentiality. Demonstrating this richness in both theory and in practice, the next chapter delves more deeply into the issues of disclosure risk analysis.

http://www.springer.com/978-1-4419-7801-1

Statistical Confidentiality
Principles and Practice
Duncan, G.T.; Elliot, M.; Juan Jose Salazar, G.
2011, XII, 200 p., Hardcover
ISBN: 978-1-4419-7801-1