Journal of Official Statistics, Vol. 32, No. 2, 2016, pp. 329–348, http://dx.doi.org/10.1515/JOS-2016-0019

End User Licence to Open Government Data? A Simulated Penetration Attack on Two Social Survey Datasets

Mark Elliot[1], *Elaine Mackey*[2], *Susan O'Shea*[3], *Caroline Tudor*[4], *and Keith Spicer*[5]

In the UK, the transparency agenda is forcing data stewardship organisations to review their dissemination policies and to consider whether to release data that is currently only available to a restricted community of researchers under licence as open data. Here we describe the results of a study providing evidence about the risks of such an approach via a simulated attack on two social survey datasets. This is also the first systematic attempt to simulate a jigsaw identification attack (one using a mashup of multiple data sources) on an anonymised dataset. The information that we draw on is collected from multiple online data sources and purchasable commercial data. The results indicate that such an attack against anonymised end user licence (EUL) datasets, if converted into open datasets, is possible and therefore we would recommend that penetration tests should be factored into any decision to make datasets (that are about people) open.

1. Introduction

The UK's Office for National Statistics (ONS) currently disseminates large numbers of datasets under end user licence (EUL). This is a restricted dissemination of the data to researchers who agree to a set of sixteen licence conditions and specifically agree not to attempt to reidentify individuals. Under the transparency agenda, ONS has considered whether some of these could be released under an Open Government Data licence. This is effectively unrestricted publication on the Internet. This is clearly a very different level of dissemination and therefore careful conceptual and disclosure risk analyses was necessary in order to understand the marginal increase in disclosure risk (if any) associated with this change in dissemination policy.

The work took place in two phases. During Phase 1 we considered the interplay of legal and statistical definitions of confidentiality, developing a detailed understanding of the differences in the licences and associated documents. This was essentially a socio-legal piece of work, which in turn allowed us to generate a set of feasible scenarios that extended beyond the orthodox intruder scenarios. Orthodox scenarios come in two basic forms: (i) *database cross match*, where an intruder attempts to link records in an *identification file* and a de-identified *target file* but does not know for certain for any given record in the identification file whether there is any corresponding record in the

[1,2,3] School of Social Sciences, University of Manchester, Manchester M13 9PL, UK. Emails: mark.elliot@manchester.ac.uk, laine.mackey@manchester.ac.uk, and susan.o'shea@manchester.ac.uk.
[4,5] Office for National Statistics, Segensworth Road, Titchfield, Fareham, Hampshire, PO15 5RR, UK. Emails: caroline.tudor@ons.gsi.gov.uk and keith.spicer@ons.gsi.gov.uk.

target database, and (ii) *fishing*, where the intruder selects records from the target database and attempts to find the corresponding person in the population. There are other variants – see Elliot and Dale (1999) for a discussion – however, the critical point is that response knowledge is not assumed. In general, it is held that, with EUL licenses, the potential costs to a researcher of attempting reidentification (e.g., career damage) outweigh the benefits of doing so. Therefore, even though it is possible that a researcher might know that a third party is in the data, such intrusions are unlikely. Under the Open Government Data (OGD) licence, the mere act of identification would not break any rules and therefore the costs of such reidentification are simply the effort required to carry it out. On top of this, the fact that OGD means effectively universal access implies the very strong possibility that somebody exists who would know that some other person was in the data (and probably for many respondents there would be somebody with such knowledge). It is generally accepted that with response knowledge, reidentification is considerably easier (i.e., the effort required is considerably less). The combination of these factors makes response knowledge scenarios far more plausible with OGD.

After a review of the report on Phase 1, ONS commissioned a Phase 2 study: a simulated attack based on an intruder who had response knowledge that an individual was in a dataset and then used publically available information (either openly available or available for a fee) in order to identify that individual in the dataset. For this stage, which we report upon here, two UK datasets were focused on: the Labour Force Survey (LFS) and the Living Costs and Food Survey (LCF). These are both microdata samples that are smaller than one percent of the UK population and contain information on individuals and their households. Some disclosure control has been applied, such as banding age and ethnic group. This study builds on previous attack simulations (e.g., Müller et al. 1995; Elliot 2009) but adds an additional step of trawling for and combining available public information (rather than simply matching two fixed datasets).

This article consists of the following sections. Section 2 reviews the existing literature on reidentification. Section 3 summarises the Phase 1 study, which describes the motivation for the attack scenario. Section 4 describes the methodological approach to the penetration test. Sections 5 and 6 describe the matching process and the results of the consequential reidentification attempts. Section 7 is the general discussion.

2. Review of Reidentification Studies

Reidentification studies come in three different forms: (i) defensive studies carried out by or on behalf of data custodians, often called penetration tests, where the goal is to assess disclosure risk associated with a proposed data release; (ii) academic studies exploring new attack forms or new potential anonymisation techniques; and (iii) demonstrative studies usually carried out by data journalists and/or academics, where the point of the study is to demonstrate that a given release is unsafe.

The earlier studies were largely of the second type. Müller et al. (1995) tested whether it was possible to link records in the 1987 German microcensus file to an administrative register. Results varied depending on the scenario assumed, but in general they concluded that "although identification is not impossible, only under special circumstances are the

chances of a successful identification larger than virtually zero"; p.149. Similarly, Elliot and Dale (1998) showed through a study linking UK census microdata to a sample survey that it was possible to reidentify some people by cross matching databases but that the correct matches were effectively hidden amongst many false positive matches. Later Elliot (2009) demonstrated that by focusing on unusual records (using the so-called fishing attack) it was possible to achieve a higher hit rate, but he also found that the anonymisation methods that ONS had employed in the test file (2001 census microdata) did effectively stymie that attack.

El Emam et al. (2011) carried out a systematic review of reidentification attacks on health data to (i) compute the overall proportion of correctly identified records, and (ii) assess whether it indicated weakness in current anonymisation methods as used with health data. On average, approximately a quarter of the records were reidentified across all studies. They concluded the evidence showed a high reidentification rate, but that this was mostly based on small-scale studies on data that were not anonymised according to existing standards. This evidence is insufficient to draw general conclusions about the efficacy of anonymisation methods.

Recent academic work has focused on new forms of data. For example, genomics data (Malin and Sweeney 2004; Gymrek et al. 2013) and social network data (Backstrom et al. 2007; Narayanan and Shmatikov 2009) have both come under the spotlight; the general conclusion drawn is that the more complex the form of data, the more vulnerable those data are to reidentification attacks.

The practical importance of these studies has been to show that care is required before data is released, particularly if it is to be released as open data. Examples where such care has not been exercised have led to the third (demonstrative) type of reidentification study.

A particularly notorious example of a demonstrative study arose from the release of a database of supposedly anonymised movie ratings by Netflix. The data were released in an attempt to improve its movie recommendations algorithm through crowdsourcing the problem, offering a $1 million prize for the best solution. For each case, a unique subscriber ID, the movie title, year of release and the date in which the subscriber rated the movie were given.

Narayanan and Shmatikov (2008) showed in their study how Netflix users could be reidentified. They were able to identify (some) users by matching their Netflix reviews with data from other sites like IMDb (http://www.imdb.com accessed 14/7/15). Furthermore, they found that if you knew a few movies a Netflix subscriber had rented in a given time period, you could reverse engineer the data and find out the rest of their viewing history. They concluded that very little auxiliary information is needed to de-anonymise an average subscriber record from the Netflix Prize dataset. With eight movie ratings (of which two may be completely wrong) and dates that may have a 14-day error, 99% of records can be uniquely identified in the dataset. For a 68% hit rate, two ratings and dates (with a three-day error) are sufficient.

The Netflix example and similar demonstrative studies involving AOL search data (in 2006), the New York taxi cab dataset (in 2014) and Transport for London bike journey data (in 2014) demonstrate the difficulties of releasing datasets that have not been thoroughly tested for reidentification risk as open data, and in particular they demonstrate the value of defensive reidentification tests.

3. Motivation for the Response Knowledge Scenario

The initial focus of the Phase 1 work was to consider a range of different materials including:

1) The Data Protection Act (1998).
2) The OGD licence.
3) The EUL licence.
4) A document provided by ONS detailing their view of the differences between the OGD and EUL licences.
5) The Anonymisation Code of Practice produced by the UK Information Commissioner's Office (ICO).
6) The standard confidentiality pledge provided by ONS to respondents.
7) The UK Government Statistica Service (GSS) Disclosure Control policy for Microdata Produced from Social Surveys.

Analysis focused on differences in the data environment in which data would exist under the two different licence forms. This has been embedded in developments of our thinking about both the relationship between statistical and legal confidentiality and the conceptualisation of the data environment (see Mackey and Elliot 2013; Elliot and Mackey 2014).

The analysis presented here is somewhat different from an orthodox disclosure risk analysis. During this phase we were trying to build a well-grounded description of the problem, its attributes and the likely and plausible consequences of a decision to change the licensing for the current EUL datasets.

3.1. Understanding the Differences Between the Licences

There are considerable differences between the EUL and OGD licences. An analysis of the licences led us to the conclusion that the following differences impact on the disclosure risk either directly or indirectly.

Restrictions on use: Clause 2 of the standard EUL provides a fairly tight definition of how the data may be processed. The OGD licence provides no restriction. As we shall see later, this is a critical factor in creating new disclosure scenarios.

Restrictions on sharing: EUL Clause 6 restricts sharing to other EUL holders (which in effect means that data can only be shared with those who already have access to it). The OGD licence (of course) places no restrictions on sharing.

Preservation of confidentiality: EUL Clause 8 imposes a specific responsibility on users to preserve respondent confidentiality. It also specifically prohibits deliberate statistical disclosure. The OGD licence does not provide any such responsibility, relying only on the Data Protection Act. The OGD license does refer to compliance with the European Directive 2002/58 on Privacy and Electronic Communications. However, this Directive does not concern individual data of the type that is in question here and therefore it is ignored henceforth.

It should be noted that the EUL does not explicitly prohibit identification (of oneself or others). However, it is hard to construct a reasonable use case where identification (of oneself or others) does not breach Clause 2 and it is hard to construct a plausible

scenario of identification of others which does not breach Clause 8. We consider therefore that identification is, for practical purposes, implicitly prohibited by the EUL.

3.2. Key Points of Interpretation of the Data Protection Act

The Data Protection Act (DPA) concerns personal data. In general, the processes of anonymisation and statistical disclosure control are designed to render the data non-personal. Personal data are defined in the DPA as data which relate to a living individual who can be identified either:

a. from that data, or
b. from that data and other information which is in the possession of, or is likely to come into the possession of, the data controller.

De-identified data – where the formal identifiers have been removed or masked – is no longer *identified* but may still be *identifiable*. The first clause clearly does not apply to de-identified data and therefore Clause b is our concern. *The key point about Clause b is that it is contextual. One cannot make judgements about whether data is personal or not simply by considering the data itself.* Whether the data is personal or not will depend upon the environment in which it resides. Each data environment has attributes which affect the personal/non-personal nature of the individual-level data contained within it. These include:

- Other data. It is explicit in the definition of "personal data" that other data is relevant.
- Data users. Data users have identifying knowledge of other individuals. They also move between data environments and carry information with them as they do so. Users have varying levels of expertise that make them more or less able to carry out the necessary data processes to enact identification.
- Data security. The better the security with which data is kept, the stronger the partition between the data environment and other data environments.
- Governance structures and processes (including licences). Governance processes create expectations about what may be reasonably done with the data. Some processing of data will make it more likely that the data will become personal.
- The intra-environment ethos. The prevailing ethos within a data environment will affect the practice of interacting with data. Behaviour and attitudes which are not necessarily precisely specified in licences come into play here. Does the prevailing ethos specify that one should *look after* data?

It is fairly clear that for all of the above attributes, an OGD licence increases the likelihood that a given dataset which relates to living individuals will be regarded as personal data because the probability of identification will be higher:

- Other data. The OG data effectively exist in the global data environment and therefore in principle could be linked to any other data. Under EUL the data could only possibly be linked to data that the researchers who are using it have access to.
- Data users: The user base will increase massively under an OGD licence compared to the EUL licence (this is the point of OGD).

- Data security. By definition there is no security with OG data – the data is open. EUL data exist in relatively controlled research environments (although these are not high security).
- Governance structures and processes (including licences). With the OGD licence, the data is unregulated as the licence is deliberately permissive; there are very few restrictions on data processes. The EUL places restrictions on what a researcher may or may not do.
- The intra-environment ethos. In the global data environment, it is probably fair to say that there is no one coherent ethos and the full range of behaviours in respect of data can be expected. In the EUL environment there is a prevailing expectation that researchers will look after data.

In essence, it is thus clear from this surface analysis that for survey microdata the OGD licence can only increase the risk that data is personal, which leaves us with the question of: *by how much*?

3.3. The OGD User Base

The overarching principle of OG data is to increase the accessibility of government data. Increased accessibility means a larger user base; larger in number with a greater diversity in types of users. Ignoring any other impacts of the OGD licence, this must increase the risk of a disclosure event. Assuming the risk is non-zero with the population of EUL users and assuming that all users are equal in terms of their risk impact, then the probability of an attempt will increase in proportion to the increase in the size of the user base. It is certainly open to discussion whether the OGD user groups are more or less risky than the EUL ones but, given that the latter is essentially a subset of the former, it is difficult to argue against the proposition that the risk would increase with the increase in size of the user base.

If this were the only problem then the increase in risk attributable to users could be managed through the orthodox trade-off mechanism – by applying more stringent disclosure control to the data. However, the problem is far more difficult than this. The OGD licence changes the nature of the disclosure risk problem, creating a whole new type of disclosure scenario.

3.4. Open Government Data Disclosure Scenarios

Work on attack scenarios for survey data includes Paass (1988) and Elliot and Dale (1999). Orthodox scenarios include *database cross match* (where two databases are linked) and *fishing* (where the intruder identifiers outliers in the data and then attempts to find those individuals in the population). When considering what additional disclosure scenarios are in scope under OGD but are not in scope under EUL, we are not concerned with technical possibility. Technically, there is nothing that a user could not do under EUL that one could do under the OGD licence. However, the licences do create formal (quasi-legal) restrictions on activity and this restriction significantly affects the shape of the data environment.

We considered six different scenarios in our analysis at this stage:

1. Self-identification
2. Spontaneous recognition
3. Spontaneous recognition augmented by subsequent response knowledge
4. Commercial data augmentation
5. Response knowledge with collusion
6. Response knowledge without collusion

There is not sufficient space to consider all of these here; suffice it to say that we considered the sixth to be the most problematic.

3.4.1. Response Knowledge of Others

Response knowledge attacks are where users who know that a particular respondent is in the OGD dataset and also have other knowledge about them use the combination of that knowledge to identify the respondent. Here we are considering a data environment that is effectively just the user themselves. In this environment and scenario, the respondent's data is personal for the user, because they have data that allow them to identify the respondent.

Given this, there is a question of whether the identification process itself constitutes a breach of the DPA. Specifically, can a user be deemed to be unfairly processing data by identifying a record in a dataset? Such an identification process does not change the status of the record concerned – it was personal and it still is. The only difference is that the user now knows precisely which record is the data for the respondent. They may not have even learnt anything new about the respondent. The scope of what constitutes "processing" within the DPA is broad and includes "alignment" and "combination", which identification processes could be said to constitute, but it is unclear about whether identification in this case would be deemed unfair or not.

The wording of the OGD licence confuses things further. It states that "you are free to exploit the information commercially by for example combining it with other information". Although we are not primarily concerned with commercial enterprise in this scenario description, the user could, with some justification, argue that s/he is not doing any more than the licence says that s/he is allowed to do.

At this stage in the process we consulted the UK Information Commissioner's Office and the ensuing discussion led to the fairly clear interpretation that the mere act of identification did not in itself constitute a breach of the DPA and did not in itself mean the user becomes a data controller. In the UK Data Protection Act (1998) the term *data controller* means "a person who – either alone or jointly or in common with other persons – determines the purposes for which and the manner in which any personal data are, or are to be processed". Similarly, EU Directive 95/46/EC defines it as: "the natural or legal person, public authority, agency or any other body which alone or jointly with others determines the purposes and means of the processing of personal data."

Of course the above situation would be radically transformed depending on what the user decides to do next. Game theoretically they have at least seven choices:

1. Do nothing
2. Use information disclosed about the respondent for some secondary purpose

3. Inform the respondent
4. Inform the data provider
5. Inform the information commissioner
6. Publicise the breach in some way
7. Some combination of the above in parallel or in sequence.

Each of these will lead to complex games with different combinations of players and each has different legal and material consequences. Specifically, some of them could lead to the user becoming a data controller. Full analysis of this would require a significant piece of work using Mackey's (2009); Mackey and Elliot (2010) game-theoretic framework and was beyond the scope of this project. However, it seems plausible that a user who accesses OG data and uses response knowledge to identify a respondent in that data could avoid any significant costs through careful strategy selection.

To summarise: (i) with EUL data the expected number of users with response knowledge will be small and the users are constrained from reidentification by the licence (and possibly sanctions), but (ii) with the OGD licence the number of potential users is greater by several orders of magnitude and therefore the number of expected potential users with response knowledge is also much greater. The OGD licence does not constrain the user from reidentification and it appears that they would not be constrained by the DPA until after the identification had occurred.

Given that response knowledge attacks are intrinsically higher risk than any orthodox attack scenario, these became the basis for our penetration test.

4. Methodological Approach to the Penetration Test

Prior to the research taking place, the approach was scrutinised closely by the ONS Ethics Committee. The Committee included a legal representative, senior staff with knowledge of social surveys and others with considerable experience of ethical issues in research. The Committee considered the experience of the researchers, in particular the fact that they had carried out research with ONS in the past, and had previously handled sensitive data appropriately. There is also a long-standing data-handling agreement between ONS and the University of Manchester regarding the latters's secure lab for disclosive data. We had also received written confirmation from the ONS Information Asset Owner for this work to go ahead. ONS Legal Services provided assurances on the legality of carrying out this research under the Statistics and Registration Service Act (SRSA), Section 38: "Use of information by the Board".

ONS staff extracted a set of fifty respondents from each of the surveys at random but excluded respondents who had not consented to further research. Respondents were selected from outside the researchers' areas of residence, in order that no respondent would be known to any of the researchers. These two files were passed to the attack simulation team at Manchester University under strict conditions, in particular that the files must be kept secure, must not be passed to anyone outside the immediate research team, and must be destroyed at the end of the work. The files were transferred securely by courier, the data encrypted and password protected. There was an additional proviso that, for the purposes of this research only, the researchers could attempt to identify individual respondents in the EUL files. The two files were:

The LFS ID file consisting of a list of 50 names, addresses and phone numbers.
The LCF ID file consisting of a list of addresses only.
The simulation consisted of four phases:

i) A search phase where intensive web-based searching was conducted on each of the 100 identifiers. On average, half a day of researchers' time per identifier was spent on this.
ii) Commercial data was purchased from a lifestyle database company corresponding to each of the 100 addresses in the identification files.
iii) The resultant information was matched against the microdata and reidentification attempted.
iv) The matches were verified by ONS.

4.1. The Search Process

In this section we describe the search method and recording system that we used. An initial pilot search was undertaken and the search toolset refined in order to maximise the data returns relative to the search effort involved.

For the cases in the LFS ID file it was necessary to verify whether the named respondent was resident at the property and in the case of the LCF ID file to identify all residents at the property at the time of the survey. For both identification files, attempts were made to gather data on all cases, including the hierarchical structure of the household. At least basic details were successfully gathered on most cases, such as approximate age and/or names of (at least some of) the other household members. On average, in the first round of searches each of the 100 cases had about three hours search time allocated to it. In the second round of searching, ten cases from each file were identified as worthy of extended search time, because the initial search indicated that they were highly visible and therefore more information about them was likely to be found. These cases received an additional 3–5 hours of search time.

4.1.1. Design Process

A search database was developed in order to systemise the search and to maximise the return relative to effort. An Excel workbook was created for each individual case.

There are a number of inherent biases associated with the use of different search engines. These biases include ranking metrics (Vaughan and Thelwall 2004; Vaughn and Zhang 2007; Bar-Ilan et al. 2009), business links associated with the searched sites and 'pay to promote' or sponsored search return services versus organic search results (Ma et al. 2010; Agarwal et al. 2011; Tarantino 2013). Some business and subscription-based websites will use this feature, thereby skewing the search results. Other sites, or individual users of sites, may opt to have their associated data blocked, restricted or removed from search engine rankings, thereby skewing data results further. Using multiple search engines does however increase the information returned. This was achieved by using meta-search tools, a selection of search engines and direct searches of specific sites such as *Facebook*, *LinkedIn*, *192* and ancestry search sites. These are detailed in the appendix.

Several test runs were carried out on pilot data to: (i) minimise the search tools used, (ii) maximise the likelihood of positive returns, and (iii) check accuracy. The most

important initial search source was *192* as this supported the cross-referencing of names with addresses and provided an age bracket, all essential basic information to assist further searching.

4.1.2. People and Business Search Tool – 192

For this crucial search tool the following information was retrieved: name, address, estimated age, estimated length of residence, data gathered from the electoral roll, recent nearby house sale prices, business details and director report information.

A recent addition to the data that one can obtain from *192* is a 'background report' on each household resident. This report is available at a relatively low cost (£29.94) and draws information into a single report from a range of public sources, that is, edited Electoral Register, company house information, D&B company listings, 118 Data Resources, Local Data Company, Land Registry, Callcredit plc, HALO Mortality file, The Insolvency Service, and The Registry Trust. The report sets out an individual's and their co-residents' (if any) name, address, length of residence, age in five-year bands, mortality, their solvency status, disqualified director status, home ownership status and whether they have any county court judgements against them. Although these sources of information are publically accessible in the UK, pulling together this information is time consuming and requires some understanding of where and how to access public data sources and knowledge of how to cross-reference sources to verify the information obtained. Thus, the *192* background report makes data collection of this type of information significantly easier. We did not use the reports in our searches but we did purchase several reports to assess the information they offered.

Other data sources were used to cross-reference the data obtained from *192* or to fill in the gaps where it was missing. *Ancestry UK* is a good source of information for gathering more accurate information on date of birth, marital status and family details. It should be noted that it was difficult to search for women or children using this site due to the ambiguity over marital status and married names. Again, if the name is common there will be too many results to search effectively. It is much more accurate with older people who have unusual names and who were less likely to migrate.

The land registry was the most reliable source for both verifying residency and home ownership. It is cheap to use (£3 per search) and has approximately eighty percent coverage. In terms of public sources, *Zoopla* was particularly useful for basic property information, as was *Google Maps*. With street view one can see the property and sometimes vehicles and make inferences about affluence and the presence of children.

The cost for using 192 is £89.94 for 100 credits allowing full access. Your credit is depleted every time you search and even searching with an address and full postcode very rarely returns a correct match within 192. Typically you need to search multiple times (approx. 4+) and you are likely to need to trawl through potentially very long lists of similar names and addresses. Credit is depleted very quickly and so 100 credits do not last very long. Individual background reports can be purchased at a cost of £29.94 each. These reports bring together information from numerous public sources such as Companies House and the Individual insolvency register. UK Ancestry was used to corroborate findings with 192 at a cost of £18.95 per month and was purchased for a period of three months.

4.1.3. Social Media

People are becoming more aware about their privacy settings when using social networking sites such as *Facebook* and *LinkedIn*. Recent studies verify that behaviour in social networking site use is changing and individuals are being encouraged to protect their privacy from different sources (see, for example Moreno et al. (2012) and Whipple et al. (2012)). This clearly inhibits intruders gathering information in this way.

With *Facebook* however, even when security settings are high, it may still be possible to gather data on location, address, employer, likes, films, pages, groups, to view photos, see posts by others on a page and so on. There were only a few cases in the study with *Facebook* accounts, but those were a good source of contextual information for those cases.

For *LinkedIn*, if security settings are activated, little information can be seen unless you are a group member. Searching requires registration and a tracing tool is used which lets a person know when you have searched for them. When the security settings are not activated a range of information is potentially available, such as workplace location, job type/title and education. We found a small number of cases where the privacy settings were not activated.

4.1.4. Difficulties Associated With the Search Strategy

It should be noted that searching in the manner described above is not without its difficulties and/or limitations. With our key search tool *192* we encountered problems:

Stability of the information: Searching within *192* does not always return the exact address/name, and multiple attempts were sometimes needed. We found that searching outside *192*, via *Google* for example, was more reliable but then you have to pay to view any information beyond address and name. It is worth noting that there is a cost for each search on *192*.

Inaccuracy of information: We found conflicting name and age profile data for a small number of our cases, where several results were returned for similar addresses with the same person and co-residents named but with contrasting age profiles. These inaccuracies made it difficult to be confident about the data as there was no way to cross-check for accuracy. Similarly, the data on directors was not always reliable when cross-referenced with business directory sources. Common names, names associated with famous people or ambiguous names, such as those that represent objects, were difficult to search as they produced a large number of returns that were difficult to cross-check for accuracy (given the time contstrains).

In terms of specific groups, it is more difficult to search for middle-aged women, especially if their marital status is unclear or unknown. Often once a ‘husband’ is identified it is easier to confirm the woman’s identity, including date of birth and so on, by tracing her through her husband’s marriage records. This is often the only way to track a woman’s birth record. Even if there are children, this group of women can be difficult to identify without knowing their maiden name to check the birth records against.

Beyond the search limitations there are two other research issues that are worth noting. The first of these is time – or more specifically, the search time used in this study. It is entirely possible that an intruder would be willing to spend more time searching than we did and it would seem more likely that the search frame would be much smaller than our 100 cases.

A second point worth making in relation to our search strategy relates to the issue of who is online, and thus who is more likely to be captured. Studies have shown that people from different age cohorts use the Internet in different ways and develop online trust in dissimilar ways (Obal and Kunz 2013). Fewer older people have online social media accounts, and those that do tend to either use them infrequently or have very little information associated with them.

4.2. *The Commercial Data Purchase*

As well as searching public data sources online for information for the 100 cases in the identification files, we also purchased commercial data from the UK lifestyle data company, CACI. The rationale for this was twofold: (i) to consider what additional information might be available through this route, and (ii) to develop first-hand knowledge and experience of the process of obtaining commercial data.

In terms of the second rationale there were two constraints, the first of which was cost. The minimum purchase from CACI is £1,200, which will get you up to 1,000 names and addresses. It typically costs thousands of pounds to access any variables associated with the names and addresses purchased and the number of variables given is usually limited to between 5 and 20. This makes it unlikely that an intruder searching for a single case is likely to consider purchasing this type of data.

The second constraint relates to the purchasing process itself, which is time consuming and requires several screening stages. The first stage requires that you explain why you want to purchase data. At the next stage, you are allocated an account manager who discusses at length why you want the data and negotiates what variables you can have access to and for what time period. The final stage requires you to sign a contract outlining the cost involved, the variables and cases requested and the intended use. Accessing the data took several weeks and many emails and phone calls. The length of this process and the hoops that the user is required to go through on top of the cost are likely to deter all but the most determined intruder.

We informed CACI that we wanted to purchase data for a study looking at what information they held and what disclosure risk (if any) it might pose. We were able to negotiate for the minimum order value the names and addresses for Sample 1 and 2 and those of any other residents at the properties. We were also given all the variables held on each resident at the 100 households. The available variables can be summarised in the three categories: demographic, lifestyle, and socioeconomic.

The key variables present in the dataset and considered useful for matching purposes were: age (five-year bands), sex of each adult, number of children in household, number of bedrooms, social grade, course occupational grouping, years at property (four bands), tenure (three bands), house type, and number of cars (0, 1, 2+).

4.3. *The Matching Process*

We initially explored the idea of using automated matching techniques. However, it was assumed that the intruder we were simulating was not a technical expert and therefore using a probability based approach was not realistic.

Even more importantly, it became evident that a non-automated approach was actually going to be more productive. An example of why this is the case can be found with Case 4 from the LFS sample. This case threw up an enormous amount of data, some of it contradictory. In particular, the crucial piece of information about Case 4's age seemed unreliable on *192.com*. This left us with several possible people who corresponded to Case 4 in the dataset. However, since we knew that Case 4 was in the sample and the information on Case 4's marital partner seemed solid, it was possible to look for household age-relationship combinations that matched Case 4's spouse's age (which appeared reliable). Fortunately most of the combinations were a priori unlikely (large age differences) and only one threw up an actual match. This was then checked for against other information about both Case 4 and his spouse.

In other cases where we found multiple matches on the available information, we were able to examine the dataset to look for an additional variable that differentiated those matches and then go out to look for information in a more targeted way for the case on that variable. This ability to go to and fro between the data, the matching information and the world made the approach much more like a piece of detective work and less like an orthodox statistical matching process. Because of the ethical and time constraints we were under on the project, we could not go as far with this as we might have done otherwise.

Once the attack simulation team had reduced the number of possible matches down as far as they could, they then made an assessment of the certainty of the best matches. This was essentially a subjective expert judgement but was based on a three factors: (i) the closeness of the next best match; (ii) prior knowledge of data divergence issues (time lag between data collection and the survey date, coding mismatches, etc.); and (iii) confidence in the data we had collected (was the information contradictory, was there doubt over whether we had found the correct person, etc.). The matcher in fact assigned a score on a 100-point scale to represent their confidence, which roughly translates into a subjective estimate of the probability of a match being correct. For the purposes of presentation here this scale is collapsed into High (70–100), Medium (50–69), Low (15–49), and Very low (0–14). It should be stressed that no algorithm was used here – the confidence level that was assigned was essentially based on expert judgement by the researcher, taking account of the above factors. A possible extension of this work would be to investigate whether this expert judgement is easily convertible into an algorithm. However, here we are simply concerned with whether matches could be achieved with minimum technical mediation.

The matches were then verified by the ONS team.

5. Matching Results: LFS

5.1. Stage 1: Openly Available Information Only

At Stage 1 we considered only openly available information that we had collected from the Internet.

In total, matches against nine records were attempted, since for the other 41 cases insufficient good-quality information was obtained to make a match attempt viable. Of the nine match attempts, eight produced a correct match, although in two cases it was one of a multiple match.

Summary information is given below. The critical point here is that if the matcher's confidence was high then the matches were successful. These are invariably cases where large amounts of good-quality information were obtained.

5.2. *Stage 2: Adding in the Commercial Dataset*

Before the Stage 1 matches were verified, the process was repeated, this time adding in the commercial data. This increased the number of correct matches to 14. Adding the commercial data did not have a completely monotonic effect on the matches; two of the matches that were correct using only the openly available information were not even attempted with the commercial data because the commercial data provided contradictory information, reducing the certainty. This nonmonotonicity may seem counterintuitive but is related to more a general phenomenon observed, for example by Elliot (2009). Essentially, increasing information has a diminishing return on the power of a set of key information (because information about people is correlated) but the impact on data divergence is linear. So at some point the noise created by the divergence exceeds the information gain from the increasing key size. Where that point is varies depending on the level of divergence, the level of correlation between the key variables and the power of the key variables. On the other hand, fuzzy and probabilistic matching techniques can reverse the process, trading lower precision for higher recall.

The headline results are that by using the openly available information, six of the 50 records were correctly matched one-to-one (12%). Using the commercial data as well pushed this up to 14 (28%). These headline figures however disguise a more significant fact: that the slope of "matchability" is quite steep; the precision rate for high-confidence matches was 100%. The reason for this steepness is primarily because of the amount and quality of the information obtained. Another factor was household size.

6. Matching Results: LCF

The process for the LCF matches was slightly different. The file was not hierarchical and so lacked that defining feature. On the other hand, it did have a low-level geographical indicator on it: the Output Area Classifier (OAC). Obtaining an OAC from a postcode is quite easy once you know how to do it, but it is certainly not obvious and would not necessarily be something that was available to a non-expert. However, a data journalist or similar should be able to obtain the information. We therefore ran two different scenarios: with and without the OAC codes. As it turned out, the OAC was an incredibly useful differentiator key.

The second feature for this dataset was that ONS only provided the simulated attack team with addresses (no names) and this reduced the certainty of the information, which was reflected in the confidence levels that we recorded.

A third difference in the process here was that where there were multiple matches of equal certainty these were recorded as single joint match. For the purpose of comparison, these are turned into effective confidences by simply dividing the total by the number of matches.

At Stage 1a (without OAC codes) there were a possible 20 matches against eight addresses. The mean effective confidence was 16%, which meant that we were predicting

Table 1. Matching attempts using openly available information against the LFS dataset

Confidence level	1-to-1			1-to-*n*		No match attempted
	Correct	Incorrect	Precision	Correct	Incorrect	
High confidence	5	0	1.00	2	0	
Medium confidence	1	0	1.00	0	0	
Low confidence	0	0	–	1	0	
Very low confidence			–			41
Overall	6	0	1.00	3	0	41

that we would obtain 3.3 matches. In fact we obtained two. This information is shown in Table 3. This file looks reasonably safe against this simple attack.

However, when the OAC code is added the situation looks very different, as we can see in Table 4. A total of 42 matches were made against 27 addresses. 16.55 matches were predicted to be correct and in fact 18 were. As with the LFS matches, the high-confidence matches were more likely to be correct.

7. General Discussion

The headline finding of this study is that an intruder could, even with partial response knowledge (an address) and using only publically available and/or purchasable information, obtain some correct high-confidence matches without the use of sophisticated matching software. The overall precision rate for high-confidence matches over Tables 1 to 4 is 91%.

This is an important finding. The data here are fairly standard social survey data and contain a mixture of mundane and sensitive information. A correct match against either file would yield income and health information about the target as well as information about other family members, including children. So beyond the obvious legal requirements, the data custodian has a clear duty of care to respondents.

There are seven caveats that must be placed on the details of our headline finding. First, the datasets were older than the public and commercial data (15 months older in the case of the LFS and 27 months in the case of the LCF). This will have increased the data

Table 2. Matching attempts combining openly available data and the CACI commercial dataset

Confidence level	1-to-1			1-to-*n*		No match attempted
	Correct	Incorrect	Precision	Correct	Incorrect	
High confidence	10	0	1.00	1	0	
Medium confidence	1	1	0.50	0	0	
Low confidence	1	1	0.50	1	2	
Very low confidence			–			42
Overall	12	2	0.86	2	2	42

Table 3. Matches against the LCF without OAC codes

Confidence level	1-to-1			1-to-*n*		No match attempted
	Correct	Incorrect	Precision	Correct	Incorrect	
High confidence	0	0	–	0	1	
Medium confidence	1	1	0.50	0	1	
Low confidence	0	2	0.00	1	1	
Very low confidence			–			42
Overall	1	3	0.25	1	2	42

divergence. In many cases this would have caused a "no match attempted" outcome and therefore will not have affected rates (against confidence), but it will have affected the number of matches attempted and this was factored into the overall confidence level.

Second, the person doing the matching was a statistical disclosure control expert. So, although he restricted himself to unsophisticated manual matching techniques, he was not able to switch off his understanding of data processes, and in particular concepts such as rareness and uniqueness. This would have made confidence estimates more accurate than might be expected for an intruder without that expertise.

Third, the study team was restricted to carrying out legal and ethical actions. We could not, for example, call the named householder. We drew a strict ethical line around our search behaviour so we did not, for example, create fake accounts, attempt to befriend anyone or pose as another person, all of which could potentially yield further information. We also decided that we would not carry out site visits as this would be potentially intrusive. A malicious intruder would not be restricted in the same way. Relatedly, a wealth of technology sources is available to knowledgeable users that could increase the likelihood of an intruder gaining access to an online account such as *Facebook* to access the data. This study did not use any such technology (for a useful background to the issue of using socialbots to hack social networking accounts, please see Boshmaf et al. 2013).

Fourth, the data gatherer was restricted in time by the need to gather information against a representative number of cases. An intruder who simply wanted to identify a single individual could focus a lot more resources on that case.

Table 4. Matches against the LCF dataset with open data using the OAC code

Confidence level	1-to-1			1-to-*n*		No match attempted
	Correct	Incorrect	Precision	Correct	Incorrect	
High confidence	5	2	0.71	3	1	
Medium confidence	8	0	1.00	1	1	
Low confidence	1	3	0.25		2	
Very low confidence						23
Overall	14	5	0.74	4	4	23

Fifth, every dataset will be different in terms of its properties (variables, sample size, data structure, etc.) and those properties will interact with the likelihood of correct identifications.

Sixth, the study is a snapshot, albeit a compelling one. The availability of data in the public domain is changing constantly, with the general trend being upwards. This will tend to increase the risk associated with this type of attack. The importance of this issue is increased by the fact that any move from EUL to OGD *for any given dataset* is a one-shot decision. Once the data are released, then the decision is effectively irreversible.

Seventhly, although we were considering response knowledge we were not able to mimic the entirety of what an individual might know about a respondent, only what is available more or less publically. It is likely that an intruder with response knowledge would also have other personal knowledge about the respondent. A potentially interesting extension to the current study would involve re-contacting respondents and asking them to nominate a friend, neighbour, or colleague and then asking the nominee to complete the survey as if they were the respondent.

Some of these factors mitigate the risks indicated by these findings, while others exacerbate them. This makes drawing general conclusions from the **specifics** of the results reported here hazardous.

Nevertheless, the general shape of the results is indicative that moving the survey datasets from end user licence to open data, without any change in their content, would significantly increase the risk of a statistical disclosure on each such dataset and make the likelihood of a disclosure event far greater. The results of the study presented here do suggest that the level of detail on geographical variables and the level of information about household structure are two issues that would need to be attended to in a data-release decision. The restrictions placed on researchers under the EUL in these two surveys are therefore necessary in order to deter an attack, and to provide ONS with some sanctions in case of an attempted or claimed disclosure. As we lay out in Section 3, the response knowledge scenario makes sense with open data but not with EUL licensing.

The financial resources required for the entire study were modest. The main cost was the commercial dataset, which cost us £1200; in addition we spent under £100 on ad hoc services on *192* and the land registry. This works out at an average of £13 per record. In fact, for the LCF study we ended up not using the commercial data as it did not add any value, so the costs there were considerably cheaper.

To summarise, the study presented here provides an illustration of the importance of carrying out well-formulated penetration tests before decisions are made about data releases, particularly irreversible ones such as the release of a file of individual records as open data.

Appendix

List of search sites used to build identifying information

Search engines	Business finder	People finder
http://www.lycos.com/	http://companycheck.co.uk/index	http://www.192.com/
http://www.clusty.com/	http://www.kompass.com/	http://www.infobel.com/en/UK/
http://www.mamma.com/	http://www.hoovers.com/	http://www.linkedin.com/
http://www.metacrawler.com/	http://www.yalwa.co.uk/	
http://uk.search.yahoo.com/	http://www.infobel.com/en/uk/Business.aspx	**Ancestry**
http://www.bing.com/	http://www.yell.com/	http://home.ancestry.co.uk/
http://www.google.com/	http://www.business.com/	
http://www.hotbot.com/	http://www.lexisnexis.co.uk/en-uk/home.page	**Facebook**
http://www.excite.com/		https://www.facebook.com/
http://www.ask.com	**Estate Agents**	
	http://www.zoopla.co.uk/	**Other**
	http://www.rightmove.co.uk/	http://www.iannounce.co.uk/

8. References

Agarwal, A., K. Hosanagar, and M.D. Smith. 2011. "Location, Location, Location: An Analysis of Profitability of Position in Online Advertising Markets." *Journal of Marketing Research* 48: 1057–1073. Doi: http://dx.doi.org/10.1509/jmr.08.0468.

Backstrom, L., C. Dwork, and J. Kleinberg. 2007. "Wherefore Art Thou r3579x?: Anonymized Social Networks, Hidden Patterns, and Structural Steganography." In Proceedings of the 16th international conference on World Wide Web, 8–12 May 2007, Banff, AB, Canada. 181–190. Available at: http://dl.acm.org/citation.cfm?id=1242598 (accessed 9 November 2015).

Bar-Ilan, J., K. Keenoy, M. Levene, and E. Yaari. 2009. "Presentation Bias Is Significant in Determining User Preference for Search Results-A User Study." *Journal of the American Society for Information Science and Technology* 60: 135–149. Doi: http://dx.doi.org/10.1002/asi.20941.

Boshmaf, Y., I. Muslukhov, K. Beznosov, and M. Ripeanu. 2013. "Design and Analysis of a Social Botnet." *Computer Networks* 57: 556–578. Doi: http://dx.doi.org/10.1016/j.comnet.2012.06.006.

El Emam, K., E. Jonker, L. Arbuckle, and B. Malin. 2011. "A Systematic Review of Re-Identification Attacks on Health Data." *PloS one* 6(12) : e28071. Doi: http://dx.doi.org/10.1371/journal.pone.0126772.

Elliot, M.J. 2009. "Using Targeted Perturbation of Microdata to Protect Against Intelligent Linkage." In Proceedings of UNECE Work Session on Statistical Confidentiality, 17–19 December 2007, Manchester. Available at: http://www.unece.org/index.php?id=14503#/ (accessed 14 December 2014).

Elliot, M.J. and A. Dale. 1998. "Disclosure Risk for Microdata Report to the European Union ESP/204 62 361–372." Available at: https://www.escholar.manchester.ac.uk/uk-ac-man-scw:19b497 (accessed 9 November 2015).

Elliot, M.J. and A. Dale. 1999. "Scenarios of Attack: the Data Intruder's Perspective on Statistical Disclosure Risk." *Netherlands Official Statistics* 14: 6–10. Available at: http://bit.ly/1ScX0cS (accessed 9 November 2015).

Elliot, M.J. and E. Mackey. 2014. "The Social Data Environment." In *Digital Enlightenment Yearbook*, edited by K. O'Hara, S.L. David, D. de Roure, and C. M-H. Nguyen. 253–263. Doi: http://dx.doi.org/10.3233/978-1-61499-450-3-253.

Gymrek, M., A.L. McGuire, D. Golan, E. Halperin, and Y. Erlich. 2013. "Identifying Personal Genomes by Surname Inference." *Science* 339: 321–324. http://dx.doi.org/10.1126/science.1229566.

Ma, Z.M., G. Pant, and O.R.L. Sheng. 2010. "Examining Organic and Sponsored Search Results: A Vendor Reliability Perspective." *Journal of Computer Information Systems* 50: 30–38. Available at: http://bit.ly/1MSpcni (accessed 9 November 2015).

Mackey, E. 2009. *A Framework for Understanding Statistical Disclosure Control Processes: A Case Study Using the UK's Neighbourhood Statistics*. PhD Thesis, University of Manchester. Available at: http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.502255 (accessed 9 November 2015).

Mackey, E. and M.J. Elliot. 2010. "The Application of Game Theory to Disclosure Events." *Proceedings of UNECE worksession on Statistical Confidentiality, Bilboa, December 2009*. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2009/wp. 40.e.pdf (accessed 09/11/2015).

Mackey, E. and M.J. Elliot. 2013. "Understanding the Data Environment." *XRDS* 20: 37–39. http://dx.doi.org/10.1145/2508973.

Malin, B. and L. Sweeney. 2004. "How (Not) to Protect Genomic Data Privacy in a Distributed Network: Using Trail Re-Identification to Evaluate and Design Anonymity Protection Systems." *Journal of Biomedical Informatics* 37: 179–192. http://dx.doi.org/10.1016/j.jbi.2004.04.005.

Moreno, M.A., A. Grant, L. Kacvinsky, P. Moreno, and M. Fleming. 2012. "Older Adolescents' Views Regarding Participation in Facebook Research." *Journal of Adolescent Health* 51: 439–444. http://dx.doi.org/10.1016/j.jadohealth.2012.02.001.

Müller, W., U. Blien, and H. Wirth. 1995. "Identification Risks of Micro Data. Evidence from Experimental Studies." *Sociological Methods and Research* 24: 131–157. http://dx.doi.org/10.1177/0049124195024002001.

Narayanan, A. and V. Shmatikov. 2008. "Robust De-Anonymization of Large Sparse Datasets." In Proceedings of the 2008 IEEE Symposium on Security and Privacy, 18–21 May 2008, Berkeley/Oakland, CA, USA. 111–125. Doi: http://dx.doi.org/10.1109/SP.2008.33.

Narayanan, A. and V. Shmatikov. 2009. "De-Anonymizing Social Networks." In Proceedings of the 2009 IEEE Symposium on Security and Privacy, 17–20 May 2009, Berkeley/Oakland, CA, USA. 173–187. Doi: http://dx.doi.org/10.1109/Sp.2009.22.

Obal, M. and W. Kunz. 2013. "Trust Development in E-Services: A Cohort Analysis of Millennials and Baby Boomers." *Journal of Service Management* 24: 45–63. Doi: http://dx.doi.org/10.1108/09564231311304189.

Paass, G. 1988. "Disclosure Risk and Disclosure Avoidance for Microdata." *Journal of Business and Economic Statistics* 6: 487–500. Doi: http://dx.doi.org/10.1080/07350015.1988.10509697.

Tarantino, E. 2013. "A Simple Model of Vertical Search Engines Foreclosure." *Telecommunications Policy* 37: 1–12. Doi: http://dx.doi.org/10.1016/j.telpol.2012.06.002.

Vaughan, L. and M. Thelwall. 2004. "Search Engine Coverage Bias: Evidence and Possible Causes." *Information Processing & Management* 40: 693–707. Doi: http://dx.doi.org/10.1016/S0306-4573(03)00063-3.

Vaughan, L.W. and Y.J. Zhang. 2007. "Equal Representation by Search Engines? A Comparison of Websites Across Countries and Domains." *Journal of Computer-Mediated Communication* 12: 888–909. Doi: http://dx.doi.org/10.1111/j.1083-6101.2007.00355.x.

Whipple, E.C., K.L. Allgood, and E.M. Larue. 2012. "Third-Year Medical Students' Knowledge of Privacy and Security Issues Concerning Mobile Devices." *Medical Teacher* 34: e532–e548. Doi: http://dx.doi.org/10.3109/0142159X.2012.670319.

Received December 2014
Revised November 2015
Accepted November 2015