Journal of Privacy and Confidentiality (2013) **5**, Number 2, 111–132

Intruder Testing on the 2011 UK Census: Providing Practical Evidence for Disclosure Protection

Caroline Tudor*, George Cornish†, and Keith Spicer‡

1 Introduction

With the recent push towards sharing greater amounts of information, the pressure is on National Statistical Institutes (NSIs) to publish more detailed datasets to broader audiences. It is of parallel importance for any such organisation to respect and protect the confidentiality of respondents' data. Assessing the risk of identification in a dataset is a challenging task and there is much in the literature on how to measure this theoretically: Hundepool et al. (2012) devote entire sections of their statistical disclosure control (SDC) book to the various methods for different types of data. However, surprisingly little is known about how easy is it in practice to identify real people from real databases, especially when they have been subjected to some form of disclosure limitation such as data swapping, described in more detail in Section 2.

Willenborg (2012) offers a perspective from within an NSI and questions the usefulness of the rather abstract approach that seems to have developed over time. He argues that a better understanding of 'intruders' and how they disclose information would allow for better protection of data without straying into overprotection. He suggests that NSIs could more usefully take an empirical approach. This paper attempts to address this issue presenting a case study of 'intruder testing' on UK Census data.

Intruder attacks on public data are widely cited in the privacy literature. There has been considerable publicity of the success of intruder attacks on non-government databases such as AOL's release of anonymous search data, and Netflix's release of 100m video-rental records, both in the U.S. In 2006, Netflix launched a competition offering million-dollar prizes to contestants who could provide algorithms that predicted viewers' movie ratings with better predictive accuracy than their own. Contestants were given access to a training dataset containing millions of movie ratings.[1] In order to protect the privacy of customers, *some of the rating data for some customers in the training and qualifying sets were deliberately perturbed in one or more of the following ways: deleting ratings; inserting alternative ratings and dates; and modifying rating dates.*[2] However, researchers from the University of Texas were able to demonstrate disclosure by combining the collection of movie ratings with public reviews published by a "few

*Office for National Statistics, United Kingdom, mailto:caroline.tudor@ons.gov.uk.

†Office for National Statistics, United Kingdom, mailto:george.cornish@ons.gov.uk.

‡Office for National Statistics, United Kingdom, mailto:keith.spicer@ons.gov.uk.

[1] http://www.wired.com/images_blogs/threatlevel/2009/12/doe-v-netflix.pdf.

[2] http://www.netflixprize.com//community/viewtopic.php?id=1537.

dozen" people in the Internet Movie Database (IMDb). They were able to identify two of the users in the Netflix data (Narayanan and Shmatikov, 2006). Similarly, the AOL case in 2006 related to the release of detailed search logs on a large number of AOL users. The database did not include names or user identities but listed only a unique ID number for each user. The release was also intended for research purposes. However, *The New York Times* was able to locate an individual known as 'AOL Searcher No. 4417749' from the anonymised search records by cross referencing them with other information.[3] There has also been discussion of attacks on genome-wide association study databases. Despite the fact that these datasets were supposed to be 'anonymous,' intruder attacks proved otherwise.

A formal intruder test carried out by data providers before release of data may be a powerful way of truly revealing whether a dataset can in fact be deanonymised. The idea of intruder testing in order to learn more about intruder behaviour has been considered in some form within NSIs—at least anecdotally. In the literature, Paass (1988) demonstrates an early simulation of intruder testing using statistical models based on the German microcensus to create synthetic target persons. He studies various disclosure scenarios such as the tax investigator or the journalist, considering the variables they might have access to. However, there has been little reported on formal attempts to use real intruders and to put this into practice in any systematic way. The only examples that could be found were that of Simpson (2011) and O'Hara and Whittall (2011). The O'Hara work involved individual-level reoffending and sentencing data from the Ministry of Justice; students acted as intruders, testing anonymised data prior to online release. There was identification as a result of matching the profile of an offender named on a local news website. Although the offender was not unique, because there was a match certain deductions could be made about his/her reoffending data. This led to the data being further aggregated before final release. Simpson's (2011) work also involved students who were tasked with examining public data on *data.gov.uk* as part of an assignment on data security. They were given free reign to explore any privacy concerns in the datasets provided. They discovered that various datasets were poorly anonymised and the paper presents an actual example of linking statistical and personal information.

This paper provides insight on how to proceed with the task of intruder testing using real intruders and on a large-scale real dataset, in a structured way. Our case study is an exercise which we believe no other NSI has carried out, on real census data and prior to publication. This exercise carried some risk—it could have generated a very different outcome if intruders were able to correctly identify all persons and then infer correct additional information about them (though one could argue that the impact would have been much less than a real disclosure claim post-publication that was shown to be correct). Despite the fact that our case study is based on a UK example, it is intended that this paper exemplify use of intruder testing as practical evidence for justifying means of disclosure control.

[3] `http://www.nytimes.com/2006/08/09/technology/09aol.html?pagewanted=all&_r=0`

1.1 Why Carry Out an Intruder Test?

In recent years, various initiatives relating to data transparency have drawn attention to the nature of data protection. In the UK, the Freedom of Information Act 2000 created a public 'right of access' to information held by public authorities.[4] This preceded the concept of 'open data,' a UK Government initiative which held the principle that certain data should be freely available for everyone to use as they wish, embodied by *data.gov.uk* bringing together thousands of public datasets in one searchable website. The underlying theme was to promote transparency in government, supporting more informed choices in policy making and improving trust.

The goals of open data in the UK are similar to those around the world. For example the U.S. Government launched a website—*data.gov*—in 2010 also to implement principles of transparency. Many other countries have launched similar sites including Canada,[5] Australia,[6] and Spain.[7] As of April 2013, New Zealand is planning a big data hub,[8] a massive expansion of data-sharing between ministries and agencies, backed by changes in the law. All of these initiatives have led to data coming under greater scrutiny for levels of release: users want all the data if possible and public organisations have to provide appropriate justification for restricting levels of access.

The disadvantage of solely theoretical arguments to justify disclosure control practices was shown in a 2005 case involving a request under the Freedom of Information Act (FOIA) by the ProLife Alliance. They requested abortion statistics from the UK Department of Health on certain categories of abnormality. In response, the Department of Health refused to disclose the suppressed figures on the grounds that the true counts were less than a pre-determined threshold of ten (zeros could be published) and subsequent concern about potential 'attribute disclosure.' Challenged by the ProLife Alliance at the High Court, the judgement laid down in 2011 was that the disclosure controls set out in the guidance were overly cautious in some circumstances,[9] and they concluded that in this case the detailed abortion statistics were not personal data. The fact that the Department of Health could give no practical evidence to demonstrate that individuals might be re-identified as the rationale for obscuring these data was key to this decision.

Commentators have since argued that *'a more systematic and critical analysis of whether a data asset is truly disclosive or not should be established in NSIs'* (Jackson, 2012). Willenborg (2012) writes *"if intrusion is better understood, data protection will be more effective. It would surpass intuition and judgements, and in the best case justify them."*

[4] `http://www.legislation.gov.uk/ukpga/2000/36/section/1`.

[5] `www.data.gc.ca`.

[6] `www.data.gov.au`.

[7] `www.datos.gob.es`.

[8] `http://www.radionz.co.nz/news/national/132496/govt-considers-new-\%27big-data\%27-hub`.

[9] `http://media.dh.gov.uk/network/261/files/2012/05/Revised-OCT-12-Data-required-for-research-purposes-bona-fide-research.pdf`.

In 2012, the UK Information Commissioner's Office, whose responsibility it is to oversee the implementation of the Data Protection and Freedom of Information Acts, published 'Anonymisation: Managing Data Protection Risk Code of Practice' (ICO, 2012). The Code sets out recommendations for how organisations can establish whether it is reasonably likely that the publication and sharing of data will result in disclosure of personal data. Consequent to the abortions case, the Code specifically cited this example in recommending that it is good practice to adopt a 'motivated intruder test' as part of a risk assessment. The intruder test would evaluate whether it is possible to identify an individual from the anonymised information, either by itself or in combination with *'other information'*.

While results of intruder testing can be used as strong practical evidence to justify NSI SDC policy safeguards in support of theoretical evidence already gathered, their usefulness is not limited to this. Along the same theme as O'Hara and Whittall (2011), the results can go a long way in refining existing disclosure control measures. They can improve knowledge of the data, helping to inform decisions regarding which parts of the data are most vulnerable and under which circumstances. The case study presented in this paper will illustrate how intruder testing results informed output checking and review of table design before final release of the census data.

2 The UK Census: The Concept of 'Sufficient Uncertainty'

The 2011 UK Census has an ambitious goal to produce local level statistics down to Output Area (OA) levels (small area geography with a minimum of 40 households and 100 individuals) as well as detailed multivariate statistics down to Middle Super Output Area (MSOA) levels; MSOAs comprising an average of approximately 20–25 OAs (a minimum of 2000 households and 5000 individuals). This raises significant disclosure risks since these tables contain many small cell counts with the potential to reveal individual respondent information.

UK SDC policy for Census has been dictated by various pieces of legislation which state the Government's interpretation of disclosure. This includes the Census White Paper, Section 6.7 (ONS, 2008) which states that *"no statistics will be produced that allow the identification of an individual (or information about an individual) with a high degree of confidence."* Section 6.8 of the Census White Paper goes on to suggest that this degree of confidence may be interpreted in terms of the uncertainty *"about the true value of small cells"* and *"as to whether or not the small cell is a true value."* Furthermore, the Statistics and Registration Service Act (UKSA, 2007), Section 39 states that personal information must not be disclosed. For the purposes of subsection (2) information identifies a particular person if, the identity of that person is (a) specified in the information, (b) can be deduced from the information, or (c) can be deduced from the information taken together with any other published information. In addition there is the Code of Practice (UKSA, 2009) accompanied by the National Statistician's Guidance: Confidentiality on Official Statistics, which prescribes that it should not be possible for either identity or attribute disclosure to occur. In particular, for 2011 the Registrars General determined that the focus of effort should be on addressing attribute disclosure and whether anything new can be learned about an individual or a particular group of individuals.

Unlike the 2001 UK Census, this time the UK Registrars General have agreed that small cell counts, i.e., cell counts of one, could be permitted in published outputs as long there is *'sufficient uncertainty'* in those cell counts and in attributed disclosures. While it is not possible to eliminate disclosure risk entirely, the aim of the policy was to create a balance between data risk and data utility. For further discussion of data risk, data utility, identity, and attribute disclosure, the reader is referred to Duncan et al. (2011).

Some of the concepts mentioned—namely *'high degree of confidence'* and *'uncertainty'*—are vague, but they are incorporated into our intruder test. For example, intruders were asked to specify a degree of confidence they had in any disclosure claim. Those claims for which the intruder can provide significant justification, i.e., with high confidence, are also mostly likely to be taken seriously by a court of law as opposed to entirely random claims and will represent the focus of this work. Moreover the interpretation of sufficient uncertainty is unclear but, as a minimum, one would expect it to mean that the majority of claims should prove to be incorrect.

2.1 SDC Policy for 2011 UK Census

To meet the legal requirements, the SDC policy evolved into two parts. One of these is to limit detail in tables at higher level geographies and to control through careful table design at lower levels. The other part of the strategy was to carry out targeted record swapping on the household data before tabulation, and to limit details provided to the public on the exact nature of the methodology (ONS, 2011). Data swapping is a well-established method in the SDC literature (see Dalenius and Reiss, 1982 for the original concept as well as Reiss, 1984 and FCSM, 1994) and has since been adapted into new variations (see Fienberg and McIntyre, 2004 and Carlson and Salabasis, 2012). The targeted swapping methodology used for the 2011 UK Census was also an adaptation that is described in more detail in Shlomo et al. (2010). In summary, households are targeted such that those that represent a greater disclosure risk across high risk/impact variables have a greater chance of being swapped. Households are paired so that they are similar across some of their census attributes but are in different geographical locations. Households are swapped at the geography small enough to make them non-risky: this new variation introduced to increase utility of the data by limiting change at the highest levels of geography (e.g., Local Authority Districts). The proportion of records actually swapped is small, and as the public do not know which records these are, a level of uncertainty can be attached to *every* cell count. Thus a cell count of one, for example, may represent a swapped person. An intruder can never be sure whether any claim of disclosure is genuine. This element of the SDC strategy is the basis for our intruder testing in seeking to measure the real level of uncertainty. The specifics of which cases are qualified as genuine claims of disclosure are detailed in Section 4.1.

When first thinking of an intruder test in this context, there was some difficulty in defining the concept of what disclosure might mean in practice. It can be contrasted with the computer security concept of penetration testing, where the aim is usually to break a protected dataset with zero tolerance of disclosure. However, with record-swapped tabular outputs (and any other SDC method based on creating uncertainty) there are many real disclosures on display, but also some fake disclosures. A fake disclosure might be a cell count of one, for example, that ostensibly reveals information about an individual but actually represents a person with matched characteristics swapped into the area. Justification for ONS SDC policy is centred on the fact that although an intruder may claim disclosure, even with high confidence, he or she does not know for certain whether it is a fake or true disclosure. Thus when thinking about an intruder test, disclosure must be thought of in terms of probabilities. Sufficient uncertainty must be thought of in terms of the overall percentage of correct claims of disclosure rather than a single case of correct disclosure. This percentage will relate to how effective the targeting of high risk records, the nature of intrusion, as well as various other processes introducing uncertainty such as non-response or capture error.

3 Case Study: Intruder Test on the 2011 UK Census

This section describes the structure for the intruder test based on assessing protected census tables as appropriate for publication. Note that microdata samples and other publications from the 2011 UK Census such as origin-destination tables were, at the time of writing, planned for later release so were not considered as part of this test.

3.1 Intrusion Scenarios

As Willenborg (2012) explains well, there is difficulty in studying real intruders and their intrusion methods. If real intruders were to be recruited, it was found that this would have involved substantial cost. Furthermore, there were ethical problems to consider given that, in this case, the census data were pre-release. Willenborg suggests *'friendly intruders'* could be employed, i.e., employees working at the statistical agency acting as much as possible as real intruders might do. These friendly intruders' task would be to see if disclosure is possible, how much effort and knowledge is needed for this, what auxiliary information is available, etc. Various intrusion scenarios would have to be studied as well as various levels of intrusion.

Paass (1988) usefully details various intrusion scenarios and simulates intruder testing using synthetically modelled target persons based on the German microcensus and the income and consumption sample. The scenarios he examines include the public prosecutor who tries to identify a businessman's property data based on additional knowledge of profession, type of income, and benefits, insurance, and business income. Other scenarios include a journalist trying to associate an arbitrary data record to the corresponding citizen, or an industrial enterprise verifying whether some of its employees have an additional occupation.

The approach taken in this case study was to simulate the intrusion scenario of a member of the public with reasonable competence to be able to assess census tables (see Section 3.2) and, similar to that of Simpson (2011), allowing them free reign to see what they could do with the data matching other publicly available information. Our intrusion scenario was also bounded by the UKSA (2007) in considering only publicly available information rather than privately obtained information, e.g., that a business might hold. It is aligned with the spontaneous recognition scenario (Hundepool et al., 2012) where personal knowledge about target units would be reasonable, i.e., age, sex and marital status. On that basis the intruders were asked to think about the following disclosure scenarios:

(i) Can they identify themselves or their household?

(ii) Can they identify someone they know, either individually or within a group, and their characteristics?

(iii) Starting with public information, can they then identify someone, or a group of people, in a table (and learn more than the public information)?

(iv) Starting with the census tables, can they identify a person or a group of people, and link this to some public information?

The National Statistician's Guidance: Confidentiality on Official Statistics (UKSA, 2009) declares that for the 2011 Census it should not be possible for identity or attribute disclosure to occur, with attribute disclosure seen to be the key disclosure risk. The intruders were therefore asked, in each of the above scenarios, to think about what new, additional information they could glean on any census respondent or household. It was stressed, however, that intruders should be free to use their own initiative and that ultimately they should try to disclose information as they wish.

3.2 Recruiting Intruders

The previous examples (O'Hara and Whittall, 2011) and (Simpson, 2011) of intruder testing have both used students in order to *'bring some unorthodox methods into the tests, and to mimic "hackers" in the real world, lacking professionalism and experience in large-scale data handling, but driven by the nature of the problem'* (O'Hara and Whittall, 2011). Although this might be an ideal approach, this was not possible to do in this case. The census data were pre-publication with very strict rules on access. Anyone accessing the data was required to be security cleared to handle material classified as 'Secret'. They were also required to have signed a Census Confidentiality Undertaking which included training in proper use of the data. This meant intruders had to be recruited from within ONS, so they had to be either staff or contractors. This did mean, however, that many of the intruders had good knowledge of census processes and census data possibly leading to better judged claims.

ICO (2012) also provides some guidance on who an intruder might be in their recommended approach to an intruder test:

- Any person who starts without any prior knowledge but who wishes to identify the individual or individuals in the information and who will take all reasonable steps to do so.
- Reasonable competence: investigative techniques may be employed and additional knowledge of the identity of the data subjects, no specialist hacking skills or equipment.
- Consideration should be given to the risks associated with the disclosure of 'ordinary' or 'innocuous' information.
- Consideration should be given to the consequences and impact.

The guidance is open to interpretation so in this case a worst-case scenario was assumed. 'Reasonable competence' was taken to mean intruders who have some methodological skills in reading and interpreting census data, and some I.T. expertise in being able to search the web.

Eighteen intruders were recruited in total. A range of intruders from administrative officers right up to director-level staff were invited to take part; among these was one external contractor. Initially there was some reticence in the intruders' willingness to take part with many citing existing work pressures. This was resolved with greater endorsement from senior management and further explanation of what was involved. Perhaps paradoxically most intruders did not seem bothered about the privacy aspects of the study. In fact, all intruders eventually agreed to take part apart from two who had issues with security clearance. As the exercise took shape, other people were actually volunteering to participate. Intruders reported they found it fascinating to look at data for their own areas potentially containing either their own household or that of friends and neighbours. It was believed that this personal factor provided strong motivation. Only two intruders had any significant experience in disclosure control. However, all volunteers were picked because they could be expected to rigorously test the data, had excellent IT skills, and were adept with external databases. The eighteen intruders were recruited from a range of local areas in England and Wales including urban, suburban, and semi-rural. The results started to become clear at this point and given time/resource availability it was not necessary to recruit any further testers.

3.3 Data and Level of Detail

The census tables used for this test were those that were likely to be the most risky, i.e., those at the lowest levels of geography and also those that were the most detailed. These were provided for each intruder for their chosen area of interest. This provision was measured against the standard of *'disproportionate amount of time, effort and expertise for an intruder to identify a statistical unit'* taken from the ONS Protocol on Data Access and Confidentiality (ONS, 2004).

A complicating factor was that this exercise was carried out before final publication of tables (pending final table design and evidence from this work). Thus a mockup of the final tables to be published was created which had the SDC methodology described in Section 2.1 applied (targeted record swapping and table design): in total 89 carefully selected tables were compiled which were mostly a combination of local characteristics, theme tables, and detailed characteristics as described in the ONS 2011 Census Prospectus (ONS, 2013). Theme tables contain many different variables (often ten or more) at the higher MOSA level on a particular theme such as a certain age group. Local characteristics contain around three or four variables but are at OA level. Detailed characteristics have a detailed categorisation of variables at the MSOA level but only on a small cross-classification of variables. There have been only very minor amendments made to the tables since creation so they should be an accurate representation of the published data. Where there were similar tables, only one or two from the set were created (due to resource issues), so the chance of intrusion via differencing between tables was reduced (though not eliminated), as was also the chance of attribute disclosure via comparison of tables where a person had already been identified in one of the tables.

It is recognised that the incidence of identification of individuals might have been higher if there had been unlimited access to the census tables. However, this was controlled for as much as possible by selecting the most risky tables with the tables provided being either OA- or MSOA-level tables (with only a few univariates at local authority level). Furthermore, to get an accurate reflection of disclosure across the country, a range of areas were included: urban, semi-rural, and suburban (influenced by the initial recruitment of intruders).

3.4 Externally Available Information

The SRSA states that account should be taken of formal public information. For initial guidance on this Elliot et al. (2009) was studied which provides details of various sources. It was found that much information could be found in comparable sources online including name, address, marital status, number of children, and household information on *192people.com*, house prices and ownership on *Rightmove.com*, and occupation on *LinkedIN*, to name a few examples. This was reinforced by the ICO Code of Anonymisation which specifies that alongside the data, *'it is necessary to consider whether other information is available that—in combination with the anonymised data—would result in a disclosure of personal data'* and gives an example of data published on the internet.

The Elliot et al. (2009) report also details barriers to obtaining public information which include cost and level of access which, as in our intruder test, determined that intruders could only do a case-by-case search as opposed to being able to use an entire record file. The final plan was to provide each intruder with access to two separate laptops: one with census data and one with unrestricted internet access (the two were kept separate for security reasons to avoid emailing out of census data). There is the possibility that this may have hindered the intruders' use of the internet for the risk scenario of linking census tables to publicly available datasets containing the population. However, obtaining entire datasets publicly is difficult and believed more likely in the scenarios of Paass (1988) where privately available information is used, such as a company might hold—a business register for example. Intruders were told they would be reimbursed up to the sum of £50 if they felt payment was needed to assist in making a disclosure; for example, to register on paid websites which provide name and address details and basic socio-demographic information on age and sex. This amount seemed appropriate given a review of the websites and what extra information might be obtained.

3.5 Conditions of Reasonable Time and Effort

Each intruder was invited to a session and was encouraged to spend around three and a half hours examining the data; an amount of time considered 'likely' and 'reasonable' in line with National Statistician Guidance. The intruders worked in a meeting room within a secure census area which required pass access. Each intruder was given a 20-minute briefing before starting their session. They were given a summary of the SDC Census 2011 policy (including a written Q&A), talked through possible theoretical examples of disclosure, supplied with a guide on 'how to find people online,' a list of the selected census tables, and their title/definition, and maps of the OA and MSOA to which the tables related (see example, relating to the ONS office postcode, in the Appendix).

Intruders were directed to write down names and addresses and the reference of the cell(s) and census table(s) that indicated any disclosure. Intruders were encouraged to think about the exercise beforehand and to bring in any extra material to help them make any disclosures. Although a few did come up with a strategy beforehand, no-one brought in any extra material.

3.6 Parameter Settings

Ideally this exercise would have examined a variety of factors. It was planned to analyse tables protected by swapping with varying parameters, i.e., different swap rates and risk variables. A further thought was to allocate intruders to more than one area for control purposes but perhaps on the other hand to allow for consideration of different layers of knowledge, both public and private. In reality, this exercise was a considerable undertaking. To reflect Willenborg (2012), *"although I think experimental work in SDL is essential, I am ready to admit that is impossible to apply on a routine basis, as it too involved, time consuming and expensive. So it can only be done in a limited number of experiments."*

3.7 Security and Ethical Considerations

Security of the census data in this exercise was taken very seriously. Intruders were asked to sign a confidentiality declaration which, although it had no legal basis, reinforced the message not to misuse the data. Intruders wrote all notes in notebooks which were taken and locked away on completion of the sessions. Laptops were chained to the desk during the exercises and all data were securely wiped after each session. In addition, intruders were made aware that any searches they made could be stored by website providers, particularly those that required registration or payment.

This intruder exercise, by definition, involved searching for personal information about people on the web. This raised ethical questions since on some websites (e.g., *192people.com*) it is possible to pay to see who has ever web-searched you. Intruders were made aware of this at the start so they could avoid these websites if they felt it was necessary. Intruders were also asked to supply names and addresses along with each claim of disclosure necessary for validation purposes. Thus the exercise required a certain level of trust on the intruders' part for the information they provided to be dealt with confidentially. Assurances were provided to intruders in the confidentiality declaration and in briefing and speaking with them. Personal details were stored on secure census computers and very limited summary details provided outside of the team directly involved.

4 Analysis of Intruder Reports

This section discusses validation and subsequent results from the intruder test, and summarises opinions of the intruders. Exact figures on number of claims crossed with the various breakdowns have been omitted to avoid compromising confidential information.

As discussed in Section 2, the concept of disclosure risk is more challenging with tables protected by record swapping since many disclosure claims will in fact be correct. This is because record swapping does not eliminate disclosure but aims to provide uncertainty around which cases (usually small cell counts) are genuine. In keeping with empirical evidence already gathered on proportions of ones and attribute disclosures that are real, the intruder testing validation (described in Section 4.1) operated on the basis of assessing the proportions of claims of disclosure that were true. Note that disclosures introduced by processes of imputation would be considered *not real* as well as those introduced by swapping.

Intruders were asked to note down a level of confidence, preferably numerical, they had that any claim was correct. Claims with high levels of confidence are of particular interest since these are the ones that the public are most likely to use in challenging SDC policy. The high confidence claims are also those that the courts are most likely to take seriously given that it would be 'reasonable' to expect the intruder to provide significant justification rather than making entirely random claims. Some confidence levels were based on what the intruder had written down in words or said; for example, 'fairly sure' was taken to imply a confidence interval of 60–79%. Table 1 defines the intruder confidence levels: this aid evolved from the words used by the intruders and their numerical association of confidence.

Overall there were more than 50 claims from the 18 intruders with between 0 and 8 claims per intruder during their 3.5 hour session. There were two types of claim—a claim of identity disclosure and a claim of attribute disclosure. This generally translated to "I can recognise somebody I already know well, such as a close friend or family member, in these data" in the case of identity disclosure and "I can find out something I didn't

Table 1: Interpreting intruder confidence levels (based on intruders' words and numbers).

Confidence Level	**Meaning in words**
0-19%	Not at all confident, complete guess
20-39%	Not very confident, bit of a guess
40-59%	Not quite sure, uncertain
60-79%	Fairly sure, reasonably confident
80-100%	Very confident, absolutely sure

already know about someone I know well, such as a close friend or family member" in the case of attribute disclosure.

There were several known reasons for certain claims being wrong such as non-response, imputation, capture processing (especially coding from either free text or multiple ticks), respondent error, as well as record swapping. Interestingly though, several claims were incorrect due to intruder error. There were a variety of reasons for this, for example, intruders typically forgot what they or their household wrote on their census form and thought themselves to be in a different category to what was originally recorded. On other occasions, intruders were incorrect when trying to work out what a family member would have put when answering (e.g., on questions about occupation or health), sometimes making incorrect assumptions. Also there were instances where the table was misread, or the variables in the table were not understood fully. The test also illustrates the difficulties of matching against public data, issues with timeliness, collection modes, and variable categories all being different.

4.1 Validation

The majority of intruders made some claims of disclosure, although two did not make any. The validation process involved five steps:

(i) The cell reference and the table relating to the disclosure claim were noted.

(ii) Using that information, the record(s) in that cell was identified in the census unit record database (post-swapping, post-imputation) in the 'ad-hoc' system.

(iii) Each record in the database contains a form identifier which was noted.

(iv) The form identifier(s) was used to select the relevant image(s) of census form(s) in the census database—(obviously pre swapping and pre imputation now, as looking at the original scanned in images).

(v) The image form was checked to see if it matched the name and address supplied by the intruder.

Note that this was a relatively rudimentary approach to validation, working on the principle that if the person/household referenced in the cell correctly related to the original census form then that counted as a correct disclosure. However, validation of these claims had the potential to be more complex if one were to give further expert thought on how to go about it. For example if an intruder made a claim about a cell count of one stating that it was their neighbour—say, person A; it could be the case that although the data correctly recorded them in the original pre-swapped, pre-imputed database, they had been swapped with person B. Therefore in the post-swapping post-imputation cell it shows person B. This claim would be incorrect under our principle when one could argue it should be a correct claim because the particular characteristics in this table are the same. This approach was considered to be the best and most straightforward because any further inferences on the person (in this case a neighbour) *are likely* to be incorrect since the person has been swapped and will not necessarily match on any other characteristics in any other table. This fits with the policy focus on attribute disclosure where new information learned is of main concern rather than merely recognising a person (see Section 2). With more thought and time the results could have been broken down to highlight the nuances of the impact of the swapping and imputation methodology (not discussed either are cell counts which do not equate to the same value in the pre-swapping pre-imputation database). It was considered that a court of law is probably more likely to work on the straightforward principle stated at the beginning of this paragraph.

It was sometimes difficult to discern whether intruders were claiming identification or attribute disclosure. An attribute disclosure occurs when a table enables an intruder to infer further, previously unknown, knowledge about a person or household. However, many intruders may not have realised that further information could be inferred, or did not actively look for this type of disclosure once they had identified someone (although this was encouraged in the briefing beforehand). There were also several claims of group identification disclosure involving identification within small cells, for example claiming a relative was represented in a cell which had a count of two or three.

4.2 Summary of Results

Table 2 shows an overall pattern where claims made with greater confidence were more likely to be correct, although unexpectedly there is a slight dip towards the claims made with certainty. Note the dip may be due to small samples in each confidence level rather than anything statistically significant. The results were then filtered based on who the intruder had identified as in Table 3. ‘Neighbour’ refers to any neighbour in the immediate area. Importantly, aside from one single case, all claims made involved persons for whom the intruder claimed to have personal knowledge. Moreover the results indicated a far greater chance of disclosing correct information about their own household or a family member emphasising the difficulty in identifying someone you know less about, which is usually the aim of an intruder. Another interesting result was that intruders who claimed to have found themselves weren’t always right.

When the claims were filtered by geography (Table 4), it was clear that being able

Table 2: Percentage of correct claims by confidence level.

Confidence Level	**Percentage of claims correct**
0-19%	0%
20-39%	11%
40-59%	38%
60-79%	67%
80-100%	47%

Table 3: Subject of the disclosure claim.

Who is being claimed	**Percentage of claims correct**
Neighbour	36%
Self/Family	61%

Table 4: Level of geography at which a claim was made.

Geography Level	**Percentage of claims made**	**Percentage of claims correct**
OA	94%	45%
MSOA	6%	50%

to identify someone at MSOA level proved very difficult with only a very small number of claims made in comparison with at OA level. Intruders often commented on how the local characteristics tables, which are all at OA level, were far more useful than any other tables despite having less detail in the constituent variables. Intruders who thought they might have identified someone in the table in an MSOA would often then look for greater confirmation in the OA tables. The results were also broken down by the type of claim made (table 5). The types of claim were Person or Household, in which a person or household has been identified, or Swapping which were fascinating cases referring to where an intruder believes somebody has been swapped in or out of the data. Again the majority of claims were incorrect within each category of type of claim. Whilst claims of disclosure were made at the household level, in the main they were concerned with a particular person.

Finally, which census tables were used to make the most claims was analysed and the associated success rates of the claims—the former shown in Table 6. A select number of tables were much more commonly used than others (giving rise to five claims or more) making up over half of all claims. These popular tables all included sex and age

Table 5: Correctness of claim by type (household/person/swapping).

Claim Type	**Percentage of claims made**	**Percentage of claims correct**
Household	16%	50%
Person	74%	46%
Swapping	10%	33%

Table 6: Census tables used to make the claims.

Table Used	**Summarised Topic**	**Claims of Disclosure**
CAS036	sex/industry/age	10
CAS002	age/sex/marital status	7
CAS028	sex/age/economic activity	6
CAS038	sex/industry/employment	5
CAS016	sex/age/health/disability	5
CAS119	sex/age/travel to work	4
CAS056	tenure/central heating	4

bar one which only included sex. It is worth pointing out that whilst most intruders made a claim in just one table, some were able to identify a person/household across several tables. Interestingly, the three most used tables for making claims show an above average percentage of claims correct in comparison with the other tables (figures not disclosed).

As specified in Section 2, the Registrars General had agreed that attribute disclosure, not identity disclosure, was to be the primary concern for the 2011 UK Census. Some attribute disclosure claims were made, but whether the figures for these are truly representative of how many could or should have been made is unclear. Of the instances where an intruder said they could identify someone in the tables, only in a tiny number of these instances were the intruders able to infer further new knowledge with not all being correct. However, this is not totally reliable, since some may not have realised further information could be inferred, or did not actively look for this type of disclosure, as mentioned earlier. The number of cases is very small and not really adequate to make sweeping judgements.

Referring back to the scenarios which the intruders were asked to consider in Section 3.1, it was very difficult for intruders to obtain data from the internet and then link it to anyone in the census tables, and only one claim of this type, which proved to be incorrect, was made. Of the others who tried to do this, they said they were unable to match up the information they had found to a low count in the tables. Some intruders did use the internet to help with identifying people, in particular to verify basic information such as names, addresses. and ages. On one occasion, the use of internet information to support a claim led to incorrect identification of a person, whereby an age obtained from an internet site proved incorrect. A possible alternative explanation is that the intruder did not take into account the difference in time between the census and the date of the test.

4.3 Intruder Comments

All intruders were asked to write any comments they had at the end of the exercise. There were many comments stating how difficult trying to identify someone was, with these intruders generally saying how their area had little variation demographically, making it difficult for people they knew or had found to stand out in the tables.

There were concerns expressed with small counts of certain ethnic groups and religions within the OA tables. They suggested that these low counts, although not identifiable to themselves, may well be to others and they perceived this to be unsafe. Coincidentally in all of these cases the counts were found to be incorrect due to either misreading the categorisation in the table or due to swapping. A final point is that intruders were asked whether they felt the data generally reflected the characteristics of their area and were prompted to ascertain the swap rate. The general consensus was that the tables looked correct for their area and that the swap rate looked 'low.'

5 Conclusions

Intruder testing was shown to be very useful. It provided empirical evidence, within the legislative framework for the 2011 UK Census, of whether the SDC policy was satisfactory. It backed up theoretical evidence already gathered and indicated whether specific strands to the methodology were working, i.e., the targeting of risk variables for swapping. It was helpful to demonstrate where further work needed to be done, highlighting vulnerabilities in the data. For example variables such as age and sex were used as common identifiers. Part of the ONS SDC strategy has required refining the format of the tables prior to final publication; results of the intruder testing informed which tables might need special attention and these were prioritised for review. Intruder testing gave indications on how easy it was to match data with public sources and (mostly) unguided by pre-defined scenarios, the creative ways intruders might try to identify individuals or groups of individuals. Intruder testing is finally very helpful to get a handle on perception of disclosure. It gave a unique insight into how the public might perceive disclosure control, if any has been applied and to what extent.

To be specific, this novel example of intruder testing on disclosure-controlled (anonymised) census data has shown that it is very difficult to re-identify respondents correctly in the 2011 UK Census and moreover, it is virtually impossible in this case to identify anyone correctly without any personal knowledge about them. In terms of the ICO motivated intruder test (as described in Section 3.2), this was a positive result since the guidance recommends to assess disclosure based on a person without any prior knowledge. Even given that the claims that were made about people the intruder would expect to know reasonably well, i.e., family, themselves, or a neighbour, the percentage of correct claims was still surprisingly low. Moreover, the majority of all claims made with 80–100% confidence were still incorrect; these were cases where the intruder was absolutely certain that they knew the identity of the person in the table cell. This work has also demonstrated that a great deal of external information about a person would be needed to be able to match them to a table cell. There was a very low instance of claims of attribute disclosure, where an intruder found out 'new' information about a person and again not all of these were correct. These attribute disclosures were all conditional on a prior identification disclosure—of which it was learned the majority of the latter were not always correct. This was deemed to be the primary disclosure risk for 2011 UK Census so in that context the result was very pleasing.

In reference to ICO (2012) on a motivated intruder test and the impact of the claims, identification of people in tables with sensitive variables such as health or disability was shown to have an even lower chance of being correct (details withheld for reasons of confidentiality of the intruders)—likely a direct result of the ONS' SDC policy of targeting high risk/impact variables (see Section 2.1).

Leaving aside the nature of the disclosure claims, it is important to remember that some claims made were correct. It is generally impossible for NSIs to publish data that retains some usefulness with zero disclosure risk, so it has always been about getting the right balance. The precise nature of the claims as discussed indicate that the level of disclosure control applied provides sufficient uncertainty where needed in the tabular outputs. This work has since been endorsed by the UK ICO *'proving re-identification risk (as we call it) is manageable.'*

The authors note there were limitations to this exercise which should be considered when making any generalised conclusions. To be specific, we refer to the socio-demographic characteristics of the intruders and their motivation given that they worked for ONS, the small sample of postcodes assessed, the fact that only a representation of the full set of output tables to be published was examined, the time spent by intruders at each session, and the availability of external information. Further work of this kind would be helpful in working towards a standardised approach for intruder testing in terms of methodology, time, likelihood, and use of external information. For example, it may be argued that such an experiment could be prolonged over a longer period for more reliable results.

To conclude, intruder testing can have all sorts of applications. It can be applied to survey as well as census data to assess levels of disclosure risk pre- or post-SDC. It can also be applied to administrative data (registration based data from government and government-funded organisations). There is a continuing drive to examine mechanisms for combining administrative data and making it available for research safely. Intruder testing can be very useful to pinpoint parts of the data or variables which are of particular concern for which greater protection is required. At ONS, collaborative work of this nature is already underway with Manchester University to design innovative intruder tests for social survey microdata—data that are currently only available under licence—and being considered for more open access. More detailed disclosure scenarios will be examined and both public and private external databases will be used. Given the push to get more data out into the public domain, NSIs need to adjust their outlook in thinking about justification of SDC policy and whether there is sufficient practical evidence not to release data rather than automatically assuming the contrary.

Acknowledgments

The authors would like to credit Dr. Frank Nolan—former Director of Census Methodology and Infrastructure who sadly passed away before the completion of this work. Frank instigated this work and was a strong proponent of intruder testing. We would also like to acknowledge the team who helped carry out this exercise: Helen Ross, David Greenway, Lucy Graham, Joe Frend, Philip Lowthian, Peter Youens, and finally Paul Jackson for his invaluable legal advice. The authors are extremely grateful for the reviewers' direction in improving this paper.

References

Carlson, M. and Salabasis, M. (2012). A data-swapping technique for generating synthetic samples; A method for disclosure control. *Res. Official Statist.*, 5:35–64.

Dalenius, T. and Reiss, S. P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference,* 6:73–85.

Duncan, G., Elliot, M., and Salazar-Gonzalez, J. (2011). *Statistical Confidentiality Principles and Practice*. Springer.

Elliot, M., Mackey, E., and Purdam, K. (2009). Data Environment Analysis Service (DEAS). ONS Internal Report.

Federal Committee on Statistical Methodology (FCSM) (1994). Subcommittee on Disclosure-Avoidance Techniques, Statistical Policy Working Paper No. 22: Report on Statistical Disclosure Limitation Methodology, Washington, D.C.

Fienberg, S. and McIntyre, J. (2004). Data swapping: Variations on a theme by Dalenius and Reiss. In Domingo-Ferrer, J. and Torra, V. (eds), *Privacy in Statistical Databases*, vol. 3050 of *LNCS*. Berlin/Heidelberg: Springer. 519.

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

Information Commissioner's Office (ICO) (2012). Anonymisation: Managing data protection risk code of practice. `http://ico.org.uk/for_organisations/data_protection/topic_guides/anonymisation`

Jackson, P. (2012). Microdata exchange and the challenges of open data and transparency. In *Expert Group for International Collaboration on Microdata Access*. Paris, France: OECD Conference Centre.

Narayanan, A. and Shmatikov, V. (2006). How to break anonymity of the Netflix prize dataset. `http://arxiv.org/abs/cs/0610105`.

O'Hara, E., K. Whitley and Whittall, P. (2011). Avoiding the Jigsaw Effect: Experiences with Ministry of Justice Reoffending Data. Tech. report, Southampton University.

Office for National Statistics (ONS) (2004). ONS protocol on Data Access and Confidentiality. Tech. report, available online at: `www.ons.gov.uk/ons/guide-method/the-national-statistics-standard/code-of-practice/protocols/index.html`. Accessed June 2013.

— (2008). 2011 Census White Paper. Tech. report, available online at: `www.ons.gov.uk/ons/guide-method/census/2011/how-our-census-works/how-we-took-the-2011-census/how-we-collected-the-information/questionnaires--delivery--completion-and-return/2011-census-questions/index.html`. Accessed June 2013.

— (2011). Evaluating a Statistical Disclosure Control (SDC) Strategy for 2011 Census Outputs. Tech. report, available online at: `http://www.ons.gov.uk/ons/guide-method/census/2011/how-our-census-works/how-we-took-the-2011-census/how-we-planned-for-data-delivery/protecting-cofidentiality-with-statistical-disclosure-conctrol/index.html`. Accessed June 2013.

— (2013). 2011 Census Prospectus. Tech. report, available online at `http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-prospectus/2011-census-prospectus.pdf`. Accessed June 2013.

Paass, G. (1988). Disclosure risk and disclosure avoidance. *Journal of Business Economics and Statistics*, 6:487–500.

Reiss, S. (1984). Practical data-swapping: The first steps. *CM Transactions on Database Systems*, 9:20–37.

Shlomo, N., Tudor, C., and Groom, P. (2010). Data swapping for protecting census tables. In *Proceedings of Privacy in Statistical Databases (PSD'2010)*, vol. 6344 of *LNCS*. Springer. 41–51.

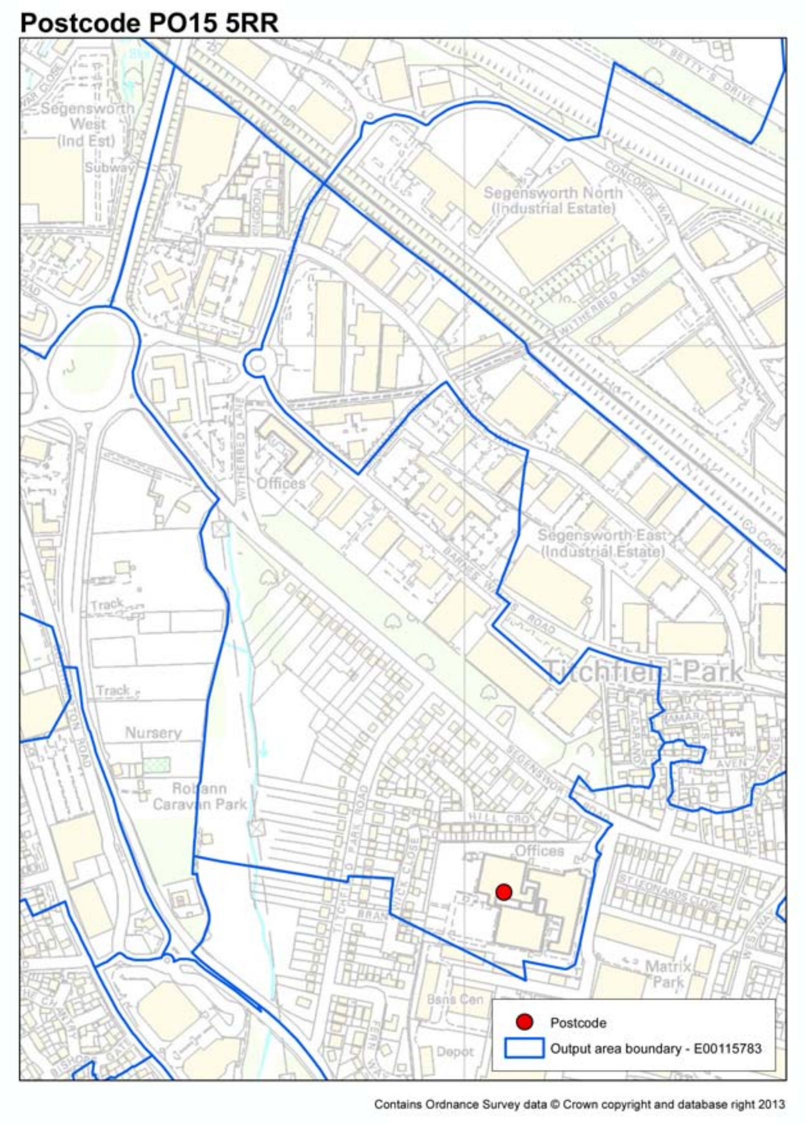
Simpson, A. (2011). On privacy and public data: A study of data.gov.uk. *Journal of Privacy and Confidentiality*, 3(1):51–65.

UK Statistics Authority (UKSA) (2007). Statistics and Registration Service Act, available online at: `http://www.legislation.gov.uk/ukpga/2007/18/contents`.

— (2009). Code of Practice for Official Statistics, available online at: `www.statisticsauthority.gov.uk/assessment/code-of-practice/index.html`.

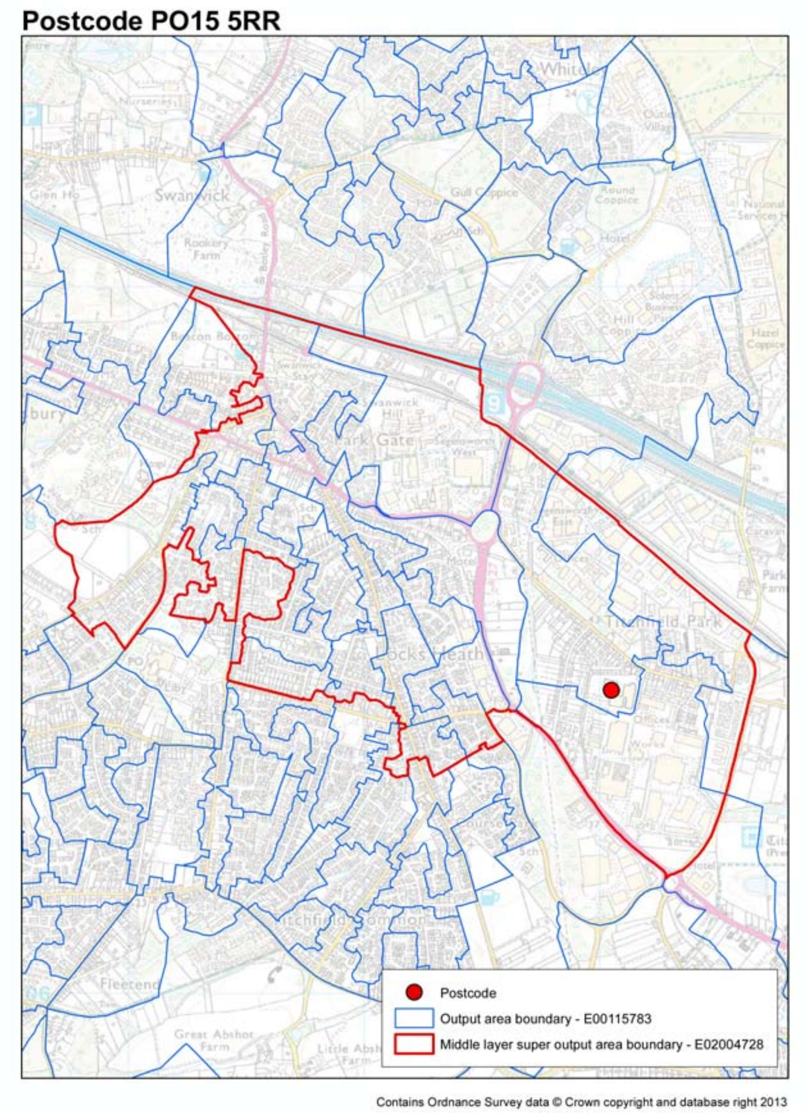
Willenborg, L. (2012). Discussion [of Skinner, C.]: Statistical disclosure risk: Separating potential and harm. *International Statistical Review*, 80(3):375–378.

Appendix



Contains Ordnance Survey data © Crown copyright and database right 2013

Figure 1: Example Output Area (OA) - (based on the Office postcode) indicating the size of this geography.


Postcode PO15 5RR
Swanwick
Park Gate
Locks Heath
Titchfield Common
Fleetend
● Postcode
Output area boundary - E00115783
Middle layer super output area boundary - E02004728

Contains Ordnance Survey data © Crown copyright and database right 2013

Figure 2: Example Middle Super Output Area (MSOA) - (based on the Office postcode) indicating the size of this geography.