Spontaneous recognition: an unnecessary control on data access?

Felix Ritchie

University of the West of England, Bristol

Abstract

Social scientists increasingly expect to have access to detailed data for research purposes. As the level of detail increases, data providers 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 models of deliberate attack on the data cover re-identification risk to an acceptable standard under most current legislation. If spontaneous recognition 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 made in data access decisions. It also shows the importance of the data provider's culture and attitude. For data providers focused on users, spontaneous recognition is a useful check on whether all relevant risks have been addressed; for those focused on the risks. For data providers primarily concerned with the risks of release, it provides a way to place insurmountable barriers in front of those wanting to increase data access.

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

Executive summary

There is a reasonable expectation nowadays that social science researchers can have access to the source data underlying published statistics. Data providers can be nervous about this when the source data is confidential, even if the data has had direct identifiers such as names removed.

One issue raised is "spontaneous recognition": the idea that a researcher using the data might recognise a neighbour, or a famous person, for example. This is a particular problem for data about companies: large public companies are assumed to be easy to recognise with just a couple of pieces of information, such as size and type of business. The amount of information available on the internet only appears to make this problem worse. The only way to prevent spontaneous recognition is to reduce the detail in the data, which may defeat the purpose of making that data available. In some cases data providers have refused to release data at all, arguing that spontaneous recognition cannot be eliminated.

This paper agrees that, indeed, spontaneous recognition cannot be eliminated; but it also argues that it is impossible to prove that it cannot exist in any dataset. This makes it useless as an aid to decision-making. Moreover, there is no legal basis to the use of this concept for stopping data access. Existing methods of confidentiality protection, using both statistical and non-statistical controls, provide more than adequate protection, and the evidence suggests that these are effective in practice as well as in theory.

Why then do data providers refer to spontaneous recognition as a potential problem? The answer is a complex manifestation of cultural factors, institutional incentives, and a defensive confidentiality literature. These encourage decision-makers to be overly risk-averse, and to seek to avoid or transfer responsibility.

This traditional "default-closed" attitude can be contrasted with the "default-open" attitude emerging amongst data providers and the confidentiality community. The default-open model uses evidence-based risk assessment (rather than hypothetical worst cases), and requires proof that a problem does exist (rather than proof that a problem does not exist).

This paper uses an example of creating a file of business data for scientific research use, adopting the arguments described in this paper. Compared to the previous "traditional" strategy, the revised approach led to a fall from 100% to less than 1% of records being perturbed: a dataset which was previously seen mostly as a teaching resource is now much closer to its original research value.

1. Introduction

Social scientists increasingly expect to have access to detailed source microdata for research purposes. The twenty-first century has seen 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) and other government departments 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 a massive expansion in the research use of microdata provided by government. The proliferation of digital information has also increased the range of data sources across different platforms, but NSI and other government data remains the most important for much research, particularly in economics and in public health.

However, this growth in data access has not always been actively driven by the data providers. As Ritchie (2016) notes, data providers, 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 provider, rather than the public benefits foregone by overly restrictive access (Hafner et al., 2015a).

The issue of *"*spontaneous recognition" 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 worries data providers wanting to allow researchers to use confidential data: no matter how trustworthy researchers seem to be, 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 this topic in the literature. Almost every manual on statistical disclosure control (SDC) or data provider's guide to data handling mentions the topic, but only as a pedagogical device before moving on to more sophisticated models.

Nevertheless, it is an important topic. Spontaneous recognition creates an additional, sometimes insurmountable, 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; this gives great power to those resistant to releasing data for re-use by 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. If any re-identification did occur spontaneously, it is an unconscious human act. It is the actions of the data user which should be governed, not the recognition itself. These actions 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 gives the definition that will be used in this paper, which differs slightly from the general definition in that it allows for mistakes. Section four then considers whether spontaneous recognition is a useful statistical concept, a legal problem or a managerial issue, and concludes that it is only the latter. Section five considers the institutional impact of the concept, and shows how it can be helpful or obstructive depending on the organisation's attitude. Finally, section six introduces a short case study, where a rejection of spontaneous recognition led to a substantially improved dataset. Section seven concludes.

For clarity of exposition, we assume that data access is being managed by the NSI. However, the issues raised here, particularly in relation to public sector culture and attitudes, are relevant for all data providers.

2. Spontaneous recognition in the confidentiality literature

The probability of spontaneous recognition is a statistical characteristic of the data. 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.

Duncan et al. (2011) give the more commonly understood definition:

"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." (Duncan et al., p. 35)

Duncan et al. (2011, p. 29) explicitly distinguish spontaneous recognition from *"*snooping" or *"*intruder" models where the user takes actions specifically to identify a data subject. Spontaneous recognition is the *"*specific form" of the more general intruder model because it does not require active searching.

In one sense, there is a large literature discussing spontaneous recognition: most of the general statistical works on SDC (e.g. 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) propose:

"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, p. 18)

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 pretty much 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 *"*intruder*"* model. In this context, spontaneous recognition disappears as an uninteresting special case: the intruder

goes straight to identification without the need for statistical analysis (although the intruder may carry out some statistical analysis to confirm the identification).

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

3. Defining spontaneous recognition

We define spontaneous recognition as

"the accidental identification of a data subject (that is, without actively searching), whether that identification is accurate or not".

This broadly follows the definition of Duncan et al. (2011), but we add the condition that the identification need not be accurate, in contrast to the implicit assumption in most of the SDC literature. That literature does of course recognise that false identification may occur, but it focuses on accurate identification for pedagogical reasons. The impact of false inference is usually treated as a separate topic, if at all.

This popular working assumption, that identification is only a problem if the true identity is uncovered, 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 provider, whether that assertion is true or not. For example, 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 defining the "data subject" whose 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, an island in the southern UK[1]. Suppose a researcher using business data from the island comes across an entity whose industrial classification describes it as a "university", and finds no other entities with that classification. There are three possibilities:

- 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 there is more uncertainty about the "accuracy" of the identification. Hence, we regard spontaneous recognition as occurring when the data user 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.

Two other definitions are necessary: identity confirmation and assertion.

"Identity confirmation" 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 from the usual 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.

[1] This is for exemplary purposes. We are not aware, at present, of any university on the Isle of Wight.

"Identity assertion" is where a user who spontaneously recognises a data subject reports his or her suspicions to someone else, again without deliberate intent to reveal information but as a human response to an interesting finding.

Since these two concepts both require action by the user after the initial suspicion has been aroused, we combine them as "identity confirmation and/or assertion".

4. A legal, statistical or management problem?

4.1 Spontaneous recognition as a statistical problem

There are three statistical risks arising from spontaneous literature, but only two are widely described in the literature:

1. 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 uniquely identify that individual. For business data, population uniques are 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.

2. 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, on learning that a researcher is using a particular dataset, may tell the researcher that she was included in a specific wave.

3. Sample uniques mistaken for population uniques

An unsophisticated user may mistake a sample unique for a population unique, and draw an inappropriate inference. This is rarely discussed in the literature for the simple reason that it has no obvious statistical solution, arising as it does from an individual's misperception.

To avoid these risks, SDC good practice normally requires that population uniques are disguised or removed; and sample uniques are avoided or limited, particularly where the number of population uniques is small. For example, Schulte-Nordholt (2013) describe Dutch public use census files as having a minimum of 1000 observations on any three-way combination of characteristics, and a minimum of five observations when observed on all household characteristics. Statistics New Zealand's old Confidentiality Protocol explicitly equated 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.

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-identification 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" (Mackey, 2013); therefore it can be treated as the less interesting special case. Mackey (2013), for example, regards spontaneous recognition as providing the starting point for the formal intruder-based review.

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

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 matching 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; in terms of behaviour, spontaneous recognition is the exact opposite of active, intruder, re-identification. It follows that cross-checking of information in the dataset to confirm a suspicion does not need to use the scenario applied in order to create the protection algorithm.

Third, it is not possible to test for spontaneous recognition when assessing the effectiveness of SDC protection in a dataset. 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, 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 (Kahnemann, 2012), 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.

Finally, 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.

To see this, consider that each data access solution can be cast as a matter of choosing control in five dimensions: projects, people, settings, data and outputs; this is the widely used Five Safes model (Desai et al., 2015; www.fivesafes.org). Any specific solution will place more or less emphasis on different elements. For example,

there are no possible controls over who can do what with a PUF; therefore, the only thing that can be done is to control the confidentiality risk in the data. In contrast, a SecUF operating through a safe centre such as that run by Eurostat has a high degree of control over the users, environment, purpose of the work, and any outputs, and so minimal restrictions can be placed on data; see Table 1.

Table 1 Subjective expectation of access controls and risk

File type	Control possible	Acceptable intruder risk	Acceptable SR risk
PUF	Data only	Negligible	Negligible
SUF	Some degree of control over all elements	More than PUF	More than PUF
SecUF	High degree of control over all elements	More than SUF	More than SUF

The more non-statistical controls that are applied to the data, the fewer controls need to be applied to the data itself. Acceptance of spontaneous recognition and intruders is implicit when anything other than a PUF is being designed.

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 is ignored.

For PUFs, 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.

4.2 Spontaneous recognition as a legal problem

If spontaneous recognition does happen, 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 providers 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 providers to take **all reasonable** protection measures, not **all** measures. Some laws explicitly absolve the data provider of legal responsibility in the case of a mistake (Green and Ritchie, 2016). It would be difficult to argue that an intruder-protected PUF is inadequately protected against spontaneous recognition (assuming 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. Are there consequences if a researcher recognises a data subject? No; 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 provider;

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 of 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 accurate or not.

The consequences of action (3) depend on the attitudes of the data provider. Data providers following Active Researcher Management principles (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 providers and irrespective of intent, is strictly forbidden and liable to penalties.

Some data providers 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 provider 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.

4.3 Spontaneous recognition as a management problem

A non-negligible possibility of spontaneous recognition is implicitly accepted in research data files, and poses no legal problems. 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 it as a statistical or legal problem.

First, it focuses on the unlawful activity: searching for identity, or speculating on identity with someone else. 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 in place to a large 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 provider 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 providers (Desai and Ritchie, 2010; Eurostat, 2016). Training on 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, for example:

"You may come across a data subject you think you recognise; we can't stop your brain making links. But don't ever try to confirm your suspicions, or talk about them; this counts as trying re-identify a data subject, which will get you into a lot of trouble. Come and talk to us if you find something you think shouldn't be there."

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:

"…and the chances are that your identification is wrong, which means you'll still be in deep trouble but you'll look like an idiot as well…"

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

- addresses only unlawful activity;
- allows data providers more flexibility to deal with problems, even for PUFs;
- encourages a community of interest amongst data providers and users;
- is consistent with current best practice training principles (Eurostat, 2016).

5. Cultures, 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 provider's cultural perspective and resulting attitudes. These reflect the way that the decision-making process is approached in the organisation, and what the organisation sees as its main priorities.

The institutional culture of the data providers can be simplified as one of the following (Hafner et al., 2015):

- 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 a very different response. It arises because of a difference in what is perceived as the initial "endowment" (Kahnemann, 2012). In the default-open model, research value is the default position, and it is being traded off for increased security; in the default-closed model, the opposite is the case. Humans value losses more highly than gains, and so the default perspective will affect the outcome.

Most NSIs notionally claim to be default-open; in the author's experience, almost everyone is default-closed. Ritchie (2014a, 2016) and Hafner et al. (2015a) note that the default-closed culture arises from three sources:

- institutional incentives in the public sector tend to be focused on the costs of individuals making bad decisions, rather than the overall loss to society of good decisions foregone; in other words, the public sector is encouraged to think defensively (Ritchie, 2014a);
- the statistical literature emphasises extreme scenarios, hypothetical examples, and worst-case risk avoidance, rather than evidence-based modelling, encouraging risk aversion in data providers who are not experts in assessing disclosure risk analyses (Hafner et al., 2015; Ritchie, 2016);
- the public choice literature, which has dominated public sector management since the 1980s, encourages data providers to view users as self-interested and so inherently untrustworthy (Ritchie, 2014a).

In these circumstances, it is not surprising that NSIs develop risk-averse cultures. In this context, statistical analysis is essential for allowing an NSI to claim an objective rationale for its decisions, despite these decisions being highly subjective (Skinner, 2012).

The Five Safes model discussed above is a decision-making framework rather than a strategic tool; nevertheless, it does illustrate how the statistical tail can wag the research dog. For a default-open NSI, focused on research, "projects", "people" and "settings" are the key dimensions; the other two imply at least a potential restriction on research, which is not a desirable outcome. SDC is something that happens when the other controls do not provide an appropriately secure solution for the users. In contrast, for a default-closed NSI, confidentiality is more important than use value, and so using the "objective" tools of SDC protection is more appealing.

For a default-closed data provider, 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 provider unwilling to release data can declare 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 provider, 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 provider, 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 (i.e. entirely unpredictable risk), then the data provider can be satisfied that the release is now no longer "shown to be unsafe" to any limit of reasonableness.

In short, for the default-open data provider 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.

This issue raises important questions about the way data providers are persuaded to allow their data to be used. Data providers 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. As Green and Ritchie (2016) note, there is very little hard evidence about genuine risks because academic research is generally a very low-risk activity; however, the operational (and cost) implications of alternative perspectives can be very large.

6. Case study: the 2010 CIS scientific use file

We conclude with a brief discussion of a specific case where an evidence-based managerial approach to spontaneous recognition led to substantially better outcomes for both data provider and researchers. The case study is discussed in more detail in Hafner et al. (2015b).

The Community Innovation Survey (CIS) has been conducted every two years since the 1990s, and collects information from businesses on their innovation and research and development (R&D) activity. It is a stratified sample – a higher proportion of large firms are included – and most of the variables are qualitative. Each EU country carries out its own survey, using the same (translated) questionnaire. The results are transmitted to Eurostat as aggregate tables, but most countries also send their microdata to Eurostat. This allows researchers visiting the Eurostat Safe Centre (ESC) in Luxembourg to have access to a pan-European SecUF. Government users tend to produce tables from the data, whereas researchers use the microdata to produce regressions and other marginal analyses.

Travelling to the ESC in Luxembourg incurs time and money costs for researchers. Distributing the data via an SUF would address some of the needs, and Eurostat creates SUFs for other datasets (such as the European Labour Force Survey). SUFs based on business data are rare because of the supposed ease of identifying large and unusual businesses, but Eurostat had commissioned SUFs on the CIS from 2004 onwards. However, the anonymisation technique employed removed a substantial amount of detail, and there was a general impression that the dataset was better suited to teaching than research.

In 2013 Eurostat commissioned a review of the procedures used to create the CIS SUFs, with a view to improving the methods and applying them to the 2010 CIS. The method used for the pre-2010 surveys assumed that business units would be easily recognisable (spontaneous recognition and intruder activity were not distinguished), and so micro-aggregated all the continuous variables on all observations to reduce the value of information obtained in the event of a successful re-identification (this would also lower the identifiability of businesses somewhat). This was the preferred alternative to distorting the identifying characteristics of the businesses, such as size or industry, but it substantially reduced the analytical potential of the dataset.

The team commissioned to review the method argued that:

- the non-statistical controls on the data (project, setting and people), plus training on outputs, provided substantial protection for the data against intruders;
- the complexity of business data provided substantial protection against accurate spontaneous recognition and against identity confirmation;
- the micro-aggregation biased the marginal analyses which were the main reason for requesting the dataset;

- employment, industry and country of ownership were identifying characteristics, but these could be put into broad categories designed to align with the typical categories used by researchers.

The source datasets had been available for some years in the ESC and in the provider countries. There was no evidence to suggest that researchers were interested in re-identifying companies, even in the small countries where big firms were more noticeable.

The research team argued that the only extant risks from the data were from researchers asserting that a particular entity had been found. The team accordingly advised that statistical disclosure control should be largely replaced by management controls: researchers were to be reminded that speculating on the nature of specific businesses was liable to (a) be wrong and (b) get them into trouble.

Micro-aggregation of turnover data was recommended in a very small number of extreme cases where it was thought that human nature might encourage researchers to make assertions or try to confirm the researcher's identity, but no other variables were micro-aggregated. The team also added a marker for whether a business was micro-aggregated or not. This suggestion caused some concern, but the team argued that researchers were always able to work out which firms had been micro-aggregated, which could lead to spontaneous recognition; putting in the marker was safer because it dissuaded them from trying to do so.

Previous versions of the data had all continuous variables micro-aggregated for all records, and each variable was micro-aggregated independently of the others. The revised method micro-aggregated less than 1% of the records, and did this for just one continuous variable (turnover). Where turnover was perturbed, other continuous variables were adjusted up or down to reduce the impact of the micro-aggregation on marginal analyses.

The result was a substantially higher quality research dataset, much closer to the SecUF. The impact on research quality was tested by running linear and non-linear regressions on both the SecUF and SUF. The results showed that the new method gave almost identical results on the two datasets, a notable improvement on the previous method. The method has subsequently been applied to later datasets.

This improvement arose solely from a change in perspective – moving from a default-closed to a default-open model. Crucial to the outcome was recognising that intruder models had no role to play in the risk scenarios once the release environment was considered, and that the risk of spontaneous recognition was best dealt with by user management.

7. Conclusion

The public good is best served by making data available for research with as little damage as possible. Data providers, 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 undergone 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 providers do raise concerns about spontaneous recognition. The authors observe this most often in relation to business data, where it seems "obvious" that any useful micro-dataset 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 data access despite there being no empirical support for the theoretical risks. With the rise of "big data" and social media, commentators have suggested that individual data may now be "recognisable".

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. Since 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 providers, 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 provider'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 a 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.

8. References

Desai, T. and Ritchie, F. (2010), "Effective researcher management", *Work session on statistical data confidentiality 2009*; Eurostat.
http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2009/wp.15.e.pdf

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.
http://www2.uwe.ac.uk/faculties/BBS/Documents/1601.pdf

Duncan, G. T., Elliot, M. and 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

Green, E. and Ritchie, F. (2016), *Data Access Project: Final Report,* Australian Bureau of Social Services, Canberra.

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", *UNECE/Eurostat Work Session on Statistical Data Confidentiality 2015*, Helsinki. http://www1.unece.org/stat/platform/download/attachments/109248612/Session%204%20-%20Various%20%28Hafner%20et%20al.%29.pdf?version=1&modificationDate=1442327222025&api=v2

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 http://www2.uwe.ac.uk/faculties/BBS/BUS/Research/Economics%20Papers%202015/1503.pdf

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

Kahnemann, D. (2012), *Thinking, Fast and Slow*, Penguin Books, London.

Mackey, E. (2013), *European Union Statistics on Income and Living Conditions (EU-SILC), case study*, 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 http://www2.uwe.ac.uk/faculties/BBS/BUS/Research/Economics%20Papers%202014/1412.pdf

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. http://www2.uwe.ac.uk/faculties/BBS/BUS/Research/General/Economics%20papers%202016/1607.pdf

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*. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_3_Schulte_Nordholt.pdf

Skinner, C. (2012), "Statistical Disclosure Risk: Separating Potential and Harm"*, Int. Stat. Rev*. v80:3 pp349–368

Spicer, K., Tudor, C. and Cornish, G. (2013), "Intruder Testing: Demonstrating practical evidence of disclosure protection in 2011 UK Census", *Work Session on Statistical Data Confidentiality 2013,* Eurostat. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_5_Spicer.pdf

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

9. Acknowledgements

This topic arose out of conversations with staff at national statistics institutes in 2015-2016. The paper is based on presentations at the 2016 Administrative Data Research Network Conference and the 2016 Conference of European Statistics Stakeholders. I am grateful to conference attendees for comments. The source presentation is http://www.ksh.hu/cess2016/pdf/cess2016_d6_5885.pdf . This is a slightly extended presentation of the conference paper. I am also grateful to Lizzie Green at UWE and to the referee for ECB Statistical Paper series; their detailed reviews improved the work substantially.

Felix Ritchie
Bristol Centre for Economics and Finance, University of the West of England, Bristol. Email: felix.ritchie@uwe.ac.uk