1 Spontaneous recognition: an unnecessary control on data access?

Felix Ritchie

Professor of Applied Economics, Bristol Business School.

Abstract

Social scientists increasingly expect to have access to detailed source microdata for research purposes. As the level of detail increases, data owners worry about ‘spontaneous recognition’, the likelihood that a microdata user believes that he or she has accidentally identified one of the data subjects in the dataset, and may share that information. This concern, particularly in respect of microdata on businesses, leads to excessive restrictions on data use.

We argue that spontaneous recognition presents no meaningful risk to confidentiality. The standard ‘intruder’ model covers re-identification risk to an acceptable standard under most current legislation. If a spontaneous re-identification did occur, the user is very unlikely to be in breach of any law or condition of access. Any breach would only occur as a result of further actions by the user to confirm or assert identity, and these should be seen as a managerial problem.

Nevertheless, a consideration of spontaneous recognition does highlight some of the implicit assumptions make in data access decisions. It also shows the importance of the data owner’s attitude: for a default-open data owner, spontaneous recognition is a useful check on whether all relevant risks have been addressed, but for a default-closed data owner spontaneous recognition provides a way to place insurmountable barriers in front of those wanting to increase data access.

We use a case study on a business dataset to show how rejecting the concept of spontaneous recognition led to a substantial change in research outcomes.

Introduction

Social scientists increasingly expect to have access to detailed source microdata for research purposes. The twenty-first century has seen a major advances in the availability of detailed social science microdata for research purposes. Two elements combined to make much more of the data collected by national statistics institutes (NSIs) available to researchers:

- secure, remote access to detailed data with few limitations on researcher use
- external pressure on NSIs to make data available

These have driven massive expansion in the use of NSI data sources; and despite the potential of ‘big data’, NSI and other government data remains the most important for much research, particularly in economics.

However, this growth in data access has not always been actively driven by the data owners. As Ritchie (2016) notes, data owners, particularly in government, are often reluctant to release data. This arises from institutional and incentive structures which encourage a risk-averse attitude to decision-making (Ritchie, 2014b). Decision-makers are supported in this attitude by the academic literature; this overwhelmingly focuses on hypotheticals and the risk to the data owner, rather than any gain to the researcher (Hafner et al, 2015a).

The issue of ‘spontaneous recognition’ (SR) illustrates the difficulties facing those trying to improve access. Spontaneous recognition occurs when a microdata user believes that he or she has identified, without trying, one of the data subjects in the dataset: a neighbour, a co-worker, an organisation, or a group of patients, for example. The identification need not even be correct: the perceived breach of confidentiality can be as important as an actual breach. This causes great concern to data providers wanting to allow researchers to use confidential data: no matter how trustworthy researchers are, they are still human and the recognition of an individual might lead to the disclosure of information about an identified data subject. Hence, data providers are often insistent on minimising the risk of spontaneous recognition.

Despite this, very little attention is paid to the topic in the literature. Almost every manual on statistical disclosure control (SDC) or data owner’s guide to data handling mentions it, but largely as a pedagogical device before moving on to more sophisticated models.

Nevertheless, it is an important topic. Spontaneous recognition creates an important hurdle to be addressed by those requesting access to data. It is not possible, in general to show that there is no risk of spontaneous recognition, and so this gives great scope to those unwilling to release data for re-use scientific researchers or the public. This is particularly true in the case of business data, where the ‘obvious’ identifiability of businesses by one or two characteristics such as size and industry has been used in the past to restrict research access to the data.

We argue that the hurdle is an irrelevant distraction: spontaneous recognition has little or no practical contribution to make in the question of whether or not a dataset should be released. It is highly unlikely that there is any lawful or ethical basis for the concept. Other SDC methods cover the reasonable requirements of law and access conditions; and if any re-identification did occur spontaneously, it is the actions of the data user which should be governed, not the recognition itself which is an unconscious human act. These are better governed by management procedures.

The structure of the paper is as follows. The next section describes the fleeting appearance of spontaneous recognition in the literature. Section three considers whether spontaneous recognition is a useful statistical concept, in terms of building risk models. Section four reflects on whether there is any legal requirement to address spontaneous recognition. Section five shows how the real problem arising from spontaneous recognition comes from the actions of the user, not from the recognition itself. Section six considers the institutional impact of the concept, and shows how it can be helpful or obstructive depending on the organisation’s attitude. Section seven concludes.

SR in the literature

The OECD Glossary of Statistical Terms defines ‘spontaneous recognition’ as

> *… the recognition of an individual within the dataset. This may occur by accident or because a data intruder is searching for a particular individual. This is more likely to be successful if the individual has a rare combination of characteristics which is known to the intruder. (OECD, 2005)*

This glossary was defined and adopted to encourage the consistent use of statistical language across countries, academics and NSIs. However, the definition given above is not widely used, largely because it includes deliberate searching (and therefore includes intruder models).

Duncan et al’s (2011) definition is closer to the more commonly understood definition:

> *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. (Duncan et al, p35)*

Duncan et al (2011, p29) explicitly distinguish spontaneous recognition from ‘snooping’ or ‘intruder’ models where the user takes actions specifically to identify a data subject. We use the more common Duncan et al definition here.

In one sense, there is a large literature discussing spontaneous recognition: most of the general statistical works on SDC (eg Hundepool et al, 2012) cover it, as do the guidelines produced by NSIs. It is often used to give examples of extreme values; for example, GSS(2014) describe it:

> *An intruder may spontaneously recognise an individual in the microdata by means of published information. This can occur for instance when a respondent has unusual characteristics and is either an acquaintance or a well-known public figure such as a politician, an entertainer or a very successful business person. An example is the “Rich List” which publishes annual salaries of high-earning individuals. (GSS, 2014, p18)*

Spontaneous recognition is then typically used to explain why extreme values or population uniques are problematic and may need to be removed from the data.

However, this is the full extent of the discussion in the SDC literature. As Hafner et al (2015a) have noted, almost all of the confidentiality literature is focused on deliberate attempts to re-identify the data, the so-called ‘intruder model’. In this context, spontaneous recognition disappears as an uninteresting special case.

In summary, in the confidentiality literature, spontaneous recognition is a useful teaching tool but otherwise not considered.

SR as a statistical problem

Defining spontaneous recognition

Following Duncan et al (2011) we define spontaneous recognition as the accidental identification of a data subject; that is, without actively searching. However, we add the condition that the identification need not be accurate.

In this we differ from most of the SDC literature, which assumes that identification is only a problem is the true identity is uncovered[1]. This is clearly consistent with legal requirements to keep data confidential, but it ignores the institutional impact. Asserting that confidentiality can be breached can have a substantial effect on the reputation of the data owner, whether that assertion is true or not. One NSI had to undertake a substantial public relations operation after it was (falsely) claimed that a multinational supermarket used confidential Census data to send out mailshots.

An additional complication is what ‘identity’ is being uncovered. Some data subjects, such as organisations, can have complex structures which makes identification a much more difficult concept.

For example, assume that there is one university on the Isle of Wight, in the southern UK[2]. Suppose a researcher using business data from the island comes across an entity whose industrial classification describes it as ‘university’, and finds no other entities with that classification. There are three possibilities.

[1] This is done for pedagogical purposes. The literature does of course recognise that false inferences may occur. However, the implications of this are rarely discussed; only accurate inference is considered.

[2] This example is for illustration only. We are not aware of any university on the Isle of Wight at present.

- This is the whole university
- This is part of the university
- This is a branch office of a mainland university; the reporting units of the Isle of Wight university are not classified as ‘university’ for some reason

Clearly with complex data subjects the ‘accuracy’ of the identification has more room for interpretation. In line with the above discussion, we treat spontaneous recognition as occurring when the data users thinks ‘I have found a data subject that I can put a name to’, irrespective of whether that name relates to an accurately identified unit or not.

A related concept is ‘identity confirmation’. This is where a user who spontaneously recognises a data subject takes active steps to confirm his or her suspicions about the data subject. For example, the researcher could cross-check with other information in the dataset, or external information. This differs for the intruder model in that the researcher has no specific interest in attacking the dataset; the researcher’s curiosity has been aroused, and the purpose of further investigation is to satisfy that curiosity.

Finally, ‘identity assertion’ is where a user who spontaneously recognises a data subject reports his or her suspicions to another, again without deliberate intent to reveal information but as a human response to an interesting finding.

Because these two concepts both require action by the user after the initial suspicion has been aroused, we combine them as ‘identity confirmation or assertion’.

Spontaneous recognition as a theoretical risk

Two obvious risks are widely described in the literature:

- *Population uniques on one or two characteristics or extreme values.*

For example, in 2014 the first female bishop in the UK Anglican church was elected. This was a high profile event, with wide newspaper coverage. Female clerics in the UK are comparatively rare, and a detailed job description or a salary range (indicating the highest paid) combined with gender could prompt a memory in the data user. Alternatively, a small geographical area might have one well-known high-value celebrity resident. Salary or wealth data may be enough to identify that individual. For business data, this is the most significant problem. Detailed industrial classification and size of business (typically employment or turnover) are assumed to be enough to identify well-known large players, such as in telecoms or aerospace.

- *Sample uniques where the sample is known.*

Sample uniques which are not population uniques on the key characteristics are not normally of concern; by definition they represent at least two indistinguishable data subjects. However, we can consider cases where the data user might have additional information about the sample, making the sample uniques into population uniques. For example, a neighbour may tell a researcher that she was included in a particular wave of the survey the researcher is using.

A third risk, less commonly documented, is that an unsophisticated user may mistake a sample unique for a population unique, and draw an inappropriate inference.

To avoid these risks, SDC good practice normally requires that population uniques are disguised or removed. For example, Statistics New Zealand’s old Confidentiality Protocol explicitly identified spontaneous recognition with population uniques (Statistics NZ, 2000, appendix B). This is why business data is commonly described as being impossible to anonymise: the variables of interest (industrial classification, size) are essential components in research, and so cannot be removed while maintaining value in the data.

Sample uniques are also avoided where the underlying number of population uniques is small; for example, Schulte-Nordholt (2013) describe Dutch public use Census files as having a minimum 1000 observations on trivariate categorisation, and minimum five observations on all household characteristics.

This practice is uncontroversial, and allows SDC advisors to concentrate on the seemingly more important problem of active, intruder, attacks on confidentiality. It is assumed that re-identiification through deliberate action must have a success rate which is no lower than that of accidental discovery. Spontaneous recognition is intruder matching without a match model; therefore it can be treated as the less interesting special case. Mackey (2013), for example, treats spontaneous recognition as the starting point for a formal intruder-based review.

Nevertheless, examining spontaneous recognition in its own right can throw light on some of the underlying assumptions.

First, intruder models are designed to give risk measures based on assumptions about behaviour. However, the risk of spontaneous recognition cannot be estimated and cannot be proved not to exist, because the personal information that leads to it is unknowable. A dataset will contain population uniques unless it is K-anonymous on all variables (that is, any combination of variables, including continuous variables, must give at least K duplicate observations); for spontaneous recognition, K=2 as no active searching is considered. A fully K-anonymous dataset has limited research value, and so in practice removal of uniques is carried out using a subset of variables which is determined subjectively (Skinner, 2012). Therefore spontaneous recognition based on an unpredicted combination of variables must be possible. For example, wages are not normally considered identifying except for extreme values; but a researcher looking at a dataset on employees might notice a promotion in the wage data, and link that to personal knowledge of a specific employee.

Second, protection measures designed to stop intruders may not be relevant for identity confirmation. By construction, spontaneous recognition arises from a combination of knowledge not foreseen in the protection algorithm. It follows that cross-checking of information in the dataset to confirm a suspicion does not need to use the scenario used to create the protection algorithm.

Third, it is not possible to test for spontaneous recognition. By construction, spontaneous recognition arises from the accidental linkage of personal information with specific data. Testing by asking users to try re-identifying data subjects (as for example in Spicer et al, 2013) cannot formally replicate the conditions for accidental re-identification (although it would clearly provide useful evidence as to whether the data protection is 'good enough').

Fourth, spontaneous recognition is implicitly accepted in research files. For public use files (PUFs; those with no restrictions on use), stringent precautions are taken against direct attack, and hence spontaneous recognition. For scientific use files (SUFs; access limited to verified researchers) and secure use files (SecUFs; access limited to controlled environments), the level of precaution is more complex as more controls are available. For example, Spicer et al (2013) discuss how the access restrictions on SUFs enable a more relaxed approach to intruder attacks. However, this implies a simultaneous increase in the probability of spontaneous recognition. In other words, for anything other than PUFs, a non-negligible level of spontaneous recognition is implicitly being accepted; see Table 1.

Table 1 Subjective expectation of SDC controls

File type	SDC controls	Acceptable intruder risk	Acceptable spontaneous recognition risk
PUF	Many	None	None
SUF	Fewer than PUF	More than PUF	More than PUF
SecUF	Fewer than SUF	More than SUF	More than SUF

Finally, the likelihood of inaccurate spontaneous recognition is not covered in the SDC literature for the simple reason that there is no meaningful way to address the problem. What is the probability that a user looking at a dataset will make assertions based on an inaccurate identification? There is good empirical evidence that humans are over-confident in their ability to re-identify data subjects, but this evidence is based on tests under known conditions where test subjects are required to express their confidence in their predictions. It is not clear how one would test whether this applies in non-test conditions where data users are under no pressure to express an opinion.

Practical implications of the statistical approach

In summary, spontaneous recognition and its consequences are unpredictable, untestable and unprovable. This means that protection based around the notion of the predictable intruder might be ineffective.

In practice, spontaneous recognition in PUFs is ignored. The evidence of a half century of anonymisation suggests that focusing on intruders seems to provide adequate protection. Inaccurate assertions about individuals do not seem problematic.

In files created for researchers (SUFs and SecUFs), a non-negligible risk of spontaneous recognition is implicitly accepted. However, in terms of behaviour, spontaneous recognition is the exact opposite of active, intruder, re-identification. To many data owners this suggests that it should be treated as a separate problem and tackled independently.

SR as a legal problem

If spontaneous recognition occurs, it is not clear that any breach of confidentiality has occurred.

In PUFs, spontaneous recognition would imply that the anonymisation procedure has failed. While the data owners would be expected to review the anonymisation procedures, it is unlikely that this would lead to legal consequences. Most data protection laws (for example, the regulation covering data management in the EU) require data owners to take all reasonable protection measures, not all measures; some laws explicitly absolve the data owner of any legal responsibility in the case of a mistake. It would be difficult to argue that an intruder-protected PUF is inadequately protected against spontaneous recognition (assuming, of course, that the intruder protection is carried out to an accepted standard).

For researcher files, a non-negligible risk of spontaneous recognition is implicitly if not explicitly approved, as noted above. What are the consequences if a researcher recognises a data subject? The answer is nothing: the researcher has been granted lawful access to the data in that state, and nothing has changed.

What happens next does matter. The researcher has four options:

1. Identity confirmation: cross-checking the data with any other information
2. Identity assertion: mentioning the fact to another researcher or a non-user
3. Identity assertion: mentioning the fact to the data owner
4. Taking no further action

Action (1) may or may not be a breach of confidentiality, but it is almost certainly a breach of the access terms to the data, as the researcher is now trying to actively re-identify a data subject (in many jurisdictions, this would also be a breach of the law).

Action (2) is also likely to be a breach of access terms: mentioning something discovered about a data subject could be taken as seeking confirmation of the identity of the data subject, seeking to provide another with identifying information, or both. It is not clear whether an offence has been committed if the identification is inaccurate, but most data access agreements ban any information being shared about data subjects, whether that information is true or not.

The consequences of action (3) depend on the attitudes of the data owner. Data owners following the Active Researcher Management principles (in essence, a shared responsibility model; see Desai and Ritchie, 2010) should welcome information about easily recognisable subjects as an opportunity to review protection measures in the light of new information. However, the authors have observed data facilities where any speculation about the identity of data subjects, even to the data owners and irrespective of intent, is strictly forbidden and liable to penalties.

Some data owners require users to report any suspected identification, and so action (4) may be a breach of access conditions. However, it is not clear how a data owner could prove that spontaneous recognition has happened, unless one of actions (1) to (3) was also taken.

In summary, spontaneous recognition by itself does not seem a breach of confidentiality on behalf of the data user. For PUFs, the fault lies with the creator of the file. For research files, any breach of law or access agreements arises from additional actions taken by the researcher. In other words, the problem arises from the actions of the users, not the statistical protection in the data.

SR as a management problem

Non-negligible possibility of spontaneous recognition is implicitly accepted in research data files, and that this poses no legal problems. Instead, breaches of confidentiality or procedures occur when the data user takes some follow-up action, identity confirmation or assertion. This clearly identifies the risk associated with spontaneous recognition as a user management problem.

This perspective offers several advantages over seeing a statistical or legal problem.

First, it focuses on the unlawful activity: searching for identity, or speculating on identity with another. It does not criminalise users for an automatic response (recognition) to some information presented to them. It penalises behaviour, not thoughts.

Second, it is likely to be easier to detect actions to confirm or assert an identity, whereas detecting whether someone has identified a data subject is impossible to know until they share that knowledge.

Third, it requires no assumptions to be made about what personal knowledge a data user might have that could lead to spontaneous recognition. It is only the outcomes that matter, not the inputs.

Fourth, it reduces incentives to damage data as protection against something which is explicitly acceptable, at least in research files.

Fifth, the protection measures are already covered to some degree. Providers of research files usually give users training or written guidelines, or both, which state that attempts to re-identify data subjects, or discuss characteristics of the data with unauthorised users, is prohibited. Data access agreements may have similar wording, although there is little evidence to suggest that users read these.

Six, the management approach can be applied to PUFs as well as research files. Thus, recognition of an individual in a PUF is no longer a failure of statistical technique, but the manifestation of a known and accepted managerial risk. The change in emphasis, from blaming individuals to corporate learning, should discourage SDC advisors from worst-case risk avoidance strategies to protect themselves.

Finally, as this is a management issue, and not a legal one, then the data owner can choose to promote positive behaviours. For example, reporting of suspected identification without penalty can be encouraged.

Best practice user training and communication encourages the development of a community of interest between researchers and the data owners (Desai and Ritchie, 2010; Eurostat, 2016). Training about spontaneous recognition versus identity confirmation or assertion can be used to reinforce messages of trust in the training. The unavoidability of human failings can be contrasted with working with the support team to ensure that no-one gets into trouble. Messages about inaccurate identification can also be pushed in training. The fictional example of the Isle of Wight university, given above, could be used to emphasise the scope for error in any assertions.

In summary, viewing spontaneous recognition as an irrelevance, and seeing identity confirmation or assertion as a problem of user management,

- focuses attention on unlawful activity only
- allows data owners more flexibility to deal with problems, even for PUFs
- encourages the community of interest amongst data owners and users
- is entirely consistent with best user training practice

Attitudes and default perspectives

Although it may have no statistical value, the concept of spontaneous recognition can have a practical impact because of the data owner's attitude.

Data owners' attitude can be simplified to one of the following (Hafner et al, 2015a):

- Default-open: release data unless the release is shown to be unsafe
- Default-closed: do not release data unless the release is shown to be safe

In theory these two positions are identical, but Ritchie (2014a) shows that the phrasing generates very different outcomes. Most NSIs notionally claim to be default-open; in practice, almost every organisation is default-closed as institutional incentives and the academic literature encourage decision-makers to focus on their personal responsibility (Ritchie, 2016).

For a default-closed data owner, spontaneous recognition offers an unbeatable hand. As noted above, spontaneous recognition arises from an unexpected combination of luck and unpredictable knowledge. It is not possible to demonstrate that this cannot happen, nor is there any evidence that it cannot happen (such evidence would have been incorporated into predictable risk). Hence, a data owner unwilling to release data can call spontaneous recognition as a risk without fear of losing the argument.

This can be done even though, for all practical purposes, the intruder model and management strategies make spontaneous recognition irrelevant. For a default-closed data owner, little expected practical impact may not be enough; it is the potential for spontaneous recognition to occur that must be demonstrably negligible.

For a default-open data owner, the reverse is true: spontaneous recognition can be a very useful check on the validity of one's risk scenarios. Considering spontaneous recognition encourages one to treat and eliminate the foreseeable risks; when the only remaining untreated risk is spontaneous recognition (ie entirely unpredictable risk), then the data owner can be satisfied that the release is now no longer 'shown to be unsafe' to any limit of reasonableness.

This issue raises important questions about the way data owners are persuaded to allow their data to be used. Data owners should be concerned about identity confirmation or assertion, but they may be unable to articulate it as the literature focuses on spontaneous recognition. Those advocating greater use therefore have a role to play in making data providers aware of the specific risk being raised.

The creation of 2010 Community Innovation Survey SUF (described in detail in Hafner et al, 2015b) illustrate the impact of the default-open/default-closed attitude. The creation of an SUF for business data is unusual, because of the assumed identifiability of the businesses, and hence, prior to the 2010 survey, all continuous variables for all observations were perturbed to reduce the attribution risk associated with a successful identification. However, a review taking a default-open approach argued that the only risk not handled by practical controls was the chance of careless assertions of identity. Accordingly, risk management focused on user training, and less than 1% of observations, the cases most likely to prompt speculation, were perturbed.

In short, for the default-open data owner spontaneous recognition offers a handy rule-of-thumb to determine whether there are any remaining untreated risks in a dataset; for a default-closed owner, it offers unlimited potential to place an unfeasible burden of proof on those wanting to release data.

Conclusion

The public good is best served by making data available for research with as little damage as possible. Data owners, facing institutional pressures which encourage them to place the needs of the organisation over the wider public good, raise concerns about both the deliberate re-identification of individuals (the intruder model) and accidental re-identification through spontaneous recognition.

The intruder model is the workhorse of statistical data protection; almost the entire literature and most practice is based on it. In contrast, spontaneous recognition has had negligible formal examination. It is used for pedagogical purposes, to demonstrate simple examples, before the intruder model takes over in formal modelling.

One reason for this is that spontaneous recognition is not easily amenable to formal modelling. By its nature, it arises from unpredictable knowledge. If that knowledge were predictable, then intruder models would be able to incorporate it. This leads to the second reason for ignoring spontaneous recognition: it is subsumed into the intruder model as a special case.

Despite this lack of theoretical or practical value, data owners do raise concerns about spontaneous recognition. The authors observe this most often in relation to business data, where it seems ‘obvious’ that any useful microdataset is going to include many pieces of information that would prompt a researcher to speculate on the identity of the business. Concerns about spontaneous recognition have in the past led to restrictions on research.

Hence the conclusion of this paper, that the concept of spontaneous recognition has no place in data protection, is important. This conclusion derives from two observations.

The first observation is that the intruder model, despite its flaws, does effectively encompass the spontaneous recognition problem: it focuses on predictable risks and allows for active searching, meaning that any remaining risk arises from luck and a complete unpredictable set of events. This is likely to meet the test of ‘reasonableness’ embodied in most data protection laws. Because the intruder model can be applied to different environments (PUFs, SUFs, SecUFs), it can incorporate spontaneous recognition in those different environments.

The second observation is that any re-identification from spontaneous recognition leads to a managerial problem. Breach of confidentiality arises from the follow-on actions of the user, the identity confirmation or assertion, not the spontaneous recognition itself. Hence any management plan must focus on the training of users and the relationship between users and data owners, not on predicting the unpredictable.

Nevertheless, an examination of spontaneous recognition can usefully highlight implicit assumptions being made in data release decisions; and it demonstrates the importance of the data owner’s attitude. With the default-closed attitude spontaneous recognition is an unplayable hand whose only value is to block access. With a default-open attitude, spontaneous recognition becomes a useful sounding board to explore the limits of our knowledge and develop non-statistical risk models to cover for the ‘unknown unknowns’.

In summary, the statistical problem of spontaneous recognition is an unhelpful chimaera encouraging the underutilisation of valuable data. The problem that should be addressed is one of identity confirmation, which is management issue. A change in both language and attitudes, a focus on the exact nature of the problem being raised, and the use of evidence can generate substantial dividends for both data providers and users.

References

Desai T. and Ritchie F. (2010) "Effective researcher management", in Work session on statistical data confidentiality 2009; Eurostat.

Desai T., Ritchie F., and Welpton R. (2016) The Five Safes: designing data access for research. Working papers in Economics no. 1601, University of the West of England, Bristol. January

Duncan, G T, Elliot, M, Salazar-Gonzalez, J (2011). Statistical Confidentiality. Principles and Practice. Springer: New York Dordrecht Heidelberg London.

Eurostat (2016) Self-study material for the users of Eurostat microdata sets. http://ec.europa.eu/eurostat/web/microdata/overview/self-study-material-for-microdata-users

GSS (2014) GSS/GSR Disclosure Control Guidance for Microdata Produced from Social Surveys. Office for National Statistics/Government Statistical Service. https://gss.civilservice.gov.uk/wp-content/uploads/2014/11/Guidance-for-microdata-produced-from-social-surveys.pdf

Hafner H-P., Lenz R., Ritchie F., and Welpton R. (2015a) "Evidence-based, context-sensitive, user-centred, risk-managed SDC planning: designing data access solutions for scientific use", in UNECE/Eurostat Worksession on Statistical Data Confidentiality 2015, Helsinki.

Hafner H.-P., Ritchie F. and Lenz R. (2015b) "User-centred threat identification for anonymized microdata". Working papers in Economics no. 1503, University of the West of England, Bristol. March

Hundepool A, Domingo-Ferrer J, Franconi L, Giessing S, Schulte Nordholt E, Spicer K, de Wolf P. (2012). Statistical Disclosure Control. Wiley.

Mackey E. (2013) European Union Statistics on Income and Living Conditions (EU-SILC). case stiudy. Mimeo, UK Anonymisation network. http://ukanon.net/wp-content/uploads/2015/09/EUROSTAT-EU-SILC-DATA-Nov-2013-pdf.pdf

OECD (2005) Glossary of statistical terms. November. https://stats.oecd.org/glossary

Ritchie F. (2014a) "Access to sensitive data: satisfying objectives, not constraints", J. Official Statistics v30:3 pp533-545, September. DOI: 10.2478/jos-2014-0033.

Ritchie F. (2014b) "Resistance to change in government: risk, inertia and incentives". Working papers in Economics no. 1412, University of the West of England, Bristol. December

Ritchie F. (2016) "Can a change in attitudes improve effective access to administrative data for research?", Working papers in economics no. 1607. University of the West of England, Bristol.

Schulte-Nordholt E (2013) Access to microdata in the Netherlands: from a cold war to co-operation projects. Work session on statistical data confidentiality 2013; Eurostat.

Skinner C. (2012) Statistical Disclosure Risk: Separating Potential and Harm, Int. Stat. Rev. v80:38

Spicer K., Tudor C. and Cornish G. (2013) Intruder Testing: Demonstrating practical evidence of disclosure protection in 2011 UK Census. Worksession on statistical data confidentiality 2013; Eurostat.

Statistics New Zealand (2000) Confidentiality Protocol. http://unstats.un.org/unsd/dnss/print.aspx?docID=141