

Confidential Data Access using Grid Computing: An Outline of the Issues and Possible Solutions.

Mark Elliot, Kingsley Purdam, Duncan Smith

CCSR, University of Manchester

Mark.Elliot@manchester.ac.uk:

Abstract. This paper discusses some of the issues arising from the use of grid computing and in particular the impact of grid computing on confidentiality and disclosure. The list of research questions needing attention is quite long. Notably however, some concern how the grid technology might be used to broaden data access whilst at the same time strengthening confidentiality. Two methodologies: automatically monitored remote access systems and data environment analysis are described.

Introduction

Individual level data is increasingly in demand by researchers, policy makers and those involved in service delivery. In social science, researchers, policy makers and practitioners are increasingly seeking to access more detailed datasets and to link information held in different datasets. Grid computing technology opens up new opportunities to enhance existing data sources and to improve data quality. Possibilities include: the cross analysis of multiple datasets, linking qualitative and quantitative data online thus creating virtual data sets; virtual spaces for collaborative analysis and manipulation of data, massive distributed data storage and data processing (so called virtual-ultra large supercomputers) for large-scale data sets and complex analyses; data mining across multiple data sets; real time data updates and single sign-on and authorisation access control.

A wide range of quantitative and qualitative data is being grid enabled. Technology will soon allow the establishment of individual digital archives, which include linked images, documents and audio. Everyone will be able to have a digital memory or “memories for life” detailing their whole lives. The computational resources of grid computing will facilitate a growth in the availability and use of individual level data. Grid computing and particularly the disclosure control methods adopted need to take account of these possibilities.

Statistical Disclosure

One of the central threats to the promise of grid computing is breaches of confidentiality arising from statistical disclosure, which 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). 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). A range of techniques for disclosure control 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 decreasing sampling sizes, perturbing, rounding, 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.

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 have important policy implications. The lowest level of geography to be released with the SAR is local authority and only down to region for the licensed use SAR. Again this severely restricts the extent to which the data can be effectively used. Moreover, Purdam and Elliot (2005) conducted a range of test analyses on SAR data before and after a disclosure control measures 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.

Only limited research has been conducted into the potential risks of statistical disclosure and grid computing. Our initial research examining linkage between records in samples from a common population suggests that the ability to distinguish true from false linkages can be significantly enhanced by the co-presence of other data and the greater the number of datasets the greater, on average, the risk of disclosure from linkage (Smith and Elliot, 2005).

Thus grid technology carries an increased risk of both accidental disclosure and disclosure as a result of deliberate attack. The key issues relate to the seamless access to multiple datasets, multiple users and increased availability of computational power. The latter poses risks through linking data and knowledge discovery techniques. Another key concern relates to the different information collected under different terms of use and the sensitivity of different variables. However, grid computing does offer several opportunities for improving disclosure control methodology: (i) the potential to assess the status of the data environment/market prior to the release of a new data set, (ii) the potential to add a more computationally intensive ongoing risk assessment post release and (iii) the ability to track user access.

Key Research Questions

Accessing grid-enabled data and the vision of the data-grid, provide many new challenges to confidentiality research. It is hard to be definitive at this stage, however, eight key research questions are:

What additional disclosure risk does the data grid present over and above normal release methods?

What different challenges do grid enabled qualitative and quantitative data (including longitudinal data) pose.

What opportunities would the grid offer for protecting confidentiality (e.g. data intrusion detection sentries)?

In what ways would existing disclosure control methodology have to be changed to take account of the impact of grid computing?

How can grid environments be used to conduct disclosure control. (Note: "grid Environments" and "grids" are not necessarily synonymous.)

How can grid environments be used to carry out data quality impact assessments?

How will data grid environments affect the way we perceive personal data, privacy and confidentiality?

How is the present law (Data Protection Act, Freedom of Information etc.) affected by grid environments?

To appropriately answer these questions requires a multidisciplinary approach encompassing socio-legal research, statistics and computer science. Such a team has been assembled at Manchester University on the back of previous collaborations and with a wealth of experience in the field. The overarching aim of this team is to develop a framework for confidential data access in a grid context. Underlying this aim there are four key objectives:

1. To develop disclosure risk assessment software for grid enabled qualitative and quantitative data.
2. To develop methods for detecting intrusive behaviour in a grid context.
3. To develop methods for assessing disclosure risk associated with analytical outputs from data interrogated behind a firewall.
4. To identify good practice in confidentiality measures for the preparation of data for the grid and the development of legal protocols for grid computing.

A potential Model for Data Access

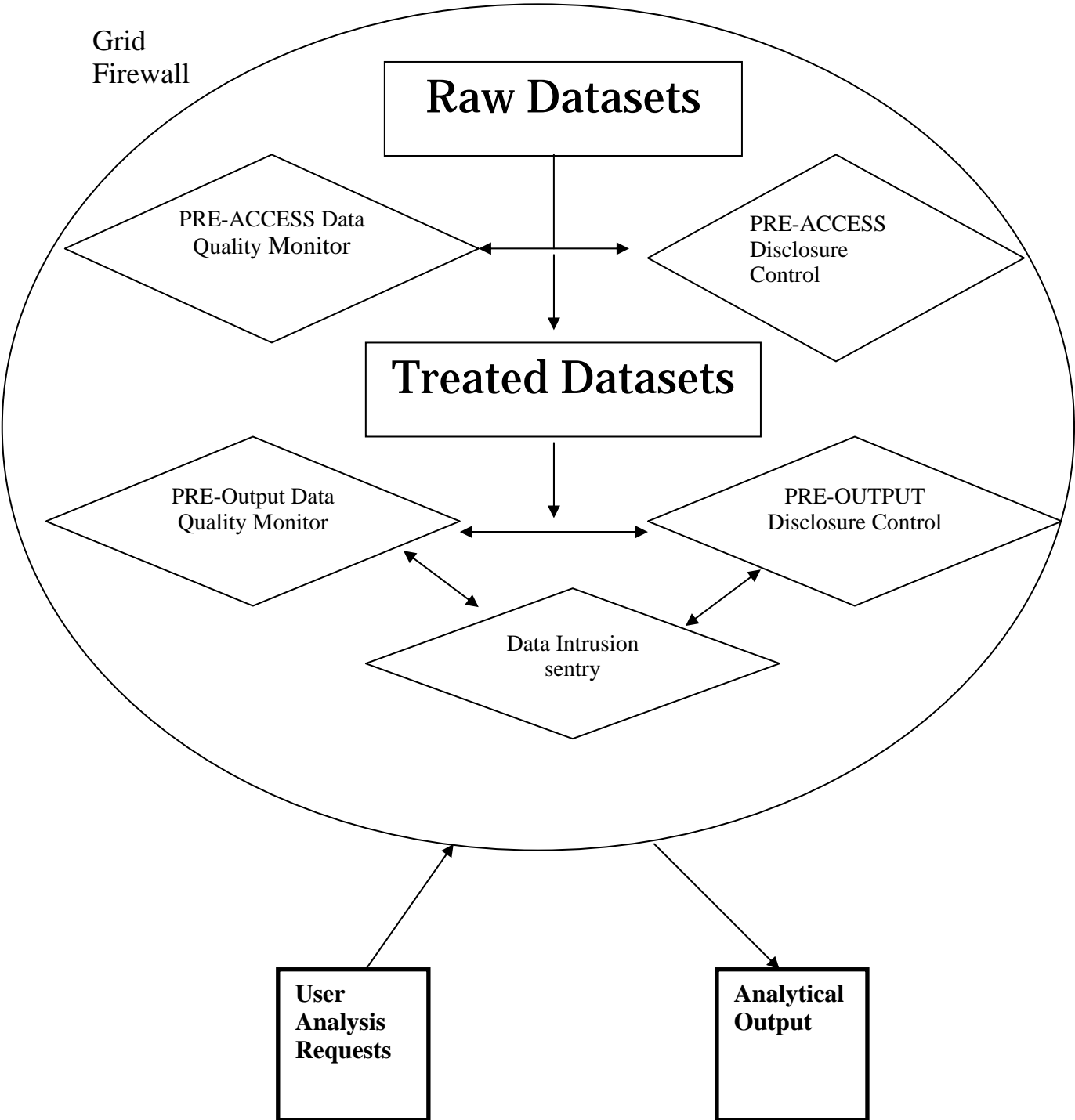
Our primary aim is to realise a situation whereby data access can be enhanced by grids rather than threatened by them. One potential model is show in Figure 1. We are not attached to the model that this figure presents, but it is at present a strong candidate. The model envisages data sitting behind a virtual firewall. Access to the data is via analytical requests. These requests are monitored by a data sentry system that serves multiple purposes, but the key under consideration at the present time, is the identification of possibly intrusive requests.

The detail of the raw data in this proposed system is potentially much greater than data users are accustomed to having available; the data has some generic disclosure control applied to it – to remove extreme instances of high disclosure risk information. However, the core of the system is in the automated analysis of the outputs. Running in parallel to this there are systems that monitor the data and output quality to ensure analytical validity for the user's request

The key components of this system are:

Pre-output Data Quality Monitor: As analytical requests are submitted to the enclosed system the output that the user will be sent can be checked for inconsistency by comparing the output that is produced against that which would be produced using the raw data. If the

Figure 1. Confidential Data Access using grid computing



results are sufficiently similar they are released, if not they are suppressed or are released with a health warning. There is a whole research stream to be initiated here to determine what is meant by “sufficiently similar” and to automate this as a real time computational algorithm.

Pre-access Data Quality Monitor: As well as assessing the data quality output impact at the output level on a case by case basis data quality is assessed at the pre-output disclosure control stage. This analysis is more extensive, potentially covering all analyses that have been submitted to the server and possibly a bank of “typical” analyses. The purpose of this stage is to assess the impact of the disclosure control on the general utility of the data. Again the method is to test whether the results obtained with the treated and untreated data are sufficiently similar. This ensures that the disclosure control that is employed at the pre-access stage, does not produce treated data that is not fit for typical purposes.

Pre-output Disclosure Control: This is relatively uncharted water in terms of disclosure risk assessment and control. Whereas traditional disclosure risk concerns data, here we are concerned with analytical output. Where output is in the form of tables traditional techniques may apply. However, regression lines, residual information, p-values, visual representations of data all carry with them a risk of indirect population unit identification and disclosure of population unit information. Furthermore, one possibility for such a system – the automatic customised linking of datasets behind the fire wall would lead to further disclosure risks. The analysis of these risks is a ripe area for research. Control methods will range from suppressing the output, through reducing the detail on the output, to perturbing the output in some way.

Pre-access Disclosure Control: This corresponds to traditional disclosure control methodology. A risk assessment is carried out on new datasets using up-to-date risk assessment methodology such as SUDA (Elliot et. al 2002) and DIS (Skinner and Elliot 2002) and mixed method disclosure control. The relationship between this module and the “Pre-access Data Quality Monitor” can be viewed as competing components of a constraint satisfaction system. One possibility is that the disclosure control component could produce multiple solutions which the DQI monitor selects from. In general it is envisaged that the disclosure control applied to the data would considerably milder than with freely released data. Some early stage systems dispense with the pre-access disclosure control altogether (O’Keefe 2005). However, the amount carried out at this stage will affect the amount necessary at the output stage, and the payoff between these two phases needs to be considered carefully.

Data Sentry: The data sentry serves multiple functions. Firstly it analyses analytical requests on the part of users against models for intrusive behaviour. If it identifies an intrusive pattern (such as requests for multiple overlapping tables) then it alerts an operator for judgment regarding further action. The data sentry also stores *a priori* refusals, i.e. request patterns which are known to be automatically suppressed, (such as the request for a table above a certain dimensionality).

Data Environment Analysis

The foregoing tackles the disclosure control/grid interface by utilising the power of grid computing to carry out complex disclosure risk/data utility analyses in real time. The concept of data environment analysis(DEA) takes this one step further. DEA takes as its start point the explosion of available personal information in the public domain, in restricted access datasets, in purchasable data, personal web sites etc . Singer(2001) and more recently Purdam et al (2004,2005) have commented on this explosion and how it is driven by collective behavioural tendencies:

1. Collect more information on each population unit
2. Replace aggregate data with person specific databases
3. Given the opportunity collect personal information
4. Link data whenever you can

The increasing availability of personal information has an increasing impact on the disclosure control problem in 3 ways.

1. As personal level information becomes more widely available, the general sensitivity of such information tends to decrease
2. As personal information becomes more available the so does the means for a would be intruder to identify individuals in an Anonymised data release.
3. As information in the wild increases so does the *amount* of identification information that a given intruder might access.

This third point is very problematic for data release agencies such as the NSI's. In practical terms it means that disclosure risk analyses focusing on the data set to be released (as is the norm) become less and less reliable. Once the information on a particular population, available to a would-be intruder exceeds the target dataset in scope and coverage, then in order to understand the risk of a given attack one would need to assess the intruder's data, which is clearly impossible. In this context understand the environment of the data release becomes critical.

One aspect of this is the development of data monitoring processes, where data collection metadata are generated through form analysis, metadata questionnaires and web-crawling software. Such work is ongoing at Manchester on behalf of the Office of national Statistics; Elliot(2005b). Such meta-data provides an understanding of what variables are available under what coverage, which could be linked with the Anonymised release sets. A second thread of research suggested by this is data environment risk analyses. Where contextual data sets are used for disclosure risk assessment both as individual sets and as linked multiple sets. The interaction between say purchasable lifestyle microdata and aggregate census or neighbourhood statistics data is particularly pertinent here; Smith and Elliot (2005). Such analyses are very computationally demanding and therefore are of themselves an ideal grid computing problem. Not only is the data environment an essentially grid computing concept, but the resources needed to conduct such data environment disclosure risk analyses, may well require grid computing level resources in order to be computationally tractable.

The potential value of both data monitoring and data environment analysis is that (i) they provide a potential to enable more appropriate understand of and description of the total real risk of disclosive events, (ii) they give description of the *de facto* attitude of our culture towards personal data, thus enabling us to make more informed decisions on such subjects as privacy and data protection law.

Concluding Remarks

There is a clear dichotomy between data quality and statistical disclosure control measures. Measures involve either the suppression (withholding) of data, or the adding of noise to the raw data before release. The exact point of 'balance' between quality and safety is contextual; in some situations it might be reasonable to allow users almost full access to raw data, whereas other situations, such as unmonitored access to 'the world', would require fairly stringent controls. Any solution must consider both quality and safety simultaneously. Solutions must also be flexible enough to be able to handle a variety of contexts.

In the interests of data quality, disclosure control should only be applied when necessary, rather than as a matter of course. Unfortunately it is not always possible to easily identify when disclosure control is necessary. Disclosure based on multiple and related datasets can require a great computational burden. It follows directly that identifying potentially disclosive sets of data is also computationally demanding (in fact more demanding, as we need to cover all possibilities, whereas a so-called intruder only needs to get lucky once). Thus, it is tempting to err on the side of caution and adopt a conservative approach to control. But whilst the power of grid computing could potentially be used by a data intruder to discover previously unknown information about individuals; it can also be used to improve data quality whilst establishing that no such inferences are possible.

As the amount of data on individual population units stored on computing systems increases, so does the threat to anonymised data releases. The possibility that such data release may come to a halt as it becomes impossible to maintain sufficient data quality whilst meeting ever more stringent disclosure control constraints, means that it is vital that creative data access solutions are uncovered. This paper has described two possible solutions, data environment analysis and automatically controlled remote access systems.

References

- Doyle, P., Lane, J. I., Theeuwes, J. J. M AND Zayatz, L., eds. (2001). *Confidentiality, disclosure and data access*. New York: Elsevier.
- Economist Intelligence Unit (2004). *From grid to great? Grid computing's corporate prospects*. Economist Intelligence Unit.
- Elliot, M. J. (2001). Disclosure risk assessment. In P. Doyle, J.I. Lane, J.J.M. Theeuwes and L. Zayatz, eds. *Confidentiality, disclosure and data access*. New York: Elsevier.
- Elliot, M. J. (2005). Statistical Disclosure Control. In *Encyclopedia of Social Measurement*. New York: Elsevier.
- Elliot, M. J. (2005b). "Data Monitoring, the Grid and Confidentiality" *Paper presented to 1st International symposium on Confidentiality, Privacy and Disclosure*. Manchester May 2005
- Elliot, M. J., and Dale, A. (1999). Scenarios of attack: the data intruder's perspective on statistical disclosure risk. *Netherlands Official Statistics*, Spring 1999, 6-10.

- Fienberg, S. and Makov, U.E. (1998). Confidentiality, uniqueness and disclosure limitation for categorical data. *Journal of Official Statistics*, 14.
- Elliot, M. J., Manning, A. M. and Ford, R. W. (2002). 'A Computational Algorithm for Handling the Special Uniques Problem'. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems* 5(10), pp 493-509.
- Flaherty, D.H. (1999). Visions of privacy: past, present, and future. In *C.Bennett and R.Grant, eds. Visions of privacy*. University of Toronto Press.
- Hakim, C. (1979). Census confidentiality in Britain. In *M.Bulmer, ed. Censuses, surveys and privacy*. London: Macmillan.
- Marsh, C. (1991). Privacy, confidentiality and anonymity in the 1991 Census. In *A. Dale and C. Marsh, eds. The 1991 Census User's Guide*. HMSO.
- O'Keefe, C. (2005) "Privacy Preserving Analytics as a Means of Enabling Access to Health Data Archives" *Paper presented to 1st International symposium on Confidentiality, Privacy and Disclosure*. Manchester May 2005
- OPCS (1992). Statement of policies on confidentiality and security of personal data. Titchfield: OPCS.
- 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. (2005). An Experiment in Naïve Bayesian Record Linkage, *Proceedings of the 55th Session of the International Statistical Institute*. Sydney.
- Skinner, C. J. and Elliot, M. J. (2002). 'A measure of disclosure risk for microdata', *Journal of the Royal Statistical Society Series B*, 64(4) pp 855-867.
- Purdam, K., Elliot, M., Pickles, S. and Smith, D. (2004). Grid Computing and Disclosure Control. *New Review of Information Networking*, Vol. 4., November.
- Purdam, K. and Elliot, M. (2005). A Case Study of the Impact of Disclosure Control on Data Quality in the UK Samples of Anonymised Records. *Proceedings of the 55th Session of the International Statistical Institute*. Sydney.
- Purdam, K. and Elliot, M. and Smith D. (forthcoming 2005) 'Disclosure control and the development of the GRID', *New Review of Information Networking*.
- Purdam, K., Mackey, E. and Elliot, M. (2004) 'The Regulation of the Personal: Individual Data Use and Identity in the UK', *Policy Studies*, Dec 2004
- Wathan, J. (2002). Disclosure Control and Household Size, *SARs online discussion forum*, <http://les1.man.ac.uk/forum/ccsr/Forum5/HTML/000012.html>