IJDC | *Peer-Reviewed Paper*

Examining Disclosure Risk and Data Utility: An Administrative Data Case Study

Michael Comerford
University of Glasgow

Abstract

The plethora of new data sources, combined with a growing interest in increased access to previously unpublished data, poses a set of ethical challenges regarding individual privacy. This paper sets out one aspect of those challenges: the need to anonymise data in such a form that protects the privacy of individuals while providing sufficient data utility for data users. This issue is discussed using a case study of Scottish Government's administrative data, in which disclosure risk is examined and data utility is assessed using a potential 'real-world' analysis.

Received 17 October 2013 | *Accepted* 26 February 2014

Correspondence should be addressed to Michael Comerford, Rm409 SAWB, School of Computing Science, University of Glasgow, Glasgow, G12 8QQ. Email: comerm@dcs.gla.ac.uk

An earlier version of this paper was presented at the 9th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v9i1.297
DOI: 10.2218/ijdc.v9i1.297

Introduction

In recent years the volume of electronic data available for analysis has grown exponentially; commentators now often refer to the data ‘deluge’ (The Economist, 2010), the data ‘explosion’ (Microsoft News Centre, 2013) and even a data ‘tsunami’ (Argonne National Laboratory, 2012). To support new data analyses also requires new data management techniques that can provide data in different formats, linked across different sources, and available in a variety of security settings. The latter, security settings, presents a particular challenge that cuts across different disciplines, including law, ethics, computing science and statistics, and as the number of potential data sources grows, so too have the concerns about data privacy and confidentiality. Ohm (2009) highlights the need to conceptualise the balance between protecting the privacy of individuals and the utility of research data; 100% anonymity results in 0% utility and vice versa. The purpose of this paper is to demonstrate an assessment of both disclosure risk, the threat to privacy, and data utility. It will be shown that choices taken in the process of maintaining statistical confidentiality need to be set in context and understood within this conceptual balance, something that data controlling organisations often struggle to achieve when the stick of sanctions for a breach of privacy is far more tangible than the carrot of providing good data utility. To achieve this purpose, a case study of governmental administrative data from the education field is referenced, while the full case study forms part of an ongoing PhD thesis in this area.

Disclosure Risk

Assessing the risk of disclosure is a complex task that is often riddled with unexplored assumptions about what constitutes disclosure, and these are often conflated with the impact of disclosure. The definition of disclosure used in this paper is:

> ‘The inappropriate attribution of information to a data subject.’ (Statistics Netherlands, 2009).

Therefore, in assessing disclosure risk the aim is to determine the risk that an intruder could obtain information that they can attribute to a data subject. A common intruder scenario is defined as an intruder with *a priori* knowledge of a data subject, matching this knowledge to an anonymised data source and learning something new about their target (Elliot and Dale, 1999). To illustrate this point, an intruder might have access to the electoral register, for example, with details of a target’s name, address, nationality and age. This knowledge could then be matched against an anonymised data source that contains health data, and if a match was found the intruder could attribute a sensitive medical condition from that source to their target. It is also worth noting here that using intruder scenarios and interrogating the data directly are complimentary methods to other forms of disclosure risk management, such as access controls and resource management, for example see Duncan and Pearson (1991).

Having defined what is meant by disclosure risk, it is also prudent to understand the measurement of this risk. Approaches to this range in their complexity and applicability: at the lower end of complexity are measures such as *k*-anonymity (Samarati and Sweeney, 1998) and its variants (Domingo-Ferrer and Torra, 2008), while in the upper echelons of complexity are models such as those proposed by Li and Li (2009). As our work forms part of a wider eScience project, the need for clearly defined and communicable concepts is a priority in order to ensure the interdisciplinary stakeholders of our research can feel confidence in their application. Therefore, *k*-anonymity is used. *K*-anonymity represents a threshold approach to disclosure risk and is defined such that a record satisfies *k*-anonymity if there are $k - 1$ other records with the same characteristics. These characteristics are often defined as a re-identification key and include quasi-identifiers such as age, gender or ethnicity. As a small example, see Table 1.

Table 1. A *k*-anonymity example.

Age	Gender	Nationality	Postcode	Satisfies *k*-anonymity
27	Male	British	G12	yes
27	Male	British	G12	yes
27	Male	British	G12	yes
43	Female	Irish	UB7	no
43	Male	Irish	UB7	no
43	Female	Irish	SW1	no
54	Female	Canadian	SG5	no

In this synthetic example the threshold is set as $k = 3$, so for a record to satisfy the *k*-anonymity requirement there must be two other records with the same characteristics. The first three records satisfy this requirement; Records 4–6 have matching ages and nationalities, however their gender and postcode differ; and Record 7 is unique across all fields except gender.

Statistical Disclosure Control

There is a growing body of literature on methods used to prevent published data inadvertently disclosing information about individuals, which is again drawn from across different disciplines, for example in computing science (Stell, Sinnott, Ajayi, and Jiang, 2009), statistics (Fienberg and Makov, 1998), and the social sciences (Skinner and Elliot, 2002). These methods are collectively referred to as statistical disclosure control. Commonly, these methods are divided into those that are perturbative and non-perturbative. Global recoding of variable values, such as the aggregation of age in years into five-year age categories, would be an example of a non-perturbative method, whereas cell suppression (the removal of values) in a data table would be perturbative. This simple dichotomy does not capture the full scope of methods: for example, those involving synthetic data (Reiter, 2002) could form a group of their own. However space precludes a full methodological survey here. In the case study presented we will use recoding and suppression to demonstrate the common approaches used by administrative data controllers in the UK context, these are often directly referred to in the statistical disclosure control policies of such bodies NHS Scotland Information Services Division (ISD) (2012).

Data Utility

Defining and measuring data utility is perhaps more complex than disclosure risk because disclosure can be neatly defined. Utility is often highly subjective, dependent on the needs of the data users. Again, there are approaches from different disciplinary areas (Duncan, Keller-McNulty, and Stokes, 2001; Purdam and Elliot, 2007; Rastogi, Suciu, and Hong, 2007). Due to the difficulty of predicting how a data user might want to analyse the data, the approach taken here is similar to Purdam and Elliot (2007) in that utility is measured by the effect disclosure control methods have on 'real-world' analyses. This also represents a more collaborative approach between data controllers and users, akin to that used in secure physical and virtual research environments, where data are not publicly released but stored and analysed using secure servers. This utility approach is also supported by large data projects, such as the Integrated Public Use Microdata Series (IPUMS) (McCaa et al., 2013). Data utility, in the case study presented below, is measured with reference to a potential research question, and the differences in the conclusions drawn from the data with disclosure control applied and the original raw data are examined.

The Data

The subject of our case study are data from the Scottish Government, drawn from administrative sources about 'looked after children'. Looked after children, in this context, are defined by the Children (Scotland) Act 1995[1], and are those in the care of their local authority.

The data are drawn from an annual survey of 'looked after children'. The survey was collated centrally by the Scottish Government from local authority administrative systems. The Government then publish aggregate statistics from the survey (Scottish Government, 2011).The dataset used had 8,185 individual records from the 2010-2011 survey, and included a range of demographic variables and variables on educational attainment, details of placements, and socio-economic background. The data were accessed on site at the Scottish Government and research outputs were vetted by government staff for possible disclosure threats. All *k*-anonymity analyses were carried out using NIAH[2], a *k*-anonymity algorithm written in java, and statistical analysis of the results was carried using the Government's SAS statistical package servers.

Constructing a Re-Identification Key

As discussed above, to implement *k*-anonymity a decision is made about the variables in the dataset that pose a risk in terms of there usefulness to an intruder. Here we have simplified the approach taken by Elliot and Dale (1999) and constructed two intruder scenarios with variables associated with each:

1. **Relatives:** Our first intruder scenario was based around a relative of a particular child looked after by the state attempting to re-identify the child's record to discover some details about their placement in care. To this intruder scenario we attribute the following potential key variables that could be mapped to our data: date of birth, gender, ethnic group, national identity, main disability and datazone (low level geography).

[1] See: http://www.scotland.gov.uk/Topics/People/Young-People/protecting/lac/about

[2] NIAH: http://sourceforge.net/projects/niahsdc/

2. **Strangers:** A stranger could be seeking information for a number of reasons, including journalistic, activist or political motivations. In legal cases involving young people, for example, efforts are made to preserve their anonymity and therefore disclosure from anonymised sources could be an avenue of attack. To this intruder group we attribute the variables: gender, age (specifically not date of birth), ethnic group (potentially at an abstracted level of detail), and local authority.

Table 2 shows the list of key variables and their relative levels of detail. Each key variable will be described in turn and any potential disclosure risk issues are presented below. Once the key variables have been addressed individually, the *k*-anonymity analysis can begin to present the disclosure risk issues of the key variables cross classification.

Table 2. Key variables for disclosure risk analysis.

Variable	Description
Lacode	Local Authority Code (3 digit)
Lacdob	Date of Birth (YYYY-MM-DD)
Lacgender	Gender (M/F)
Lacethnicgroup	Ethnic Group (17 categories - 2 digit)
Nationalidentity	National Identity (9 categories - 2 digit)
SParlCon	Scottish Parliamentary Constituency (2 digit)
Datazone	Datazone (small area geography)
MainDisability	Main Disability (2 digit)
StudentStage	Student Stage e.g. Primary 1 (coded as 2 characters)

Geographical details

The geographical variables give us an indication of where risk might be concentrated. It should be noted that it would be unusual to release data with geographic areas smaller than local authority (LA) for this type of data. However, the presence of three geographic variables gives us the ability to see what the effect of different levels of geography have on disclosure risk in this particular case. For our initial sweep of key variables we consider the number of records with a given value as a crude indication of disclosure risk. For example, LA 330 (Orkney) had only 23 cases; the potential for re-identification given knowledge of the LA alone is greater than for LA 260 (Glasgow City) that carried 1,604 cases.

Scottish Parliamentary Constituency (SPC) provides a different geographical distribution with a larger number of areas. However, constituency 67 (Shetland) has the lowest number of cases but at 16 this does not look very different from the 23 cases for Orkney at the LA level. Datazones provide a particular challenge because of their relatively small size in the dataset: there are around 3,000 datazones and 98% of those had less than ten records associated with them.

It should be noted that different geographical partitioning with a similar number of areas can have completely different risk profiles. With this comparison of LA and SPC we can also start to see the counterintuitive effects of geographic detail on disclosure risk highlighted by Elliot et al. (1998) and Witkowski (2008). For example, a group of records with particularly unique characteristics present at the datazone level might also be particularly unique at LA level, despite the significant aggregation of geography. In addition, if multiple geographical variables were released, or variables that could act as a proxy for geography were included, such as the Scottish Index of Multiple Deprivation (SIMD) or Urban-Rural indicators, then particular attention should be paid during any disclosure risk assessment in case records can be better located by an intruder combining their knowledge of geographic details (Steel and Sperling, 2001).

Demographic details

Age is a common key variable across a number of datasets and it is relatively easy to obtain from public sources. Given that we were working with education data concerning young people, the data was fairly homogeneous. However, immediate areas of concern, with regard to statistical disclosure, are concentrated in the upper end of the distribution with both ages 20 and 21 having counts of less than 100 records.

Gender itself does not normally present a disclosure risk, but it can be a mitigating factor when cross-tabulations with other key variables are taken into account. As with age, gender is a fairly easy variable for an intruder to know *a priori*. The distribution of gender in this case is 54% male and 46% female. To demonstrate the potential problem of combining the gender key variables with others, Figure 1 shows gender in combination with age for the City of Edinburgh LA. In terms of a potential intrusion scenario, knowledge of a target's gender can potentially half the ambiguity associated with a match. From Figure 1, there are 72 individuals aged 14; if the target is a female, the target group is reduced to 38.

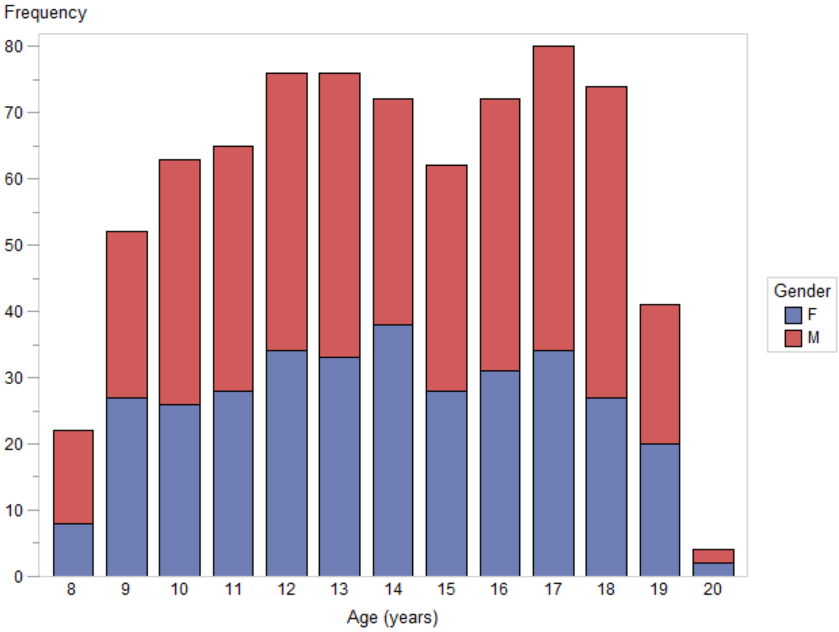

Figure 1. Gender distribution by age for City of Edinburgh local authority (N=759).

Ethnicity data presents a particular problem for disclosure control in Scotland especially, because ethnicity in Scotland is relatively homogeneous, with the majority of the population falling into a 'white' category with few outliers. This data is often removed from research datasets because the disclosure risk outweighs the analytical utility unless the study has a particular focus on ethnicity. This lack of data sources has been highlighted, especially in the health field (Ranganathan and Bhopal, 2006). It is possible to counter this argument about homogeneity. The categorisation of ethnicity into a large number of very specific groups, which are subjectively assigned, potentially introduces a significant amount of ambiguity for an intruder to navigate. Even if the level of detail is reduced to ‘white’ and ‘non-white’ an intruder cannot be certain which category their target might be assigned.

National identity presents a similar challenge to ethnic group because of the relative homogeneity of the data. However, the ability for an adversary to know national identity *a priori* is less obvious than for ethnic group because of the greater subjectivity in the assigned value. It is again worth noting that the suppression of a variable such as ethnic group or national identity is not a universal rule. A study interested in the attainment of children from different ethnic groups or national identities might not be interested in gender or sub-national geographies, for example. Therefore, the level of detail required for effective analysis could be negotiated for specific research projects.

Throughout this case study greater attention is paid to a general public use case and the typical variables, such as age and gender, which are common to the majority of analyses, but this need for flexibility should not be forgotten.

Main disability is perhaps not a commonly analysed variable. However, given that this is an exploratory analysis its disclosive potential is considered. The scenario here is that that a target for re-identification has a disability known to the adversary that could aid their attack. The disability categories include visual and hearing impairments, physical disabilities, learning disabilities and social, emotional and behavioural difficulties. However, the majority of individuals have no disability recorded. Figure 2 shows a scenario in which the intruder knows the age, local authority and that the individual has a physical disability. The ambiguity is reduced significantly, in some cases down to one record.

Our last key variable is student stage, an indicator of what level of schooling the individual receives. This has been included as a proxy for age, because they are strongly correlated, and if student stage and age were released together it could undermine any disclosure control carried out on the age variable. If student stage is stacked against age, when age is recoded into five year bands, student stage can be used to partition the age bands into their constituent parts. This is more ambiguous than releasing age in years because of the possibility of individuals with different ages from the norm for their stage of education. However, it demonstrates the ability to undermine disclosure control methods using proxy indicators.

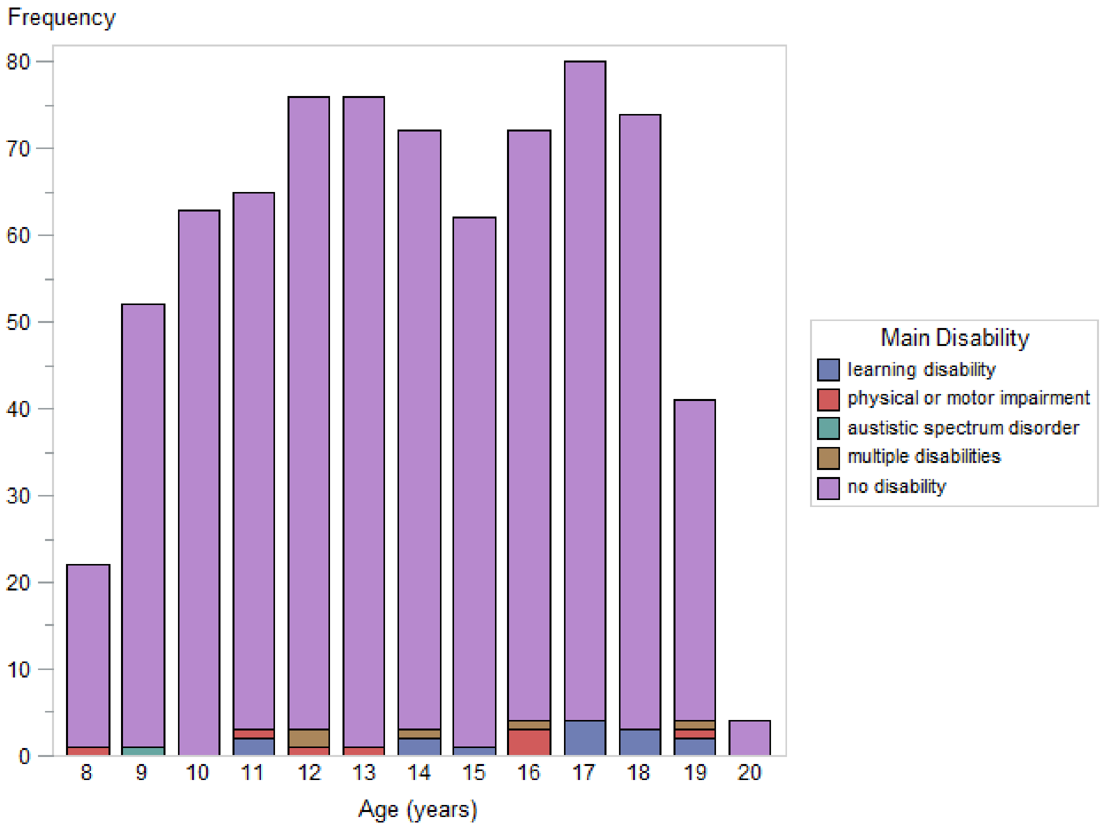

Figure 2. Main disability by age for City of Edinburgh local authority (N=759).

K-Anonymity Analysis

NIAH partitions the dataset into two parts: the 'safe' partition that satisfies *k*-anonymity and the 'at risk' for those records that do not. This done for a specified set of key variables and for a given threshold. Table 3 provides some summary details for a *k*-anonymity analysis of all the key variables by local authority and a threshold of $k = 3$.

From Table 3 the first point to note is the number of records deemed to be 'at risk' of disclosure (with the threshold set at $k = 3$) in this instance 40% of records. $k = 3$ is a relatively loose threshold (i.e. an indicator that the data are not deemed particularly sensitive or disclosive), the logic being that for every record there are at least two other records with the same values across the key variables. However, at this stage the data is still 'raw' data without any disclosure control applied. Therefore, this position of 40% should decrease as we continue our analysis. By adjusting the threshold to $k = 5$ the picture becomes worse, with 57% of records flagged at risk, and subsequently 80% of records are flagged at $k = 10$.

Different variations of key variables are inserted into the *k*-anonymity model and those that have the largest effect on the number of records flagged at risk are noted as possible candidates for disclosure control. In this case, the following disclosure control was deemed to reduce the disclosure risk to a level at which the work could progress to the stage of comparing data utility:

- Age was recoded into five year age bands, with those younger than ten bottom coded, e.g. 11–15.

- Main disability was recoded into a binary indicator of whether or a disability was recorded.
- National identity recoded into a binary indicator of ‘Scottish’ or ‘Other’.
- Student stage and ethnicity were suppressed for all records.

This disclosure control resulted in a significant reduction in the number of at risk records; only 4% were now flagged at risk for $k = 3$.

Table 3. Summary of NIAH output for all key variables and local authority, $k = 3$.

	Original Data (N = 8185)			NIAH Safe Output (N = 4922)			NIAH At Risk Output (N = 3263)		
Variable	–1 σ	Mean	+1 σ	–1 σ	Mean	+1 σ	–1 σ	Mean	+1 σ
Age	10.7	14.0	17.3	10.5	13.7	16.9	11.0	14.4	17.8

	Original Data (N = 8185)		NIAH Safe Output (N = 4922)		NIAH At Risk Output (N = 3263)	
Variable	Mode	% of N	Mode	% of N	Mode	% of N
Gender	Male	54	Male	53	Male	55
Ethnic Group	1. White	92	1. White	99	1. White	88
National Identity	1. Scottish	76	1. Scottish	91	1. Scottish	53
Main Disability	84. No Disability	80	84. No Disability	91	84. No Disability	63
Student Stage	S3	10	S3	11	SP	12

Data Utility

To provide a comparison dataset, the remaining 4% of records that do not satisfy $k = 3$ are suppressed and therefore only a safe partition was used to compare with the original data. The exploratory research question is to explore what factors affect looked after children achieving above or below the median percentage attendance at their educational institution. The variables age, gender, local authority and placement type are used as exploratory variables, and a binary indicator of whether a record has above or below the median attendance (c. 92%) was constructed. Table 4 provides the results from a series of logistical regression models for the two datasets.

As a backdrop to these comparisons, a linear regression using the raw data was carried out for the effects of age in years on percentage attendance, which resulted in a significant estimate of –0.021 with an R^2 of 0.1096. Therefore, for every unit increase in age a 0.02 unit decrease in the percentage attendance is expected. This model suggests that as ‘looked after children’ get older, their percentage attendance in education declines.

Table 4. Models 1-6: Influences on having lower than median attendance, safe and original data (*p*-value in parentheses).

Dataset	Model 1 Safe	Model 2 Orig.	Model 3 Safe	Model 4 Orig.	Model 5 Safe	Model 6 Orig.
Age 11-15	-0.0258 (0.4203)	-0.0405 (0.1887)	-0.0251 (0.4339)	-0.0402 (0.1926)	-0.0249 (0.4371)	-0.0400 (0.1947)
Age 16-20	-0.6100 (<.0001)	-0.5916 (<.0001)	-0.6091 (<.0001)	-0.5912 (<.0001)	-0.6080 (<.0001)	-0.5899 (<.0001)
Age 21-25		0.00155 (0.9953)		0.000363 (0.9989)		-0.00359 (0.9892)
Gender (F)			-0.00140 (0.5542)	-0.00728 (0.7518)	-0.0150 (0.5269)	-0.00815 (0.7236)
No. of Placements					0.0532 (0.0721)	0.0553 (0.0578)
Pseudo R^2	0.1018	0.0926	0.1019	0.0926	0.1024	0.09

Comparing Models 1-6, it is also shown that age (with the less than ten years old category excluded for reference) has a negative effect on the binary median attendance indicator, though it is only significant for the 16-20 age group. An obvious impact on the data utility shown in the models is that all records with an age of greater than 20 have been suppressed in the safe dataset. In Model 6, the significance of the number of placements is stronger for the original data, and there is a change in sign for the 21-25 age group (although this is not significant). Lastly, the pseudo R^2 values have remained static as each model carries a figure of approximately 0.10.

Conclusions

Having established our exploratory research question, it has been demonstrated that this type of sensitivity analysis is important when data are changed or manipulated to satisfy statistical disclosure control conditions. Given the possible variations in analysis, it is difficult to predict the global effects of different disclosure control methods without probing the dataset beyond the surface level. In the example, these analyses would lead an analyst to very similar conclusions, although there are possible variations in the relationship between age and attendance, and the potential for a significant result when considering the number of placements. It should be noted that the research question was designed to avoid using variables that had been heavily affected by our disclosure control choices. This was predominately due to a need for actual results for comparison.

The data utility that has been lost should be remembered: ethnicity and student stage have been suppressed, and national identity and disability have been heavily recoded. If a research question looking at the effects of ethnicity, disability or national identity had been chosen, it would have hit an analytical dead end rather quickly. This examination of disclosure risk and data utility reinforces the need for data controllers to be able to work through any number of variations of disclosure control, sensitive to the type of release they intend to make and to the needs of the research community. This work also provides further evidence for the need for a conceptual risk vs. utility balance in practical applications of disclosure control across data access systems.

Acknowledgements

This work was supported by the Economic and Social Research Council [Grant ES/H032150/1].

The author would also like to acknowledge the assistance of the Scottish Government's Health, and Education Analytical Services Division for providing access and support for this research.

References

Argonne National Laboratory. (2012). *New institute to tackle "data tsunami" challenge.* Press Release. Retrieved from http://www.anl.gov/articles/new-institute-tackle-data-tsunami-challenge

Domingo-Ferrer, J., & Torra, V. (2008). A critique of *k*-anonymity and some of its enhancements. In *ARES 08. Third International Conference on Availability, Reliability and Security, 2008* (pp. 990–993). doi:10.1109/ARES.2008.97

Duncan, G.T., Keller-McNulty, S.A., & Stokes, S.L. (2001). Disclosure risk vs. data utility: The R-U confidentiality map. *CHANCE*, *17*, 16–20.

Duncan, G.T., & Pearson, R.W. (1991). Enhancing access to microdata while protecting confidentiality: Prospects for the future. *Statistical Science*, *6*(3), 219–232. doi:10.1214/ss/1177011681

Elliot, M., & Dale, A. (1999). Scenarios of attack: The data intruder's perspective on statistical disclosure risk. *Netherlands Official Statistics*, *14*, 6–10.

Elliot, M.J, Skinner, C.J., & Dale, A. (1998). Special uniques, random uniques and sticky populations: Some counter-intuitive effects of geographical detail on disclosure risk. *Research in Official Statistics*, *1*(2), 53–67.

Fienberg, S.E., & Makov, U.E. (1998). Confidentiality, uniqueness and disclosure limitation for categorical data. *Journal of Official Statistics - Stockholm*, *14*, 385–398.

Li, T., & Li, N. (2009). On the trade-off between privacy and utility in data publishing. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 517–526). Retrieved from doi:10.1145/1557019.1557079

McCaa, R., Muralidhar, K., Sarathy, R., Comerford M., & Esteve, A. (2013). Analytical tests of controlled shuffling to protect statistical confidentiality and privacy of a ten per cent household sample of the 2011 census of Ireland for the IPUMS-International database. Paper presented at the Joint UNECE/Eurostat work session on statistical data confidentiality, Canada. Retrieved from http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_2_McCaa.pdf

Microsoft News Centre. (2013). *The big bang: How the big data explosion is changing the world.* Microsoft. Retrieved from http://www.microsoft.com/en-us/news/features/2013/feb13/02-11bigdata.aspx

NHS Scotland Information Services Division. (2012). *Statistical disclosure control protocol* (No. 2.2). Retrieved from http://www.isdscotland.org/About-ISD/confidentiality/

Ohm, P. (2009). *Broken promises of privacy: Responding to the surprising failure of anonymization*. Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=1450006

Purdam, K., & Elliot, M. (2007). A case study of the impact of statistical disclosure control on data quality in the individual UK Samples of Anonymised Records. *Environment and Planning A*, *39*(5), 1101–1118. doi:10.1068/a38335

Ranganathan, M., & Bhopal, R. (2006). Exclusion and inclusion of nonwhite ethnic minority groups in 72 North American and European cardiovascular cohort studies. *PLoS medicine*, *3*(3), e44. doi:10.1371/journal.pmed.0030044

Rastogi, V., Suciu, D., & Hong, S. (2007). The boundary between privacy and utility in data publishing. In *Proceedings of the 33rd International Conference on Very Large Databases* (pp. 531–542). Retrieved from http://dl.acm.org/citation.cfm?id=1325913

Reiter, J.P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics - Stockholm*, *18*(4), 531–544.

Samarati, P., & Sweeney, L. (1998). *Protecting privacy when disclosing information: K-anonymity and its enforcement through generalization and suppression*. Technical report, SRI International. Retrieved from http://epic.org/privacy/reidentification/Samarati_Sweeney_paper.pdf

Scottish Government. (2011). *Children looked after statistics 2009-10*. Retrieved from http://www.scotland.gov.uk/Publications/2011/02/18105352/0

Skinner, C.J., & Elliot, M.J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 855–867. doi:10.1111/1467-9868.00365

Statistics Netherlands (2009). *Glossary on Statistical Disclosure Control*, Version: May 2009. Retrieved from http://neon.vb.cbs.nl/casc/glossary.htm

Steel, P., & Sperling, J. (2001). The impact of multiple geographies and geographic detail on disclosure risk: Interactions between census tract and ZIP code tabulation geography. *Bureau of Census*.

Stell, A., Sinnott, R., Ajayi, O., & Jiang, J. (2009). Designing privacy for scalable electronic healthcare linkage. In *CSE'09. International Conference on Computational Science and Engineering, 2009* (Vol. 3, pp. 330–336).

The Economist. (2010). Technology: The data deluge. *The Economist*. Retrieved from http://www.economist.com/node/15579717

Witkowski, K.M. (2008). Disclosure risk of geography attributes: The role of spatial scale, identified geography, and measurement detail in public-use files. *ICPSR Working Papers Series, No. 2.* Retrieved from http://hdl.handle.net/2027.42/58626