User-centred threat identification for anonymised microdata

Hans-Peter Hafner[1], Rainer Lenz[2] and Felix Ritchie[3]

[1] right.basedonscience

[2] Faculty of Automotive Engineering and Production, Koln University of Technology, Arts and Sciences

[3] Bristol Business School, University of the West of England Bristol

Corresponding author: Felix Ritchie, Bristol Business School, University of the West of England Bristol, Coldharbour Lane, Bristol, UK BS16 1QY. Email: felix.ritchie@uwe.ac.uk. Tel: +44 (0) 117 32 81319

Abstract: When producing anonymised microdata for research, national statistics institutes (NSIs) identify a number of 'risk scenarios' of how intruders might seek to attack a confidential dataset. This approach has been criticised for focusing on data protection only without sufficient reference to other aspects of confidentiality management, and for emphasising theoretical possibilities rather than evidence-based attacks.

An alternative, ‘user-centred’ approach offers more efficient outcomes and is more in tune with the spirit of data protection legislation, as well as the letter. The user-centred approach has been successfully adopted in controlled research facilities. However, it has not been systematically applied beyond these specialist facilities.

This paper shows how the same approach can be applied to distributed data with limited NSI control. It describes the creation of a scientific use file for business microdata, traditionally hard to protect. This case study demonstrates that an alternative perspective can have dramatically different outcomes as compared with established anonymization strategies; in the case study discussed, the alternative approach reduces 100% perturbation of continuous variables to under 1%. The paper also considers the implications for future developments in official statistics, such as administrative data and ‘big data’.

Key words: statistical disclosure control, risk management, statistical data confidentiality, data anonymisation

1. Introduction

Government bodies collecting and publishing data are increasingly required to produce research datasets from the same sources used for aggregate statistics. Allowing access to this microdata effectively leverages the investment in data collection. As data collected by government are typically confidential, the dataset is rarely released 'as is' but has confidentiality protection measures applied to it.

National statistical institutes (NSIs) carry out this function to a greater or lesser degree and have sponsored much research on reducing re-identification risk in datasets. There is a large academic literature to support such processes, as well as automatic tools such as μ-Argus (http://neon.vb.cbs.nl/casc/mu.htm) and SDCMicro (http://cran.r-project.org/web/packages/sdcMicro/index.html).

However, there is also a strong perspective about the way that the tools should be used. NSIs tend to be risk-averse [1], more comfortable (in our experience) with the 'policing' than the 'sharing' approach to data access and focused on the statistical product rather than the use to which it is put. This

leads to “best practice” models that emphasise the protection of data in extreme circumstances. We refer to this as the ‘data centred’ approach, and it dominates the literature on this topic.

In recent years a small but growing literature has challenged the data-centred approach to risk assessment [2]. The challenge is based on both theoretical grounds and on decades of empirical evidence about how intruders and researchers actually use such data files. The ‘user-centred’ approach to risk management focuses on the circumstances in which data is used (how, where, why, by whom), placing the primary emphasis on factors other than inherent risk in the data. This is effectively switching the objective function and restriction within the optimisation problem [3]: in the user-centred approach the objective is to maximise analytical validity subject to not exceeding some predefined level of data protection, and vice-versa in the traditional model.

Changing the perspective can make substantial improvements to the utility of the dataset while preserving the nature of the data. This can also satisfy

NSI objectives as the model leads to greater security and improves the prospects for positive user engagement.

This user-centred perspective was developed in the context of (remote) controlled access research facilities. In these, it is straightforward to demonstrate that the user centred approach is both more secure and more cost-effective, and the approach is increasingly seen as best practice. Such facilities have been one of the great success stories for NSIs in recent years, allowing unprecedented access to confidential data for research.

However, most research use of NSI data is still via datasets distributed to researchers to use on their own machines. Even business microdata have been disseminated. These tend to be more identifiable than individual data, so the creation of SUFs based on business surveys meeting national legislation on data protection is much more complicated. Nevertheless, DeStatis in Germany, for instance, has distributed SUFs and PUFs containing cross sectional and (since 2008) longitudinally linked business microdata [4, 5]. Despite concerns about the increasing vulnerability of distributed data [6], this is unlikely to change in the near future as there are significant cost

advantages to the NSI and benefits to society in distributing low-risk datasets, and researchers like having microdata on their desktops.

At first glance, distributing data does not seem suitable for the user-centred approach. By definition, the NSI has limited practical control over how the data is used once it leaves the NSI, and so traditionally NSIs have minimised the perceived risk in the data itself through statistical disclosure control (SDC) methods. This paper argues that this is a costly error: thinking about how researchers could use the data (in contrast to how potential data intruders could attack the data) can bring substantial gains to both the NSI and the user.

We illustrate with a case study the creation of a 'scientific use file' (SUF) from multinational business survey data. As noted above, business data is generally much more identifiable than individual data, and so the production of business SUFs is rare and much more likely to involve perturbation of the data [4]. We demonstrate that an alternative perspective can have dramatically different outcomes: in this case, from 100% perturbation of all continuous variables to perturbation of under 1% of values for just one single

variable. This change owes nothing to new statistical information, but everything to a change in perspective about risks and the use of evidence.

The next section summarises the standard approach to dataset protection. Section 3 critiques this, and proposes an alternative strategy. Section 4 is a case study in applying the alternative model to the creation of an SUF from confidential business microdata, and the resulting impact. Section 5 considers the lessons learned and the implications for wider developments in NSI outputs. Section 6 concludes.

2. Common approaches to anonymisation

The literature on statistical disclosure control (SDC) has developed over many years. It is large, coherent, and flexible. The ESSNet Handbook on SDC ([7], edited and published as [8]) was the result of several EU- and Eurostat-sponsored projects to describe the current state of the art in SDC with a general purpose review of the whole field. It succeeds largely because there is a clear and relatively uncontentious canon of results. Researchers revise models and provide analysis of the effectiveness of different methods, but the broad approach is largely unchallenged; similar approaches are described

in [9] or [10]. The seminal text of Willenborg ([11], revised 2001 and 2013) illustrates the incremental development of SDC theory.

The Handbook notes that microdata protection should be based upon knowledge of the use of the data, the access requirements, the potential for an intruder to match external datasets, and the structure of the data itself. Risk scenarios are based upon actively searching for an individual, possibly using record linkage (see [12], for example). It is possible to generate estimates of the likelihood of re-identification of an individual, given an appropriate set of assumptions. These probabilities can then be used to compare alternative data protection methods.

This approach has three near-universal features:

- A malicious 'intruder' or 'adversary' with the resources and motivation to breach data security
- A focus on worst-case scenarios
- The use of univariate measures of 'utility' lost by SDC measures

These are easily justified. If a dataset is protected against a deliberate, malicious attempt to re-identify data, it must also protect against accidental or non-malicious attempts which are less motivated. Similarly, worst-case

planning is justified as protection against less-serious cases. Finally, univariate metrics are the only objective measures as multivariate analysis involves subjective decisions about which variables to include.

The ‘intruder’ assumption provides the rationale for an attack; it is almost universal, but may not be explicit. Of the eleven papers published in a leading SDC journal *Transactions in Data Privacy* in 2017-18, three talk of ’intruders’, four of ‘adversaries’ and one of ’attackers’; two further papers build mathematical models without reference to an attack scenario.

Worst-case planning assumes that an intruder has effectively, unlimited time and resources, plus additional information to re-identify data. This ‘additional information’ is usually assumed to be a dataset containing some of the same data subjects and much the same information as the source but with the target variables (identifying or attribute, depending on the attack scenario) missing. Both datasets are assumed to be accurate; as [13] notes, this over-estimates disclosure risk but avoids introducing another level of dataset-specific subjectivity into the model. Automated tools such

as mu-Argus and sdcMicro use the source dataset as the ‘attack’ dataset, creating a pure theoretic ‘worst-case’ model of limited practical relevance.

It could be argued that the real worst case is where the intruder has some specific private information on respondents, but this is unhelpful. By its nature, that additional information is unknowable, which means that it cannot be modelled and thus every SDC procedure might be differently affected by it in unknown ways. For the methodologist, this makes it impossible to prove the effectiveness of any particular technique, or even to demonstrate superiority over other techniques. For the same reason, spontaneous recognition is used for pedagogical purposes but not for scenario modelling [14].

Whilst changes in the probability of detection can be described in a straightforward manner, changes in the utility of the data are harder to quantify as this depends upon the likely uses of the data. Sophisticated analyses on the effect of various anonymisation methods - applied to both discrete and continuous variables - on the analytical validity of microdata can be found in [15] and [16]. However, much of the discussion of ‘utility’ focuses

on univariate measures, such as perturbation of mean, median, and percentiles, or distributional measures. For example, Fletcher and Islam [17] present a range of complex multi-variable metrics for utility loss, but these are still essentially predicting low-dimension perturbed tables (to be fair, the authors do not claim to be defining definitive measures, but rather an additional set of useful indicators; they acknowledge that the choice of measure must reflect the data manager's goals).

In summary, the theoretical basis for SDC is well-founded, coherent and largely uncontested. The common agreement on the use of intruders, worst-case scenarios and univariate impact metrics has allowed for a consistent treatment of methodologies, so that the various pros and cons of different methodologies have been repeatedly analysed. This in turn has encouraged the development of software to automatically provide objective estimates of disclosure risk and the effect of protection measures.

However, problems occur when applying this theoretical foundation to practical problems of data management.

3. Critique of common perspective

There are three major concerns about the way microdata protection is implemented in practice. Two can be seen as failures to use evidence; the third is a case of failure of the theoretical framework for decision-making.

3.1 Focus on data protection

Microdata sets are classified into 'public use files' (PUFs, available without restriction to anyone), 'scientific use files' (SUFs, available to accredited users only), or 'secure use files' (SecUFs, available to accredited users within an environment controlled by the NSI; sometimes referred to as 'controlled access files'). It is questionable how much attention is paid to these different surrounding conditions. Implicitly, most SDC models assume the dataset is public, as the intruder threat is unlimited and there is little or no discussion of non-statistical controls such as licensing or data management.

For SecUFs and SUFs the malicious intruder is a difficult case to make. Good practice requires the removal of direct identifiers (names etc.) from such files, so identification is only possible indirectly, implying some effort on the part of the researcher. For SecUFs this effort is monitored and can be limited.

For SUFs, the effort is not monitored but still required [12]. In both cases, accreditation procedures and contracts are used to ensure appropriate use of data.

In both cases it is clear that, if intruder threat is a genuine risk, the problem lies with accreditation procedures and not with anonymisation. As demonstrated below, restrictions on the data to guard against intruders are likely to be ineffective and damaging to researchers. In contrast, better accreditation tackles directly the problem, the non-trustworthiness of users. Accreditation is also easier to manage: speculating on the possibilities of matching databases is much more nebulous than checking whether a researcher genuinely has a social science degree and is employed by a university.

There is also an indirect benefit: users are wary of the impact of anonymisation on quality; for example, in the case of the CIS data discussed below, researchers reported that the confidentiality protection had made the data too unreliable for genuine research use. Replacing anonymization by accreditation, at least in part, gives users more confidence in the analytical

validity of results and so it is more likely to make them accept the necessity of a proportionate level of detail reduction.

Such evidence as there is suggests that intruder modelling is highly unrealistic. There are no cases (to the authors' knowledge) of *malicious* misuse of SecUFs or SUFs in the ways identified by standard risk scenarios. There is ample evidence of researchers *making mistakes*, or *circumventing procedures* - but not to deliberately de-anonymise the data. The deliberate misuses were all the result of researchers ignoring or trying to reorganise processes for their own convenience.

Even such non-malicious outcomes are rare. Over ten years, one controlled facility saw three deliberate acts of misuse and another ten or so genuine mistakes, set in the context of some six thousand user visits; discussions with managers of SUF and SecUF releases across the world suggests that this outcome is the norm. The most egregious case involved a group of researchers downloading a dataset piecemeal over a number of months through a flaw in the control systems; but it is worth noting that the

researchers did not do this to re-identify data, but for the convenience of having data on their desktops.

It could be argued that no NSI would willingly share information on a deliberate breach because of the poor publicity, but the relatively small size of the international data protection community militates against such a case; when problems do occur they are freely discussed amongst the community in the spirit of improving outcomes. One could also argue that successful malicious breaches have occurred but remain undiscovered, which is theoretically true but not practically helpful.

In summary, for SUFs and SecUFs empirical evidence suggests that factors other than protection of the data dominate the likelihood of successful protection; such non-data control measures have a forty year record of demonstrable effectiveness.

For PUFs, it could be argued that intruder threat is a genuine risk, as potentially it only needs one person in the world to have sufficient malice or prurience to try to breach confidentiality protection. As the PUF is either openly circulated (e.g. simply by download) or delivered with low restriction

(e.g. requirement to register as user before download), the potential attackers include not just all living individuals, but all future attackers and in all future states of the world.

However, protecting against any attack by any person at any time in the future is an impossible standard, and no law requires it. In practice, all NSIs explicitly or implicitly accept the "reasonableness" argument, and make subjective judgments about what is proportionate. In such judgements the value of the data, highly perturbed and/or hidden becomes relevant. The German 'de facto anonymisation' rule makes this explicit in law: a dataset is deemed to be non-disclosive if the cost of extracting identified information from the data exceeds the value of that information. More recent legislation, such as the EU General Data Protection Regulation 2016 or the UK Digital Economy Act, have moved away from defining breaches of the law in terms of outcomes (such as identifiable datasets); lawfulness is now embodied in the procedures which govern the data access, not in the data itself.

3.2 Worst-case scenarios and spurious objectivity

Using ‘worst-case scenarios’ makes sense in the context of methodological research, where the aim is to compare methods using a common framework wherever possible. Such assumptions allow the *relative* effectiveness of methods to be assessed fairly, which is essential for developing understanding of the effect and effectiveness of different techniques. It does not however follow that worst-case scenarios must be used in practice, for four reasons.

First, any NSI must balance costs against benefits; otherwise, the confidentiality problem is easily solved by not releasing the data, full stop. Ideally, the full range of expected costs and benefits would be assessed, but focusing on the worst-case scenario for the cost requires a corresponding increase in benefit, preventing the release of more and better microdata for the scientific community. As noted above, no law requires confidentiality protection to be valued against all other criteria; implicitly, and increasingly explicitly, legislation requires that data owners justify their decisions as ‘reasonable’.

Second, there is no evidence to suggest that typical worst-case scenarios ever manifest themselves. Consider the popular assumption that (almost) exactly the same data as that held in the dataset is available to an intruder. It is well known that there are large differences between data from official statistics and external commercial databases (eg. [12], [18]). Although the aggregate statistics produced from such datasets may be similar, at the record level there is a poor correlation between units even when an exact match is possible [19]. Gregory [20] demonstrated that matching record-level personal data to social media was much less successful than predicted. Studies ([6], [21]) note that the use of administrative data by NSIs does mean that an external agent has, potentially, access to exactly the source data underlying the protected dataset, However, there are many practical problems with this theoretical perspective; indeed, much of the empirical research on data linkage is focused around *improving* match rates rather than reducing identifiability. In summary in practical cases there is always some natural protection for the data that adds to the protection achieved by anonymisation procedures.

Despite this, NSIs often try to replicate the worst case scenario favoured in academic papers. It can be time consuming and expensive to generate a realistic external data source. For a statistical match, the commercial data have to be purchased from different sources, and the identifiers often have to be harmonized manually; for a manual test such as [20], there is the difficulty of getting reviewers, and the omnipresent criticism that a 'pass' for the anonymisation might be because the reviewers were not expert enough or did not have enough resources.

Hence, a common technique is to match the anonymized data to the original survey data, with the latter pretending to represent 'external' data; this is how software such as μ-Argus creates risk assessments. This fabricated worst case scenario clearly should not be treated as a 'real-world' test, but in practice the risk estimates generated by such models can be given substantial weight.

Third, worst case scenarios are typically not that: they are the *mathematically tractable* worst cases. A realistic worst case might the

unplanned release of some of the original data on the internet, against which no anonymisation can protect.

Finally, worst-case scenarios are no less subjective than other models. Skinner [22] argues that claims of objectivity in risk assessment are misleading; the framing of the risk assessment is decided by the NSI on subjective criteria. For example, the ESSNet Handbook describes a potential 'conservative and worst case scenario' with only one known external data source being used for matching and with design, but not response, weights available. Clearly, both assumptions are debatable, and an NSI adopting these assumptions is making a subjective decision.

Once it is recognised that worst-case scenarios are (a) inefficient for society and not required by legislation (b) not supported by evidence (c) mathematically convenient rather than true 'worst case' and (d) as subjective as any other modelling base, their use in decision-making comes into question.

3.3 The default position

The default perspective of most NSIs is *defensive*: no data can be released unless they can be shown to be 'safe'. The protection of the NSI is the key objective. However, the NSI could take the *public benefit* perspective: that data will be released unless they present a demonstrable disclosure risk. This does not conflict with meeting national legislation on data privacy – legal responsibilities are unchanged - but public benefit is now the objective.

Many organisations formally support the 'public-benefit' approach, but this does not necessarily happen on the ground. One author regularly addresses groups of data professionals, and, in shows of hands, respondents typically overwhelmingly agree that the public benefit perspective is preferable; however, when asked about their organisation, similar numbers believe that the defensive perspective is their organisation's normal position. Functionally, "release unless not allowed to" and "do not release unless allowed to" are identical in legal and statistical terms; it is in the psychology of the data controller that they differ. NSIs typically have insufficient user input to influence the discussion and overcome security concerns (the 'diffuse benefit and concentrated cost' often associated with lack of government action [23], [24]). Although in theory both positions should

make the same recommendations about data access, [3] demonstrates how these two perspectives will, in operational situations, generate different outcomes, with the former almost certain to restrict data access much more.

This defensive perspective is reflected in the lack of discussion in meetings and the literature. A major gathering for SDC researchers, the biennial Worksession on Statistical Data Confidentiality (WSC), has until recently taken an almost exclusively defensive approach. In 2013 two sessions on data access were organised, with only ISTAT [22] taking a user-centred approach; all other papers explained how they were 'opening up' data access (that is, the default is 'no release') and this should be seen as a bonus for users (this may be changing; the 2015 WSC was also largely defensive, but devoted a half-day to presenters with a default-open perspective; and the 2017 WSC presented user-centred papers across the range of topics).

The use of unrealistic scenarios is a consequence of this defensive stance, and arises from a misunderstanding of legal liability. Unrealistic assumptions are defended on the grounds that all practical measures need to be taken to protect the data (which, of course, can also lead to excessive caution over

statistical aggregates). This is unlikely to be true. All statistical legislation leaves the level of protection as something to be determined in specific context: for example, the German construct of De Facto Anonymity (German Federal Statistics Law §16(6)). Legislation does not require worst-case planning, recognising that it is unlikely to be good for society: designing strategy based on extreme hypothetical outcomes imposes costs on society which a more balanced view of likely outcomes would avoid.

More importantly, the most recent legislation (such as the New South Wales Public Data Sharing Act 2016, the UK Digital Economy Act 2017, and the European General Data Protection Regulation which came into force in 2018) explicitly allows a range of non-statistical measures to be included when assessing confidentiality protection. In a world of multi-dimensional control options, extreme caution in one single domain is easily challenged in court.

A more helpful discussion might be “what does the spirit of the law intend?” Here, laws tend to be more explicitly relativist, making reference to ‘reasonable’ expectations, and balancing risks and benefits. This is more likely to be explicitly pro-release (for example in the current European

Commission regulation covering data access, or in the UK Freedom of Information Act). In short, while the public-benefit and defensive perspectives are both likely to be consistent with the letter or the law, the public-benefit argument is more likely to be consistent with the spirit of the law.

4. Case study of an evidence-based risk assessment: the 2010 CIS

The previous section noted the problems of the traditional approach to anonymisation: a focus on theory rather than evidence and on data rather than environment. An alternative perspective is provided by the EDRU framework: evidence-based, default-open, risk-managed, user-centred [26], which tries to address the above criticisms of the traditional approach. We now use a case study to show how such an approach can have radically different outcomes.

The dataset used in the case study is the Community Innovation Survey (CIS), a business survey carried out in all EU countries. Eurostat distributes a subset of country files, anonymised as scientific use files for research purposes. Uses have been very small and the perception exists among the research

community that the existing anonymisation method has created, at best, a teaching dataset rather than a research resource.

In 2013 Eurostat commissioned a review of the protection strategy to create the 2010 CIS SUFs. Whilst the review recommended a number of significant changes (described in [27]), in the following we focus on the risk scenario and its consequences for protection mechanisms.

Detail reduction for microdata typically follows five stages:

1. Identify user needs
2. Identify the user environment and risks
3. Evaluate risks
4. Determine relevant risk scenarios
5. Apply protection measures

The case study follows the same five stages, but the EDRU approach tackles each stage in a different manner.

4.1 Identify user needs

The study analysed 11 research papers using the CIS in SUF and SecUF form in different countries (official documents from NSIs or other government

departments were also analysed but these consisted exclusively of simple tabulations). In addition, the authors could draw on observations from nine years running a controlled environment, where researchers could carry out unrestricted analyses on the CIS. Finally, a non-systematic Google Scholar search was carried out. These confirmed that the overwhelming use of the CIS by researchers (as opposed to government agencies who hold the source data) was marginal analysis, particularly linear and non-linear regression. A key objective was therefore to retain validity in marginal analyses.

4.2 Identify the user environment and risks

Users have the CIS SUFs under their control, with instructions to store the data safely and not attempt to re-identify companies, but with no effective mechanism for monitoring whether the instructions are complied with. There is no evidence of malicious use of SUFs by genuine researchers. There is evidence of accidental and deliberate misuse which has the consequence of breaching confidentiality rules or procedures [28].

For business data, the most identifying information (company size and industrial sector) is also the most analytically useful. As a result, the most

commonly identified risk is spontaneous recognition of outliers; for example, a researcher provisionally recognising the largest company in a particular industrial sector, and then either publicly speculating on the identity of the firm or trying to augment or compare the SUF with data from external sources.

However, this is a management problem not a statistical one [14], best addressed through licensing and training. Note also that it is not a risk that a researcher spontaneously notes the characteristics of an observation and muses on the company identity but does not follow up - there has been no disclosure to an unauthorised person, and no deliberate attempt to identify a company.

A second risk is that the researcher may circulate the data inappropriately. This may be deliberate: the researcher may share the data with a colleague who has not signed the appropriate data access licence. It is more likely to be accidental: for example, on a shared folder without checking who has access permissions, or taking an authorised backup on a memory stick and then losing it.

The third area of risk is the output produced by the researcher, where a mistake on the researcher's part might lead to the publication of identifiable information. In general, outputs from genuine research are low risk, but there are a large number of categorical variables in the CIS and the interest in them makes the potential for disclosure by differencing larger.

There is the risk of group disclosure. The categorical variables in the CIS make saturated or empty cells more likely: for example, there may be many cells in a table where all companies undertake a specific form of innovation. However, most of the CIS categorical variables are targets, not identifiers; someone may want to know "has your company made a product innovation in the last three years?", but this is not information that can help to identify the respondent.

Finally, there is always the risk from a misperceived output; for example, a naïve reader of a paper could assume that a statistic refers to a single company even if it does not. In this case, the risk is not to confidentiality but to the reputation of the organisations collecting and distributing the data.

4.3 Evaluate risks

Re-identification risk arises from publicly available classification data (company size, head office location etc) and from extreme values in continuous attributes, such as very high turnover. However, practical experiments done by the authors and others in this field (for example, [18], [29]) suggest that exact matching on continuous variables is not a practical concern, although a broad search on industrial classification and location might be more effective. Identification may arise from matching to external databases, but the sampling frame is the Eurostat-compliant business register, which is designed to reflect economic activity with *statistical* accuracy, not *financial* accuracy. As is well-known (eg. [19]), NSI business registers are difficult to reconcile with publicly available accounting information which makes extensive use of financial engineering. These factors provide considerable uncertainty about which companies are included in the data, and in which organisational form.

Much of the data in the CIS, while useful for research, has a low disclosure value. For example, it is a breach of information supplied in confidence to be able to identify that Company X has engaged in product innovation over the period 2008-2010; but it is of negligible commercial value (the 'innovation' is

not specified and could be anything from repackaging to a complete new product). Much more detailed, and useful, information is available in company accounts, patent applications, press releases, and so on, all of which will directly identify the company. Understanding the data requires access to the metadata (such as the original questionnaires); this is not impossible for an unauthorised person to find, but it adds an extra stage. In short, exploiting record-level data requires considerable work for relatively little value, and so is unlikely to be a target for hackers (or curious individuals finding a memory stick on a train).

In summary, re-identification is unlikely to have sufficient certainty to be worthwhile: a successful and informative match is theoretically possible but the practical problems are large, and the value dubious. Most importantly, matching requires the researcher to actively search for the company; it is not an outcome of spontaneous recognition. The SUF licence agreement forbids attempting to identify any respondent; evidence on researcher behaviour suggests this is credible. Therefore, it appears that the risks of deliberate disclosure associated with researcher inquisitiveness are of a very low order.

4.4 Determine relevant risk scenarios

This led to three credible risk scenarios:

I) A researcher publishes a magnitude table with one or two observations in a cell
II) A researcher comments on the dominance of one unit in some table's cell
III) A researcher comments on the dominance of one unit in the dataset

These are all expected to arise as a result of error on the part of the researcher who doesn't intend to publish confidential information. This differs substantially from the usual risk scenario which assumes deliberate action to re-identify companies in the *microdata*. The scenarios used relate to mistakes in the *outputs* of the researchers.

4.5 Apply protection measures

The scenario analysis suggested that the protection measures need to guard against human error in interpretation and publication. This was achieved largely by[1]

- grouping head office location
- banding employment
- averaging the turnover values in the upper size class by industry and country (microaggregation) if, and only if, the company could be placed in that size class with certainty and if it dominated the class

The user documentation stated clearly that the data had been adjusted, to maintain statistical validity but to reduce certainty over the value or identity of any specific observation. A microaggregation marker was added

---

[1] The detailed analysis and the full set of measures taken is described in [27]. This limited-circulated document may be requested from Eurostat. The description given here has been approved for a general scientific audience.

to the dataset to emphasise this point. The research team also created new variables for employment and turnover growth to support analysis.

5. Impact of the revised risk assessment strategy

As noted above, the 2010 CIS methodology review recommended a wide range of changes to the anonymisation strategy, not all of which are relevant here. However, the way the risk scenarios were defined had important implications for the anonymisation strategy.

The previous anonymisation strategy stated that deliberate misuse was not deemed to be a risk. However, the logic of this position was not followed through: no other explicit risk was identified, and yet all observations were deemed to be potentially problematic. Disclosure risk was to be addressed by microaggregation of all continuous variables and the coarsening of categorical information (global recoding), which, it was argued, also reduced the need to test for and address dominance problems. The conceptual framework used in that case was defensive: ‘apply protection until it is safe to release’.

In contrast, the revised risk scenario implied a very different protection model. As the key risk was identified as accidental disclosure, only measures to tackle dominance and small cell count were put in place, apart from a global recode of employment. In effect, only observations at risk were perturbed; the conceptual framework was 'apply protection only if it is demonstrably necessary; otherwise release'. Moreover, in deciding observations at risk, the team took account of (1) the known disparity between published and surveyed employment data, and (2) the sampling rate; both of these provide additional arguments as to why the data was inherently safe.

Microaggregation was used to implement the anonymisation, with some adjustment of other continuous variables to maintain covariances with perturbed turnover data. Table 1 below summarises the old data-centred and new user-centred approaches.

[Table 1 here]

Table 2 displays the impact of the adjusting only at-risk records defined on the user-centred method:

[Table 2 here]

Overall only 0.53% of turnover values were microaggregated; this number includes neighbours of at-risk observations, who were not at risk but microaggregated to provide cover for their neighboured at-risk responses. There was some small effect on means; maxima and minima were affected as these are single values which microaggregation will perturb by design. However, the medians and most percentiles were unaffected by the changes. The project team also carried out linear and non-linear regressions and found that the anonymisation procedure changed coefficients estimates by under 5%, for all estimates significant at the 10% level.

The method was accepted by Eurostat and approved by participating Member States after a methodological review. The 2010 CIS was released for users in 2016. The method was also subsequently applied to the 2012 CIS microdata, also now available. Unfortunately, there is no evidence that the improved data quality has increased usage: demand has been stable since 2016 but it remains one of Eurostat's lesser-used datasets.

6. Discussion and conclusion

Microdata protection is a well-established mature field with a great deal of advice for NSIs trying to make confidential data accessible to a wider audience. However, users often express concerns that data protection occurs without consideration of the user experience. More recently the data protection community has also begun to question the profoundly conservative outlook found in NSIs[2]. Current research into confidentiality protection is still focused on the data-centred worst-case statistical analysis which has dominated thinking for the last half–century.

This paper has argued that rethinking the problem may go a long way towards resolving difficulties, as well as being more in tune with the spirit of recent legislation. This has been repeatedly demonstrated in SecUF contexts, and to some individual SUFs (eg 16]), but this is the first time it has been applied systematically to a business data SUF. Anonymising distributed microdata appropriately does present more statistical challenges, but the core messages still hold: (1) understand the non-statistical risks first, and the remaining statistical measures are likely to be both safer and substantially

less harmful to the data; and (2) hard empirical evidence is a much better base for decision making than unrealistic worst-case scenarios.

The example discussed, of creating scientific use files for the European CIS data, shows that a change in attitude can have significant consequences. A business dataset was used as the model because the high identifiability of business microdata leads to a high degree of perturbation in the traditional model.

No new methods were developed: protection was a combination of two established methods, recoding and microaggregation. The difference came in the default perspective of the research team; the use of evidence in assessing disclosure risk; a public-benefit interpretation of what counted as 'reasonable' protection; and an explicit allowance for non-statistical protection measures in the access environment. The end result was a dataset with more protection for the most risky observations than under the previous method, but with much less impact on data usage (very little of the data was perturbed at all). In addition, the strategy was also able to tackle dominance problems which had not previously been resolved. The 'do not

disturb' strategy of Ichim ([16], [30]) came to a similar conclusion about the ability to improve both security and utility by changing the perspective.

Future trends in data are moving away from surveys to administrative data sources and social media, which present new problems. For example, PUFs based on administrative data may be re-identifiable by administrative staff in the supplier organisation who have access to the original data [3]. In November 2015 a workshop on SUFs from linked social data was held in Berlin at the German Institute for Economic Research (DIW) on risk scenarios. The opinions expressed were very diverse, and the meeting concluded with an agreement only that systematic research on this topic is required (see http://www.diw.de/suf-workshop2015 for the presentations and a summary in German; the summary discusses Big Data on page 5).

Changes in data use and availability imply that the problem of matching to external databases will become much more prevalent, at least in theory, but the importance is much less obvious. It is clear from this paper that simply reducing content in NSI datasets to prevent theoretical problems is likely to produce data of increasingly unacceptable quality. Given the amount of

perturbation needed to protect against matches when the range of potential matches is continually increasing, non-statistical protection mechanisms grounded in evidence may be a more productive route forward [31]. This is certainly in keeping with a larger role for non-statistical protection in recent laws such as the European General Data Protection Regulation.

Acknowledgements

We are grateful for comments received from referees and participants at the 2014 Conference of European Statistics Stakeholders. This work draws on an independent research project carried out under a framework contract on methodology for the European Commission (Eurostat). This paper was written after project completion, and no inferences should be drawn about Eurostat's policies or views on the matters discussed. All opinions expressed are those of the authors. All errors and omissions remain our own.

References

[1] Hafner H-P., Lenz R., Ritchie F and Welpton R. 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; 2015. https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2015/0/Paper_9_Session_4_-_Various__Hafner_et_al._.pdf

[2] Tam S. and Kim J. Big Data ethics and selection-bias: An official statistician's perspective. Statistical Journal of the IAOS 34 (2018) 577–588 DOI 10.3233/SJI-170395

[3] Ritchie F. Access to sensitive data: satisfying objectives rather than constraints. J. Official Statistics 2014; 30(3):533-545, September. DOI: 10.2478/jos-2014-0033

[4] Lenz R, Zwick M. Business Micro Data in Germany: Access and Anonymization, presented at the International conference of the Royal Statistical Society ‚Statistics in a changing society – 175 years of progress',

Edinburgh 2009. Published in: Journal of Applied Social Science Studies 129 (4): 645-653, 2009.

[5] Brandt M, Lenz R, Rosemann M. Anonymisation of panel enterprise microdata – Survey of a German Project, in: Domingo-Ferrer J, Saygin Y (Eds.): Privacy in Statistical Databases, Lecture Notes in Computer Science, vol. 5262, Springer Heidelberg, 139-151, 2008.

[6] Ritchie F, Smith J. Confidentiality and Linked Data. 2018 in [31]

[7] Hundepool A, Domingo-Ferrer J, Franconi L, Giessing S, Lenz R, Longhurst J, Schulte Nord-holt E, Seri G and De Wolf P. Handbook on Statistical Disclosure Control, ESSNet SDC, 2010. http://neon.vb.cbs.nl/casc/.\SDC_Handbook.pdf

[8] Hundepool, A, Domingo-Ferrer, J, Franconi, L, Giessing, S, Schulte Nord-holt E, Spicer K. and De Wolf P. Statistical Disclosure Control. Wiley Series in Survey Methodology. New York, 2012.

[9] Domingo-Ferrer J, Sánchez D. and Soria-Comas J. Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based

Inter-model Connections. Morgan and Claypool: Synthesis Lectures on Information Security, Privacy, and Trust, 2016.

[10] Templ M. Statistical Disclosure Control for Microdata: Methods and Applications in R. Springer-Verlag: Berlin, 2018.

[11] Willenborg L. Statistical Disclosure Control in Practice. Springer: New York: Lecture Notes in Statistics, 1996.

[12] Lenz R. Measuring the disclosure protection of micro aggregated business microdata - an analysis taking as an example the German Structure of Costs Survey. J. Official Statistics 2006; 22 (4):681-710.

[13] Willenborg L. de Waal T. Statistical Disclosure Control in Practice: v. 111. Spinrger: New York: Lecture Notes in Statistics, 2013.

[14] Ritchie F. Spontaneous recognition: an unnecessary control on data access? European Central Bank Statistical Papers 2017; no.24, August

[15] Ronning G, Sturm R, Hoehne J, Lenz R, Rosemann M, Scheffler M and Vorgrimler D. Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten. DeStatis: Statistik und Wissenschaft, Band 4, 2005.

[16] Ichim D. Community Innovation Survey: a Flexible Approach to the Dissemination of Microdata Files for Research. Proceedings of the Q2008 European Conference on Quality in Official Statistics 2008.

[17] Fletcher S and Islam MZ. Measuring Rule Retention inAnonymized Data – When One Measure Is Not Enough. Transactions On Data Privacy 2017; 10: 175–201.

[18] Hafner H-P. Die Qualität der Angriffsdatenbank für die Matchingexperimente mit den Daten des KSE-Panels 1999 – 2002. Mimeo. IAB: Berlin 2008.

[19] Evans P and Ritchie F. UK Company Statistics Reconciliation Project: final report. Department of Business Enterprise and Regulatory Reform 2009; URN 09/599.

[20] Gregory M. DECC's National Energy Efficiency Data Framework – Anonymised Dataset. Presentation to the Turing Institute 2014. http://www.turing-gateway.cam.ac.uk/sites/default/files/asset/doc/1608/Gregory.pdf

[21] Ritchie F. 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 2016.

[22] Skinner CJ. Statistical disclosure risk: separating potential and harm. International Statistical Review 2012; 80(3):349-368

[23] Moore MH. Break-Through Innovations and Continuous Improvement: Two Different Models of Innovative Processes in the Public Sector. Public Money and Management 2010; 44 January.

[24] Ritchie F. Resistance to change in government: risk, inertia and incentives. Working papers in Economics no. 1412, University of the West of England, Bristol, 2014.

[25] Ichim D and Franconi L. Istat experience on releasing multiple microdata files stemming from the same survey. Work session on statistical data confidentiality 2013. Luxembourg: Eurostat 2014.

[26] Green E and Ritchie F. Data Access Project: Final Report. Australian Department of Social Services 2016 June. http://eprints.uwe.ac.uk/31874/

[27] Hafner H-P, Lenz R and Ritchie F. Revised creation of CIS Scientific Use Files: Final Report. Eurostat: Luxembourg June 2014.

[28] Desai T andRitchie F. Effective researcher management. Work session on statistical data confidentiality, Bilbao 17-19 December 2009, Luxembourg: Eurostat, 2010.

[29] Bauer T, Bachteler T, Bender S, Huber M, Schnell R, Dundler A and Engel D. Why are firms treated in a different way than individuals? – Results of a re-identification experiment with real data. Discussion paper. Ruhr-University Bochum 2009.

[30] Ichim D. Microdata Anonymisation of the Community Innovation Survey Data: A Density Based Clustering Approach for Risk Assessment. Dokumenti Istat 2, 2007 http://www3.istat.it/dati/pubbsci/documenti/Documenti/doc_2007/2007_2.pdf

[31] Roarson G, editor. Privacy and Data Confidentiality Methods – a National Statistician's Quality Review. Office for National Statistics:

Newport https://gss.civilservice.gov.uk/guidances/quality/nsqr/privacy-and-data-confidentiality-methods-a-national-statisticians-quality-review/

Tables

Table 1 Comparison of old and revised risk assessment

	Data-centred method	User-centred method
Variables at risk	All continuous variables	Employment and turnover
Observations at risk	• All observations	• Employment: all • Turnover: only responses in small groups with dominant observations, non-sampled, with certainty over employment band
Disclosure measures applied	• Employment: banding • All other continuous variables: microaggregation, independently for each variable	• Employment: banding • Turnover: microaggregation on at-risk observations and neighbours • Other continuous variables: adjustment consistent with microaggregation

Table 2 Impact on turnover of user-centred approach

Country	Records changed	Change in mean
BG	0.30 %	0.33 %
CZ	1.44 %	1.28 %
EE	1.64 %	1.11 %
ES	0.17 %	0.30 %
FR	0.24 %	0.38 %
HR	0.09 %	0.00 %
HU	0.91 %	0.95 %
IE	0.62 %	0.33 %
LT	1.23 %	4.08 %
LV	0.97 %	0.63 %
NO	1.72 %	2.90 %
PT	0.91 %	1.70 %
RO	0.41 %	0.21%
SI	1.28 %	1.00 %
SK	2.44 %	2.66 %
Total	**0.53 %**	

Notes: Number of records is percentage of country totals. 'Change in mean' is the weighted average difference in means across all the employment size-industry sector domains in a country.