Formalizing the Selection of Key Variables in Disclosure Risk Scenarios

Elliot, Mark
University of Manchester
Oxford Road
Manchester M13 9PL
E-mail: mark.elliot@manchester.ac.uk

Mackey, Elaine
University of Manchester
Oxford Road
Manchester M13 9PL
E-mail: elaine.mackey@manchester.ac.uk

Purdam, Kingsley
University of Manchester
Oxford Road
Manchester M13 9PL
E-mail: Kingsley.purdam@manchester.ac.uk

Introduction

It is now generally accepted that a precursor to carrying out statistical disclosure risk analysis is the grounded generation of appropriate key variable sets. Work by Paass (1990), Elliot and Dale (1999) laid the ground work for an approach based on attack scenarios. This approach has been further extended by Mackey (2009) who describes the notion of the data environment. Elliot et al (2010) show how this notion might be formalised and describes a pilot system - the key variable mapping system. This paper reports on this work, incorporating the notion of data accessibility and other parameters effecting the degree of effort required by a would be data intruder who wished to use external data sources to identify individual population units within anonymised data sets. This is the first attempt to metricise the likelihood of a disclosure attempt and the work provides an indicator of how we might move beyond the data-centric approach (using just the properties of the to-be-released data to estimate risk).

Elliot and Dale (1999) describe an 11-point system for analysing statistical disclosure scenarios. Taking a quasi-criminological view they express that first one must consider the means, motives and opportunities that a would-be *data intruder* might hav;. only by considering why a would-be intruder would attack an anonymised dataset can we construct some measure of the prior likelihood of them make such an attempt. This view arose from consideration of Marsh et al's (1991) intuitive formulation of disclosure risk for microdata:

$$p(\text{identification})=p(\text{identification}|\text{attempt}).p(\text{attempt})$$

This formulation underlies the majority of attempts to model disclosure risk for individual level microdata. However, the focus has been on the conditional part. In other words p(attempt) has never been explicitly incorporated in such models (precisely because it has such operational complexity) and so, erring on the side of caution, data stewardship organisations have tended to work on the basis that p(attempt)=1.

Underlying Marsh et al's formulation is an implicit assumption that the disclosure process will be one

of linking (or matching) known information to the anonymised target dataset. There are two consequent assumptions here: (i) that the known information includes unique formal identifiers and (ii) that the known information includes some information which is also found on the target database – this information is usually referred to as the *key variables.* It is assumed that the linkage of the known information to the target dataset would be done through these key variables.

Elliot and Dale"s (1999) scenario scheme outlined a general set of principles for a conducting scenario analyses the output from which would be a set of such key variables. These were based on a mixture of rational analysis and *ad hoc* data collection. Elliot et al (2010) argued for new empirically based approach to producing key variable lists, which they term *Data Environment Analysis* and which captures in far more breadth and depth the information available to potential data intruders.

Data Environment Analysis

Data Environment Analysis (DEA) is a unique approach developed at the University of Manchester with funding from the Office of National Statistics (ONS). The goal of DEA to investigate, catalogue, categorise, and document available data in identification databases (those which could be used to link to target anonymised datasets in order to inform disclosure scenarios for data release).

Prior to this work there has been no (other) formal mechanism, within SDC, which allows the identification and classification of what additional, external, information might be utilizable by a would-be data intruder. This has meant that a key element of the scenario structure the *means of an attempt* (which centrally revolves around what key variables is potentially available to a would-be intruder), is based largely on informed guesswork. In the absence of such a formal method, we have hit a barrier preventing further development of our understanding of how a disclosure might occur. Without such understanding well grounded scenarios are impossible and we need well grounded scenarios to produce reasonably accurate disclosure risk measures to avoid disclosure management decisions that are either too conservative or too liberal (which can lead to (i) potentially risky data are released; (ii) valuable low-risk data are not released; (iii) data releases are of limited utility because of the damage caused by data protection methods).

Another way to look at this is that there are two overarching themes captured in prior conditions of any scenario analysis: (i) *is it likely* - which is assessed (using Elliot and Dale classification scheme) by considering an plausible intruder „motivations" and „opportunities for attack" and (ii) *is it possible, and if so how* - which is centrally assessed considering what additional information an intruder would require to successful identify and/or disclose new information about respondents from a data release. It is this second element, the how a disclosure might occur question, that provides the rationale to DEA work.

We have identified that within a given DEA cycle there are two phases: (1) to investigate, catalogue, categorise and document what additional information may be available to an intruder and (2) to develop grounded disclosure risk scenarios.

Data Environment Analysis Methods

Data environment analysis uses several interrelated methods. For each method the principle is to capture metadata which can inform disclosure risk decision making. Elliot et al (2010) describe the various methods by which such data can be collected:

1. Form Field Analysis – used for capturing information held on *restricted access databases* and involves obtaining and cataloguing the data entry fields and data use specifications of paper and online forms from companies, organizations, political groups, clubs, charities, government departments etc.

2. Public domain data harvesting –an intensive search of all data in a given environment available in principle to any person within a given population (such information might include: directories, professional registers, vital life events data, self published data - such as blogs, land registry information, estate agents house sale information, electoral registers, commercial datasets and so on).

3. Web Searching – the researcher searches the web for online resources and entry points for services, for example online shopping. Often this requires entering data for a fictitious person into a web form up to the point where payment details are requested.

4. Consultation with commercial data suppliers - commercial data companies increasingly hold detailed individual level information. Such information is often imputed from consumer surveys and combined with census and administrative data.

5. Attack Resourcing Simulation – an independent researcher is given a set of search parameters to identify as much information in the public domain about one or more variables that are in a particular target data set, within a particular time-frame.

6. Security practices case studies - these involve establishing links with organizations holding individual level information and developing case studies of present and future plans for information gathering and data handling practices.

7. Social network studies. Social networks can be defined by patterns of self disclosure; Elliot (2010) and therefore it is possible to use social network analysis to establish how much and what information individuals routinely know about other individuals within given social groups (work colleagues, neighbours etc.).

All of these methods of data collection are important, and different methods inform different scenarios and different aspects of those scenarios. The initial focus of our work to date was form field analysis as it is this method that has enabled the structured development of the Key Variable Mapping System (KVMS) which is described later in the paper and in detail in Elliot et al (2010).

Form field analysis works on the assumption that if a form (paper or electronic) asks for a given piece of personal information then that information will be stored on a database of individual records (which could form an identification file in an attempt to attack an anonymised data file). The second assumption is that the data will be stored at the level of detail that it is collected. These are not strong assumptions and therefore it is plausible to infer that each form provides veridical metadata for the correspondent database held by the organisation that collecting the information.

So, in form field analysis then the forms themselves are the raw data and each form provides information about the content of a database and that information itself becomes a record within a *meta-database*. The rows in such a database are the forms and the columns are possible variables and their codings. So, each possible coding of a variable will be represented as a column in the metadatabase. The various possible codings of a variable are not independent and can be arranged in graph structures such as that shown in figure 1 which shows part of an example graph for the variable "employment status".

In the scheme used here, each coding has a simple classification code consisting of a set of letters followed by a number. The letters simply represent the construct captured by the variables in this case employment status is coded as ES. The numbers are in a pseudo decimal form. The digits before the decimal point refer to the number of categories in that coding. So a "2" here means a variable with two categories (such as "employed" and "not employed"). The first digit after the decimal point is a placeholder which indicates whether the node is actually captured data (0) or a harmonization coding (1). The final two digits are simply used to distinguish between coding for which the preceding information is identical. So ES3.101 and ES3.102 are two different harmonization codes with three categories.

The management of this meta-database is a complex process. To understand this consider that in response to each question on a new form there are three types of development that can happen:

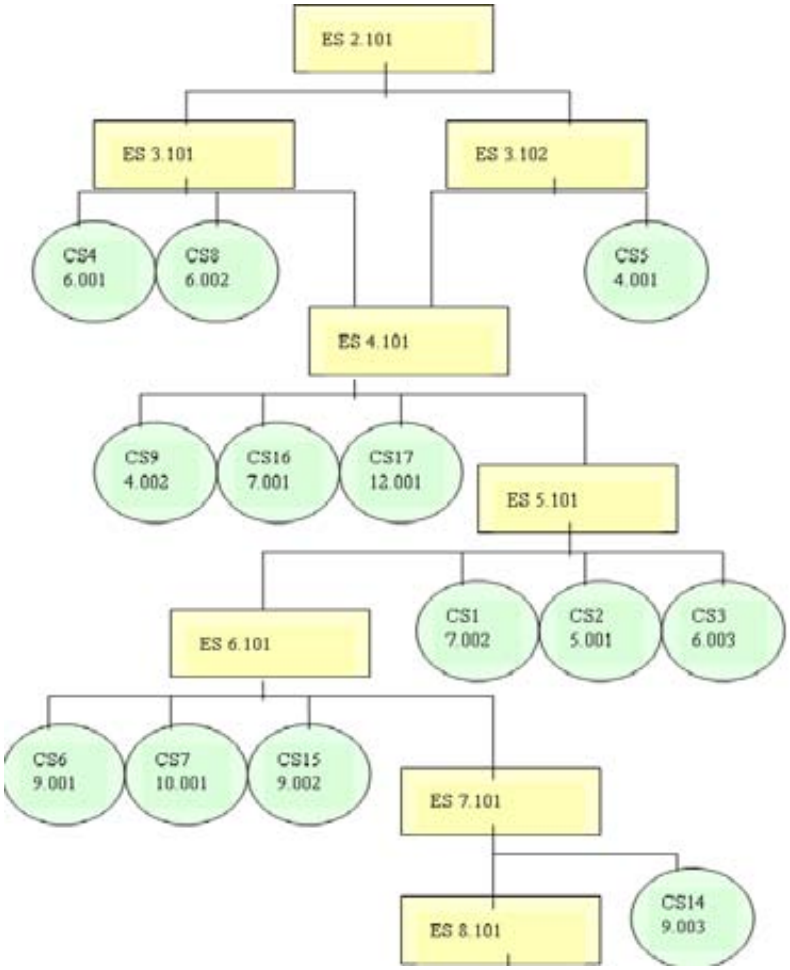
1. *Assimilation to an existing structure* – here the question maps directly onto an existing coding and so the process is one of simply adding the form to that section of the database and indicate the code.

2. *Accommodation of an existing structure* – here the question does not directly map on to an existing code meaning that at least one new coding will have to be created. The new code will then have to be placed into the graph for that key variable. This can be extremely complex and can require the creation of harmonization codes which link the new coding to one or more existing codings.

3. *Creation of a new structure.* Occasionally a question is asked on a form which might result in the decision to create a new key variable.

With each new form the metadatabase develops and the overall structure, which in effect is a picture of the data environment, is enriched.

Fig. 1. Example Taxonomic Graph for Employment Status Variable from Elliot et al 2009.



KEY:

ES (boxes) represents the harmonized codes created for Employment Status necessary to link together two or more captured codes. CS (circles) represents the codes captured directly in the forms collected and catalogued in the database for Employment Status.

The number before the decimal point represents the number of categories in that coding. The number after the point completes the unique identifier for that coding.

The Key Variable Mapping System (KVMS)

KVMS has been developed to enable the production of precise key variable specifications. Its overall function is to generate a key variable set which is the metadata intersection of two data sets (or indeed classes of datasets). KVMS compares codings of key variables across different datasets using a target dataset and either a single dataset or a summary of similar datasets. It is straightforward to add new target data sets of interest as they become available.

KVMS, written in Visual Basic, sits on top of the meta-database. The system works over a set of stored

variable codings; as outlined in section 2, these codings are continually developed as new meta-data is added to the system. The codings are structured as a set of taxonomic graphs such as the one shown in Figure 1. The nodes in each graph are in two types:

1. observed data nodes – these correspond to coding systems actually used in collected meta-data.
2. harmonized data nodes which are codings produced by harmonizing two observed data nodes.

Sometimes, an observed data node can also be a harmonized data node. For example, because most external data sets collect Date of Birth information, the variable Age invariably harmonizes to whatever age coding is used on the target dataset.

When, asked to map two datasets, KVMS starts at the observed data nodes corresponding to the coding systems employed in both target and data environment datasets and then proceeds up the graph until it finds the node where the two paths to the top of the graph join. The join is the harmonized coding between the two datasets. It repeats that process for all variables. The full set of harmonized codes is then the key variable set for the pair of datasets.

An alternative and potentially more useful analysis is to use all of the meta-data for the set of forms, at present the system allows for two sets of classes: the sector in which the data is collected (e. g. banking or supermarket) and the purpose for which the data is collected (e.g. finance application). Expanding this to allow for other forms of classification would be relatively simple. In order to use this set analysis, the user is required to enter a prevalence threshold value between 0 and 1, this indicates the proportion of datasets of a given classification that is required to have a given coding before it is considered. The system then moves up the graph from the coding system of the target dataset until it arrives at a classification which meets the threshold. A value of zero means that if any form has a variable coding then it is considered. In effect, this will mean that the system will select as the key variable, the most detailed harmonization code between any dataset within the set and the target dataset. On the other hand a value of zero means that a coding will only be considered if all forms within the set either have that coding or harmonize to it.

What is Public Data?

As described in the previous section, the initial development of the KVMS focused heavily on the capture and coding of restricted access data. The notion of restricted access implies a dichotomy: restricted and unrestricted. However, this is a simplification; at the very least this notion could be extended to a continuum of accessibility extending across both sides of the dichotomy. But really the situation is even more complex than this. In our recent work we have started to formalize this complexity.

Our focus here has been what is termed “publically available data”. By publically available – we simply mean that there is no legal or security restriction to a person obtaining a given set of data. This definition allows for several different types of data and so as a first step we should attempt to generate a typology of public data.

On a very general level we consider there to be two types of public data: formal and informal. Formal public data would include such sources as: public records (for example, the Electoral Register and share holder lists, professional occupation lists, telephone directories, and consumer profile data). Informal public data would include the growing wide range of sources of individual level information such as social networking websites and information obtained through personal and local knowledge.

Formal Public Information refers to information collected and released by government and non-government agencies and bodies. This is the type of data that in the UK the *Statistics and Registrations Services Act* indicates that a statistical agency such as the ONS must take account of when thinking about disclosure risk. It is collected largely for administrative purposes such as registers, documents and certificates. Thus the individuals to which the data pertains are compelled to provide the information either by law (e.g. birth, death, marriage certificates) or because they otherwise could not participate fully in society or benefit socially or commercially. For example, one needs to be on an Electoral Register to vote and in addition, this is also used as an identity check by credit agencies. Formal public information is

potentially available for large populations. The personal information collected is in general identifying and/or matching information such as name, address, date of birth, unique identifier, dependent information, house type, professional qualification etc.

Informal public information often arises from some sort of self-disclosure. Self-reporting such as, for example, in the form of an autobiography or information made available on-line poses a number of challenges in relation to the estimate of the risk of statistical disclosure. Not least because information provided by a person that pertains to the individual releasing it may be a rich source of identifying, matching and sensitive information. However, in contrast with formal information sources it is usually given on a voluntary basis and available for small numbers/ disparate members of populations. Examples may include online CVs, personal web-pages and information provided on social networking sites but may also include the deliberate publication of completed public records. On a pragmatic basis there is a good case for arguing that National Statistical Agencies should not have to take account of this (disparate) source of information not least because individuals provide the information of their own free will.

Another sort of informal information, which might also be regarded as public abut non-voluntary, is information that might be obtained though observation. My neighbours might through observation or inference know of the type of house I live in and whether I drive a car.

A grey area in this classifications concerns secondary reporting of (Formal and Informal) Information, which themselves become a type of public data. Specifically, this refers to information collected (and potentially released) by a third party and reported from publicly available sources such as court records and registers and /or self reported sources. Information maybe drawn together from one or other data type (i.e. official public source and/or a self-report source) and from many sources within the data type(s). Examples might include the reporting of a court case.

There is potential for secondary reporting to raise disclosure risk problems for National Statistical Agencies such as is in relation to what has been termed jigsaw identification where identification is made through the gathering of information through numerous sources. The law may protect indirectly against jigsaw identification such, as for example, by court injunctions and the right to remain anonymous. The different sources of information which could be used for identification are not disclosive in themselves.

Operationalising the Accessibility of Formal Public Data

We attempt to capture the concept of data accessibility in terms of the effort that would be required from a member of the public to access and obtain information from public data sources. In an attempt to simplify this we construct a simple meta-variable – resource costs - with three levels: 1. High to medium resource cost; 2. Medium to low resource cost; 3. Low resource costs. The notion of resource costs conceptualizes how easy it is to access/obtain data. In our first pass at this it incorporates the following: (a) the mode of access; (b) the scope of access to data: whether one can access a whole database or single cases; (c) the cost involved in accessing and obtaining the data.

(a) Mode of access: four categories were identified and then graded using a very simple numerical system. This system graded the categories 1- 4 according to the effort required to access data, with 1 representing the least amount of effort needed and 4 the most amount of effort.

Mode of access	Rank
Available online and from information centre	1
Available online only	2
Available from a local centre	3
Available from a centralised information centre	4

(b) Scope of access: this refers to the amount of data one potentially has access to. Three categories were identified ranging from access to the whole database to access to only single cases. The categories are

graded 1-3 according to the level of access, with 1 representing the most access (we hypothesise that the greater the scope of access the less effort will be required to find information about a particular individual) and 3 the least access.

Scope of access	Rank
Whole database is accessible	1
Case by case search only (returns multiple similar names)	2
Case by case search only (returns one name only)	3

(c) Cost: This refers to the economic costs associated with viewing/obtaining data. The categories are graded 1-3, with 1 representing minimal cost and 3 the highest cost.

Cost of Access	Rank
Minimal or no cost	1
Medium cost	2
High cost	3

If we return to the data accessibility dimension of „resource costs" then the ranks previously identified could be summed to produce a simple metric, illustrated in table 4.

Table 4. Simple Resource Cost Metric

Resource cost summary	Resource Index
High – medium resource cost	8+
Low - Medium resource cost	5-7
Low - resource cost	3-4

Although very rudimentary, this coding scheme is the first step to parameterising data accessibility as part of our disclosure risk scenarios. Using this, we can set up different scenario descriptions for intruders of varying levels of resource and this will lead to different sets of key variables. We can also consider whether the set of publically available data (give a certain set of resources) would enhance data on any given restricted dataset. This is important because if data is available, at a given resource cost RC, then it should added to the output key for any scenario for which available resources exceed RC.

Building on this work, the intention is to encode public data metadata for use in the KVMS in the same way as we can restricted access data. However, the inputs regarding its use in scenario creation will parameterized on the basis of resource costs rather than prevalence. We can then see building up a key variable picture as a multi stage process involving what public information could potentially be combined with what restricted access information.

Concluding Remarks

The work reported here forms part of a portfolio of ongoing research which attempts to capture, describe and metricise the data environment. This process is important in enabling data stewardship organizations such as national statistical institutes to make principled decisions about data dissemination of research data products. As the work goes forward we are planning to run studies on informal public data, particularly that known via personal knowledge. A second key point of interest for data stewardship organisations is how information is changing over time; the importance of understanding data environment

trends will be crucial to longer term planning of data dissemination.

REFERENCES.

Elliot, M. and Dale, A. (1999): “Scenarios of Attack: A Data Intruder‟s Perspective on Statistical Disclosure Risk," *Netherlands Official Statistics*, 14, 6-10.

Elliot, M. (2010) “Privacy, Confidentiality and Disclosure: conflicts and resolutions”, Paper presented to Angela Dale‟s retirement Colloquium; Manchester, June 2010. http://www.ccsr.ac.uk/events/adc/index.html

Elliot, M. J., Lomax, S., Mackey, E. and Purdam, P. (2010) „Data Environment Analysis and the Key Variable Mapping System‟ with In J Domingo-Ferrer and E Magkos (eds) Privacy in Statistical Databases. Springer; Berlin.

Mackey, E. (2009) “A Framework for Understanding Statistical Disclosure Processes: A Case Study Using the UK‟s Neighbourhood Statistics.” Phd Thesis submitted to the University of Manchester.

Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D., and Walford, N. (1991): "The case for samples of anonymized records from the 1991 census", *Journal of the Royal Statistical Society series A* 154 305-340.

Paass, G. (1988): "Disclosure risk and disclosure avoidance for microdata", *Journal of Business and Economic Statistics* 6(4): 487-500.

Purdam, K., & Elliot, M. J. (2002): "An evaluation of the availability of public data sources which could be used for identification purposes - A Europe wide perspective", CASC project report. Manchester: University of Manchester.

Purdam, K., Mackey, E. and Elliot, M. (2004) The Regulation of the Personal, Policy Studies, Vol 25, No 4, Dec 2004 pp 267-282