

Disclosure Risk and Grid Computing

Mark Elliot, Stephen Pickles, Kingsley Purdam, and Duncan Smith

University of Manchester

2005

Abstract

Grid computing raises new issues in respect of the confidentiality of individual data. Different data sets are likely to have been collected under different terms of use and they are also likely to contain variables that have different levels of sensitivity and disclosure risk. Multiple dataset access and the increased computation power in grid environments also increases the potential for the identification of unique records. This paper provides a review of the key confidentiality issues raised by grid computing and reports the results of consultations with key stakeholders and the findings of exemplar disclosure risk experiments. Establishing effective disclosure control measures in grid environments is vital to ensuring the participation of data depositors in sharing both their data and computational resources.

1. Introduction

In this paper we provide an overview and initial assessment of the additional risks of statistical disclosure posed by the grid. The research involved examining the current availability of individual data over computer grid environments, in order to produce a scenario analysis of the disclosure risk arising from such data. We compared existing confidentiality agreements across key anonymised data sets that might, in the future, be grid-enabled. We identify methods of disclosure risk assessment that take account of the additional disclosure risks posed by grids and grid available data.

The following research methodology was employed: [1] We collected and analysed confidentiality agreements across a number of case study surveys and pilot data sets that might be used as part of the grid data resources. Interviews were conducted with survey managers to examine the existing disclosure control methods, to consider the potential they see for grid computing and to identify any specific concerns they may have about disclosure control. [2] A review of publicly available data in the UK was conducted, using online search techniques and form field analysis. The forms collected reflected the types of form that the general public might be required to complete in the course of their dealings with these organisations. Making the plausible assumption that data collected through these forms was stored in a database, a set of variables that the organisations might have

reasonably generated through the forms was produced. [3] Finally, we investigated the disclosure risks stemming from table linkage and record linkage, particularly where (possibly publicly available) sample data are available to an intruder. This replicates a grid disclosure risk scenario.

1.1 Grid computing and the data environment

Grid technologies include computational grids for high performance computing, access grids (supporting collaboration amongst dispersed researchers) and data grids (supporting the movement of data and analyses). They open up new opportunities to enhance existing data sources and data quality to inform research, policy and service delivery. Possibilities at present include: linking data and data sets online thus creating virtual data sets; distributed data storage and data processing for large-scale data sets and complex analyses; data mining across different data sets; real time data updates and single sign in and use authorisation.

Through the e-Social Science pilot project programme (being funded by the ESRC) a range of qualitative and quantitative data is being grid enabled, including: individual level demographics (age, gender), health, education (school performance), criminal record information, risk assessments, service access, geographic information, field research information, financial transaction information, telecommunications data, service provision, financial market information, expert comment,

ethnobiology, video images and environmental modelling data. At the core of many of the projects is the facilitation of simultaneous access to multiple data sets. Another ESRC project, Convert Grid, integrates aggregated census area data and individual level commercial lifestyle data to develop an interoperable system. In addition to the ESRC driven efforts, many more grid related projects are being funded by the Joint Information Systems Committee (see www.jisc.ac.uk).

There is no limit to the type of data that can be grid enabled. Technology will soon allow the establishment of individual digital archives that include linked images, documents and audio. Grid computing, and particularly the disclosure control methods adopted, need to take account of these possibilities.

1.2 The disclosure risk issues

Individual level data is record level information about specific individuals (though it may also be linked to aggregate level information). It can include demographic information, consumption habits, health information, interests, communications, movement and biometric information. Biometric information includes information about appearance, social behaviour, what the person does (e.g. key stroke dynamics) and natural physiography (e.g. iris scans, voice prints, DNA patterns and body scans¹).

One of the central threats to the promise of grid computing is breaches of confidentiality arising from statistical disclosure. This has been defined as “the revealing of information about a population unit through the statistical matching of information already known to the revealing agent (or data intruder) with other anonymised information (or target dataset) to which the intruder has access either legitimately or otherwise” Elliot (2004). Statistical disclosure can occur in a number of (related) ways. The most serious form of disclosure is termed attribution, or prediction disclosure. In essence, attribution takes place whenever a so-called data intruder can logically (or with a very high degree of certainty) infer something that was previously unknown about a particular population unit (usually a person, but could be an organisation). Possibilities for attribution can be assessed via table linkage; and it can be shown that such inferences are only possible when tables contain zero counts (Smith and Elliot, 2003). Attribution can also occur by

identifying an individual with a microdata record when the microdata record contains previously unknown information. The risk is greater when the record can be ‘expanded’ via linkage to records in other datasets.

Protecting confidentiality in statistical databases has been extensively considered in relation to survey data and particularly census data, for example; Flaherty (1979); Hakim (1979); Elliot (2001); Feinberg and Makov (1998); Marsh et. al. (1991); OPCS (1992); Singer (2001). Different disclosure control techniques are employed in relation to different data sets, data releases and across different countries. Techniques are based around reducing data specificity and distorting the data, and include reducing sampling fractions, perturbing, rounding, record swapping and adding noise (see Marsh et. al. 1991; Doyle et. al. 2001). Such techniques impact on the usability and quality of the data. Yet the research in this area is limited. For example the Individual Samples of Anonymised Records (SARs) from the 2001 Census contains less information than was the case in the 1991 because of concerns about disclosure risk. Such concerns have also resulted in considerable delays in the release of the data, and at the time of writing the Household SAR has still not been released. A number of the variables on the Individual SAR have either been removed or reduced in detail and recoded. Such measures can limit the usefulness of the data and have caused concern amongst users. For example, as Wathan (2002) has argued, suppressing information on households with 7+ members and multi-ethnic households would not only lead to a bias in the data regarding ethnicity, overcrowding and number of children, but would also have important policy implications. The lowest level of geography to be released within the SAR is local authority level, and regional level is only available for the licensed use SAR. Again this severely restricts the extent to which the data can be effectively used. Moreover, Purdam and Elliot (forthcoming) replicated a set of published analyses on 1991 SAR data before and after disclosure control using the ARGUS system had been applied. They found that disclosure control measures had a significant impact, not only on the usability of the data but also on the conclusions stemming from the analysis.

Statistical organisations such as the Office of National Statistics are concerned that claims of disclosure by data intruders could lead to a loss of public confidence, and a reluctance to participate in, for example, census surveys. It

¹ In the USA a recent government employment survey of workers at an air force base, a full body scan was included on the database alongside other variables such as age and gender.

can be argued that there is always some degree of uncertainty over the correctness of any available data. But as this uncertainty would not generally prevent a claim of attribution, it is generally ignored for disclosure control purposes. Although attribution concerns inferences that can be made with certainty (subject to assumptions regarding data quality) it can be the case that inferences that can be made with high degrees of confidence are also a concern. This is also the case with record linkage. Exact inference can be computationally demanding, and approximate inference via simulation may sometimes be the only feasible option (for both intruder and risk assessor).

Specific disclosure risks are raised in relation to the grid enablement of medical and patient record data.

The ability to link patient records to create comprehensive medical and research datasets is at the heart of the introduction of e-Science in healthcare. The Clinical E-Science Framework (CLEF) research programme places patients and their histories at the centre of clinical practice and research. (www.clinical-escience.org). CLEF specifically involves the integration of distributed heterogeneous databases of descriptive text, image, genomic, and quantitative information.

The UK Biobank established in 2003 aims to build a major resource to support a diverse range of research that will improve the prevention, diagnosis, and treatment of illness and promote health throughout society. It is focused on the linking of individual DNA records with a range of socio-demographic and lifestyle information (www.ukbiobank.ac.uk). It is proposed that the BioBank will hold 500,000 individual records across a specific age group. The project will follow the health of the volunteers for up to 30 years, the creation of such a database raises a number of significant statistical disclosure issues; see Biobank (2003).

The degree of detail and accuracy required by medical practitioners and researchers presents new disclosure control challenges. These include the different risks associated with highly detailed patient notes, treatment plans and outcomes, named practitioners alongside quantitative medical research data. Other factors include the need to continually update records from multiple access points. In addition, medical data is arguably potentially more sensitive than social survey data, and its secondary use raises a number of ethical and data protection issues. As a result, any disclosure would constitute a crisis for the profession and potentially for the patient.

Recent examples have highlighted some of the risks of disclosure. For example, an image of a skin complaint, which had been used without the patient's permission in the British Medical Journal was recognised by the patient. Secondly, using sources such as lists of medical professionals who conduct abortions along with local knowledge, it became possible to identify a person who had undergone a late abortion.

Only limited research has been conducted on the issues of confidentiality and disclosure raised by the grid enablement of data. Cole et. al. (2003) in their scoping study of the grid enablement of data sets highlight a number of general issues in relation to confidentiality and the grid, but it is clear many of the issues remain unresolved. Grid technologies carry the risk of both accidental disclosure and disclosure as a result of deliberate attack.

2. Findings

2.1 Form field analysis

We reviewed the availability of individual level data and possibly disclosive linking data already available in the public sphere, largely from commercial data companies but also from public registers. Over a two-month period in 2003 we analysed one hundred paper forms used by organisations and companies from the financial, retail, insurance, marketing, education, health and local government sectors. The forms analysed reflect the types of form that any individual in the UK might be required to complete in the course of their dealings with these organisations. From the sample of one hundred forms a total of 590 variables could have been generated, many of these duplicate the type of demographic information that is found on anonymised datasets released by National Statistical Institutes (NSIs). Ninety-one of the organisations were registered with the Information Commission. The table below summarises the percentage of forms requesting key demographic and socio-economic information.

Some of the forms requested a range of outwardly unrelated information to the stated purpose of the form. One such example is that of a survey, ostensibly about pets, undertaken on behalf of numerous organisations including AOL, Procter & Gamble, CAFOD and Lloyds Bank Insurance Services. Along with the usual identifying information details on the respondents' dependants (i.e. number of dependants, their DOB and gender), their income, banking and insurance arrangements

were also requested. Information on financial liabilities included the number and type of credit cards the person had, the type and amount of repayment they were making, their overdraft facilities and personal loan arrangements, their investments and mortgage arrangements (i.e. the property they owned and how they paid for it), house (whether they had a security system and, the value of their utility bills) and car details (the household number, their car type and how it was purchased). Information was also collected on respondents' lifestyle (smoking and drinking habits and dress size), sport and leisure interests (types of music listened to and books/newspapers read, betting habits and accounts held), shopping and travels habits as well as supported causes (including questions on who they donate to and how, and the political party to which they belong).

Table 1.

Variable	% of forms
Age	69%
Sex	24%
Marital Status	28%
Nationality	14%
Country of Birth	7%
Length of time at current address	15%
Previous address	17%
Religion	6%
Education	8%
Employment Status	44%
Occupation	32%
Health	14%
Number of Children	28%
Tenure	19%
House type	4%
Relationship	30%
Number of rooms	4%
Number of cars	6%
Income	18%

All the organisations that specified a privacy agreement on their forms (49% of the sample) stated that they shared the information they collected with others within their organisational network. The organisations in this “network” were not always made clear. Of these organisations, 14% (all from the private sector) stated that they also shared the information with others outside of their organisational network, with 6% transferring information out of the EU. As outlined below, the completion of such forms often leads to data being transferred to third parties; often a data warehouse specialist. Consumer information can be held indefinitely, and thus lifetime profiles of consumer purchases and behaviour are being built up in data warehouses (authors' interview with data

warehouse manager, 2002). A much larger scale study of the information gathering processes is required. This would allow analysis across different sectors and would look at data quality and use. However, it is already clear that the scale and detail of information being collected and stored is considerable, and is also likely to continue to increase.

This analysis has only provided a snapshot of the scale and type of information gathering. The authors are developing this methodology further and have secured funding from the ONS to conduct a one year pilot study of such information gathering. Initial findings from this study have revealed that types of data collected continue to increase, and there is only limited awareness amongst expert data users of the extent of individual data commonly available both in the public domain and in restricted access datasets.

2.2 Statistical disclosure risk assessment experiments

Two specific approaches to risk assessment were investigated. The first concerning table linkage, and the second concerning record linkage. The primary function of these experiments was to establish whether adding additional overlapping datasets to a data environment would increase the disclosure risk associated with a base set of data.

Experiment 1 considered how the release of sample data (in addition to a release of rounded tables) reduces uncertainty over the cell counts in the population cross-classification over all variables. As our ‘base’ table we took the 3-way table AGE × ETHNICITY × GENDER from the 1991 sample of anonymised records (SAR). Specifically the release of all three 2-way tables of the 3-way table was considered, with an additional table comprising a sample from the full 3-way table. The cell counts in the 2-way tables were rounded to the nearest multiple of three. Various population tables (corresponding to different geographical areas) were used, along with a variety of sampling fractions. Intruder uncertainty was measured using an entropy-based measure. This measure differed from previously investigated measures in a number of respects. The entropy-based measure was based specifically on the upper and lower bounds that an intruder could place on the base table cell counts. Generally, tighter bounds represent a greater risk of disclosure. The entropy-based measure is necessarily a non-decreasing function of the sample size. However, we found that the measure only decreased slowly over sampling fractions from

0% to 80%, representing only a moderate tightening of the cell bounds. Even a sample fraction of 90% left a reasonable degree of uncertainty over the true cell counts. These results were consistent over all the populations considered.

The results demonstrate that samples with quite high sampling fractions can be released, whilst maintaining a reasonable degree of intruder uncertainty over population cell counts. Of course this uncertainty is only measured in terms of cell bounds, so is mainly applicable in situations where exact disclosure is the issue. That is, where we are concerned about discoveries that an intruder can make with certainty. The following experiment deals with a situation where inferences that can be made with high probability are the issue.

Experiment 2 considered the degree to which an intruder's attempt to link records across distinct samples of data from a common population could be enhanced by naïve Bayesian methods. The samples had a number of common variables that the intruder could use for matching purposes. The specific scenario of attack considered was when the intruder could find a record in the first sample that had only one possible match in the second sample. We found that even a very naïve approach, which avoided the usual computational complexity of exact Bayesian inference, could be used to generate reasonably accurate match probabilities. The degree to which this was possible was assessed by ranking the unique matches by their estimated probabilities and by their known (to us) correct match probabilities, and testing the significance of the rank correlation. The results were highly significant.

The general conclusion was that naïve Bayesian linkage methods can be effective from an intruder's point of view, although uncertainty over the population counts provided protection against linkages being able to be made with certainty. This also suggests that naïve Bayesian methods might be useful for risk assessment. A common problem in risk assessment is discovering exactly what an intruder can infer given enough computational resources. It is very useful to be able to estimate this efficiently, and with reasonable accuracy.

Correctly linking records between samples from a common population was significantly enhanced by the co-presence of other tables. The tendency was for match probabilities to be improved as new information was added to the data environment. But tables that provided new constraints on the counts in the population cross-classification of the common variables

turned out to be significantly more informative than others. Recovering these counts exactly allows exact match probabilities to be calculated. In particular, counts of one imply match probabilities of one.

2.3 Case Studies - Grid awareness and approaches to confidentiality and disclosure control.

A review of survey confidentiality policies was carried out and interviews were conducted with key survey managers and ESRC e-science pilot projects. The focus was on existing disclosure control methods, the potential use of grid computing and the identification of any specific concerns about disclosure control.

At the time of the review there was only limited awareness of the development of grid technology amongst key survey managers. The potential of grid computing for social science was not yet recognised by major data suppliers such as the Data Archive, the ONS or key stakeholders in the wider social science community². In our consultations with survey managers our interviews on several occasions had to be prefaced by an explanation of what grid computing is. There was a perception that desktop computers and the internet provided enough computational power and data access. There seemed to be a view that grid technology was more relevant to pure science applications such as physics. One interviewee commented that there was a lot of rhetoric about the grid in relation to social science. An experienced data user in the commercial sector has questioned who will use the grid and how non-specialised it will be (Dugmore, 2004).

As a consequence of this level of awareness of grid computing, the issue of disclosure control and grid computing was obscure to many interviewees. As one commented, "*we will be better placed to cope with the challenges and opportunities of the grid when they have consolidated the various confidentiality agreements they employ for different data sets*". The organisation is at present moving towards minimum standards and content in data access agreements.

Although the ESRC e-Science pilot projects are still developing the enablement of different types of data, some common themes can be identified. All of the projects are based around the use and linking of multiple datasets, and drawing on the computational power of grid

² The ESRC scoping report on grid enabled datasets drew a similar conclusion. See Cole et. al. (2003: 7)

computing. All the pilot projects had a clear vision of the potential of the grid, particularly in relation to the ability to build new virtual integrated data sets, speed up analyses (e.g. interactions and hold state) and ultimately develop new areas of research.

One of the key challenges of data linking is that information in different data sets and different types of data, such as text and image, are likely to have different levels of sensitivity, disclosure risk and specific terms of collection and use. The ESRC pilot projects described different security and confidentiality issues. These related to commercial interests and intellectual property rights, data sharing agreements and computational access. On the whole, disclosure issues were seen as important, but often described in terms of being a barrier in gaining access to data and developing data sharing agreements. It was clear that, in many ways, a grid was thought to offer a solution to disclosure issues by creating a virtual safe environment of users. Yet one project stated that they were developing micro-simulation data because it which was free from usage restrictions.

There was only limited awareness of statistical disclosure risks and how grid technology might increase such risks. There was also limited consideration of the disclosure risks of data released in anonymised form and other outputs from grid analysis such as statistical models. One of the projects which is focused on developing architecture for the grid stated that their software will not allow the linking of records. For other projects linkage is a key goal.

Bespoke formal agreements still had to be put in place with the organisations sharing data. This was proving complex for some projects, and more so if agreements were across national boundaries where the legal frameworks differed. For example, in one project prior agreement had to be established as to under which legal jurisdiction any disputes would be resolved. Another project described how prior knowledge of the individuals involved and established relationships of trust was the only way in which the data sharing got off the ground. After this the arrangements had to be backed up by specific legal agreements based on hypothetical scenarios. These had been very time consuming to draw up. Because of the difficulty in getting agreement on even hypothetical outcomes one project has kept medical companies out of their grid pilot projects. Another project commented that setting up data sharing agreements from scratch, without any prior relationship, would have been almost impossible. One of the

projects commented that it was for this reason that it was important to use real data rather than synthetic data. Much of the work in developing a grid related to working with different organisations and building the appropriate agreements.

Confidentiality and data sharing agreements go beyond considering the risk of identifying individuals to considering an organisation discovering new value in a data set and the question of who had ownership of this. In fact, across the pilot projects this seemed to be a more definitely developed area of regulation. Organisations and researchers are often reluctant to share their data because of this. Data sharing agreements may be one of the key barriers to the development of grids. One project commented that access had only been granted to a particular dataset on the basis that the project would add considerable value to the data. Another project commented that they felt that security issues in relation to the grid are still emerging. Another project commented how they had developed their own agreements with organisations, but that there should be agreements specifically established for grid enabled data. They commented, "*it is not sufficient to rely on agreements based on trust, as individuals in organisations change, as does the climate and legislation for data protection*". These need to be developed on a global level. As Foster et. al. (2001) state the grid is not simply about unrestricted access to resources, but about controlled sharing.

In relation to the development of grid technology one interviewee commented that there will need to be extensive training and user support to ensure that data linking is done correctly. It is notable that JISC is funding a fast track e-Social Science training and awareness programme.

Many grid development projects which address concerns about the infrastructure for data sharing are ongoing. For example, the Data Documentation Initiative is focused on developing a standard for technical documentation and formatting for social science data. All data will have XML tags and therefore are readable by computers. Codebook information is encoded into databases that share a known structure and a specification language across many bodies of data. The Dublin Core Metadata Initiative (in the UK see the E-Government Metadata Standard³) is an open

³ These standards are specifically designed for public sector data. See <http://www.govtalk.gov.uk/schemasstandards/metadata.asp>

forum engaged in the development of interoperable online metadata standards that support a broad range of purposes and business models. The development of a metalanguage which can be used in the linking of datasets is ongoing. Yet as Cole et. al. (2003) state “*while these standards are relatively powerful and extensible they only scratch the surface of what is need for a comprehensive description of social datasets*”. See Watson (2003) for further general discussion. The Open Grid Services Architecture Data Access and Integration project is concerned with constructing middleware to assist with access and integration of data from separate data sources via the grid. A range of access protocols, user certificates and authentication initiatives are being developed by JISC as part of the wider development of grid architecture; particularly what is termed the three As: authentication, authorisation and accounting. However, despite covering some important issues none of these initiatives directly address the issue of statistical disclosure and informational privacy.

As well as there being a pressing need to examine the technical disclosure risk issues arising from grid computing, the legal implications of the grid enablement of data are also require urgent consideration. The Data Protection Act legislates that data cannot be used for any purposes, other than that for which it was collected, without the subject’s knowledge. Data use has to be compliant with the Data Protection Act. Yet what constitutes personal data remains unclear, and regulation varies across different jurisdictions. The question of what constitutes personal information is currently proving a challenge to data protection enforcement in the UK and across the EU in relation to compliance (authors’ discussion with Information Commission, 2004). The ruling in the recent Durant case whereby the Court of Appeal decided that to qualify as “personal data” under the Data Protection Act the information had to be biographical in a significant sense to a person. It could not be just the person’s name. Durant vs. FSA 2003 (EWCA Civ 1746) (8th Dec 2003). This provides a further complexity to the legal definition and regulation of personal data.

It has been argued that the protection of privacy has inhibited service provision and research. For example, the national charity Cancer Research UK has claimed that the requirement to protect confidentiality through anonymisation under the Data Protection Act is having a detrimental effect on research (see The

Observer 5.10.3). As a result, Cancer Research UK is campaigning for an exemption from the Data Protection Act for medical research. Clearly, legal principles of data use in grid environments need to be drawn up.

3. Conclusions and policy implications

Grid infrastructures must, like other data organisations, take account of the increased availability of individual level data and the possibly disclosive linking with data in the public domain or held on restricted access databases.

Statistical disclosure risks are posed by linking data, knowledge discovery techniques and outputs that are released outside a grid environment. Our research found that across a number of ESRC grid projects there appears to be only a limited awareness of statistical disclosure issues. There seemed to be an assumption that a grid environment constitutes a safe data setting. Also, the disclosure risk impact of information that may become available later in grid development has implications for our decision making now. The decision over whether to include a given variable or not should be based critically on the sensitivity of the information represented by that variable. Sensitivity analyses should be ongoing, in order to assess public opinion about particular information. In a grid context there is scope for real time assessment and reassessment of the context for data release.

There several strong policy recommendations that we can make based on this and other ongoing research:

1. There has been only limited consideration of the legal framework and good practice protocols for the grid enablement of data and data sharing. Data sharing and agreements are often ad hoc and based around individual relationships of trust. Compliance with data access protocols and agreements needs to be backed up by systematic physical inspection of sites and data handling practices.
2. Data holders should review their confidentiality policies to take account of the new possibilities offered by grid computing. Development of good practice and legal agreements for data sharing in a grid environment is essential. This needs to take account of ongoing developments in the legal recognition of privacy.
3. Ongoing monitoring of the data available in the public domain is also needed. Such

monitoring would enable the construction of plausible attack scenarios so that disclosure risk can be measured in a meaningful way.

4. Data providers should be consulted regarding their views on individual data use and access, in order to maximise data availability.
5. Development work should be conducted with data collectors concerning the incorporation of grid data use cases and disclosure control into the survey design process and terms of data collection.
6. Disclosure control should not just be something that happens once the data have been collected.
7. Methods for analysing information loss resulting from the disclosure control of grid enabled data sets must also be developed. The development of software tools for maximising data release is of paramount importance.
8. Computational statistical methods and statistical disclosure control software must be developed for controlling the querying of anonymised data in environments containing multiple overlapping datasets.
9. Further research is required into the scope for actively assessing risk live across a grid.

These measures would enable the development of a standard set of guidelines, components, and services to ease the creation of new repositories and the bringing on line of new grid enabled datasets.

References

- Better Information (2000) Cabinet Office, Policy Action Team.
- Biobank (2003) Ethics and Governance Framework, Biobank, Manchester
- Birchard, M. (2004) The Birchard Enquiry. Cabinet Office.
- Cole, K., Schurer, K., Beedham, H. and Hewitt, T. eds. (2003) Grid enabling quantitative social science datasets - a scoping study, Economic and Social Research Council.
- Common Data, Common Sense (2000) Audit Scotland.
- Corti, L. and Wright, M. (2002) MRC Population and Data Archiving and Access Project, (Draft) UK Data Archive.
- Data Archive (2004) Depositing Data - Legal Issues, Data Archive 2004
- Data Sharing Advisory Group (2000) Gee, J, Director, Counter Fraud Services, Department of Health, Privacy and Data Sharing Meeting Minutes.
- Data Standards (1999) Government Data Standards Working Group, Data standards catalogue V1.0, September 1999. <http://www.iagchampions.gov.uk/guidelines>
- Doyle, P., Lane, J., Theeuwes, J.J.M and Zayatz, L. (eds) (2001) Confidentiality, Disclosure and Data Access, New York: Elsevier.
- Dugmore, K. (2004) Linking Data, ESRC Research Methods Seminar, Concluding comments.
- Elliot, M. J. (2001) 'Data Intrusion Simulation: Advances and a Vision for the Future of Disclosure Control.' Statistical Journal of the United Nations 17. 1-9.
- Feinberg, S. and Makov, U.E. (1998) "Confidentiality, Uniqueness and Disclosure Limitation for Categorical Data", Journal of Official Statistics 14.
- Flaherty, D. (1979) Privacy and Government Data Banks: an international perspective, London, Mansell Scientific.
- Foster, I. Kesselman, C. and Tuecke, S. (2001) The Anatomy of the Grid, International Journal of Supercomputer Applications 2001.
- HAKIM, C. (1979) Census confidentiality in Britain. In M.Bulmer, ed. Censuses, surveys and privacy. London: Macmillan.
- Marsh, C., Skinner, C. and Arber, S. (1991) The case for the Samples of Anonymised Records from the 1991 Census, Journal of Royal Statistical Society A, 154 305-340.
- Modernising Government (1999) Cabinet Office. (Cmnd 4310).
- OPCS (1992) Statement of Policies on Confidentiality and Security of Personal Data, Titchfield: OPCS
- Privacy and Data Sharing (2002) The Way Forward for Public Services, PIU, London, Cabinet Office.
- Purdam and Elliot (forthcoming) A Case Study of the Impact of Statistical Disclosure Control on Data Quality in the UK Samples of Anonymised Records.
- Singer, E. (2001) Public Perceptions of Confidentiality and Attitudes Toward Data Sharing by Federal Agencies in P. Doyle, J. Lane, J.J.M. Theeuwes and L. Zayatz (eds) Confidentiality, Disclosure and Data Access, New York: Elsevier.
- Smith, D. and Elliot, M. (2003). An Investigation of the Disclosure Risk Associated with the Proposed Neighbourhood Statistics. ONS Report, 2003.
- Wathan, J. (2002) Disclosure Control and Household Size. SARs online discussion forum <http://les1.man.ac.uk/forum/ccsr/Forum5/HTML/000012.html>
- Watson, P (2003) Databases and the Grid, UK E-Science Core Programme