Institution: University of the West of England, Bristol

Unit of Assessment: 15 – General Engineering

Title of case study:

Better and faster guarantees of respondent privacy when releasing public statistics

1. Summary of the impact

Several National Statistics Agencies (NSAs) in Europe now use tools based on UWE research to ensure published tables are protected from hacking attempts to breach data privacy. Provision of high-quality data to policy and decision makers is so important that supplying it to NSAs is often mandatory for organisations and individuals. In return, NSAs, such as the UK's Office for National Statistics (ONS), must guarantee a degree of confidentiality. Our research has benefitted ONS, its clients and data providers, by exposing serious flaws in existing methodologies and techniques for protecting confidentiality and by creating tools for (i) auditing and (ii) protecting large complex tables.

2. Underpinning research

National Statistics Agencies (NSAs) supply information that governments, business and public services use to develop policies and make better decisions. This information is derived from data provided by "respondents": organisations, companies and people. Respondent's data must not be published in a form vulnerable to hacking – such that data privacy (and/or anonymity) is breached. This is a legal duty upon NSAs. For tables containing "magnitude" data (e.g., financial turnover), this means a respondent's data cannot be calculated within a given margin of error (typically 10%). The preferred approach is to leave blank certain "sensitive" cells of the table, but this is almost always not enough to protect privacy. A table usually contains sub-totals. As a result, additional cells must also be "suppressed" to prevent hacking via mathematical techniques. However this means that less data is published.

Guaranteeing protection while maximising information that is published is known as the Cell Suppression Problem (CSP). While easy to solve for very small tables through "exact" mathematical optimisation, increases in size and complexity can rapidly cause this approach to fail, so in practice NSAs tend to use rough-and-ready "heuristic" approaches. However, these can greatly over-suppress. UWE research (a collaboration between staff in this Unit and in Computer Science) has developed efficient methods that protect large tables from mathematical hacking, and yet maximise the information published.

The research was carried out at UWE by Dr *Alistair Clark* (Principal Lecturer 1998-2011, Associate Professor 2011-present) and Dr James Smith (Research Fellow/Senior Lecturer 1996-2007, Associate Professor 2007-present). *Clark* has substantial experience of mathematical optimisation applied to production scheduling, supply chains and manpower rostering. He has long researched with Smith at the interface of exact and heuristic approaches, hybridising mathematical and computing techniques to create tools that produce high-quality near-best solutions. In 2006, the Office for National Statistics (ONS) approached them to undertake research into the CSP for large complex tables, subsequently supported by EPSRC funding for Martin Serpell, now a permanent early-career researcher at UWE whom they jointly supervised for a PhD. Their collaboration resulted in the development of:

- "Unpicking" algorithms to rapidly "attack" protected complex tables, revealing serious issues with existing tools used by ONS and NSAs. Protection techniques not only over-suppressed, but also consistently left complex tables vulnerable to hacking (unlike UWE's new methods). This research resulted in section 4's impact points (a, c, f) where ONS took action to improve its service to end-users.
- A more efficient mathematical formulation (R1), allowing larger tables (up to 40,000 cells) to be protected (see impact in section 4(d)).
- Novel methods that pre-process tables to greatly reduce problem size (R1,2), again allowing

larger tables (up to 200,000 cells) to be protected. Section 4 describes the impact of this improved ability, detailing ONS initiatives to strengthen their SDC methods, resulting in impacts 4(d, e).

- New hybrid approaches combining mathematical methods with heuristic approaches. These outperformed all existing approaches on the huge variety of data tables published by ONS (R3) (impacts 4(b, d, e)).
- Evidence showing that the best choice among several alternative methods depended on the type of table being protected, resulting in new robust, flexible and best-performing methods (R3).
- Combined methods to better protect large complex tables while publishing more information for end-users (R3) (impacts 4(d, e)).

3. References to the research

Publications

R1 Serpell, M., Clark, A., Smith, J. and Staggemeier, A. (2008). Pre-processing Optimisation Applied to the Classical Integer Programming Model for Statistical Disclosure Control. *Lecture Notes in Computer Science*, 5262, 24-36. http://dx.doi.org/10.1007/978-3-540-87471-3_3

R2 Serpell, M. Smith, J., Clark, A. and Staggemeier, A. (2013). A Preprocessing Optimization applied to the Cell Suppression Problem in Statistical Disclosure Control. *Information Sciences*, 238, 22-32. http://dx.doi.org/10.1016/j.ins.2013.02.006

R3 Smith, J. E., Clark, A. R., Staggemeier, A. T. and Serpell, M. C. (2012). A Genetic Approach to Statistical Disclosure Control. *IEEE Transactions on Evolutionary Computation*, 16(3), 431-441. http://dx.doi.org/10.1109/TEVC.2011.2159271

Grants

*Improvements to Cell Suppression in Statistical Disclosure Contro*l, PI Clark, CI Smith, Office for National Statistics, 2005-06, £20k

Evaluation of Heuristic Approaches to Statistical Disclosure Control, PI Smith, CI Clark, Office for National Statistics, 2007, £5k

Mathematical Modelling of Statistical Disclosure Control, PI Smith, CI Clark, PhD student Serpell, EPSRC Mathematical CASE award with UK Office for National Statistics, 2007 – 2011, £72k

4. Details of the impact

Our research has benefitted the UK's *Office for National Statistics* (ONS), its clients and data providers, by exposing serious flaws in existing methodologies and techniques for protecting confidentiality and by creating tools for (i) auditing and (ii) protecting large complex tables. UWE's research enabled ONS to handle: (i) larger tables than previously, for example two-dimensional (2D) tables with over 1,000,000 cells; and (ii) complex tables, for example 3D with up to 200,000 cells, and smaller 4D tables.

UWE's research findings were verified in extensive testing at ONS and by the developers of the *tau-Argus* SDC software in 2011 and 2012. Their confirmation of the weaknesses in existing methodologies has led to the following (S1):

a. The incorporation of UWE's unpicker algorithm as an auditing tool into *tau-Argos* version 3 since Sept 2012. This tool is used by NSAs worldwide. It is also the standard recommended by the ONS to UK data providers such as the NHS.

b. A policy decision at ONS in 2012 to change their methodology for protecting magnitude data.

c. A decision by ONS in 2012 to incorporate UWE's unpicker algorithm as part of ONS's standard working practice to validate and protect tables.

d. A decision in 2012 to incorporate UWE's protection tool within ONS's tool set, accompanied in spring 2013 by internal ONS funding for the necessary development work, with initial deployment in 2013.

The research findings also motivated ONS to provide additional funding to ensure further impact:

e. The award in May 2013 of internal ONS funding to complete the roll-out of the full suite of UWE tools and an integrated work-flow that protects the Business Related Employment Statistics from autumn 2013,

f. The award in spring 2013 of ONS internal funding for the development work needed to complete the incorporation in early 2014 of UWE's unpicker algorithm into the desktop workflow management system at ONS's Business Statistics Methods Unit.

There are three groups of beneficiaries of UWE's research:

1. *Respondents* to data surveys (such as individuals and businesses), that supply the source information and who have now been protected by guaranteed confidentiality. Fortunately, to date there have been no (highly publicised) breaches, although the existence of increased computer power and sophistication make this ever more likely.

2. *Data providers* (not ONS and NSAs), such as local and central government agencies, the NHS, businesses and other organisations. The new ability to rapidly validate their tables prior to publication has allowed them to meet their duty of trust with confidence. The release of the updated *tau-Argus* software (and, in ONS's case, the direct deployment of UWE's programs) meant that many of the world's NSAs are now able to meet their legal obligation to protect the confidentiality of the information provided to them (while noting that to date there has never been a (widely publicised) problem to have undermined confidence). Data providers now also benefit from faster automated tools for creating protected tables that replace manual and/or ad-hoc methods used for larger and/or more detailed datasets. This enables more frequent and flexible release of valuable data into the public domain.

3. *Data users* (for example strategic planners and policymakers in business, health, and government), who now benefit from improved access to more detailed information. Providing greater data availability is in line with the *National Data Strategy* developed by the *UK Data Forum* who has stated that "the value to the UK of good data has never been greater". This has also benefitted the official *Open Data Initiative* that aims to make available as much government data as possible.

The unpicker algorithm has now been deployed at ONS and has since 2012 also been incorporated within the *tau-Argus* tool maintained by the *Central Bureau of Statistics* (CBS), Netherlands (S2), and used by NSAs worldwide. This has had a significant trans-European and worldwide impact given that *tau-Argus* is the most popular SDC tool in use throughout the world. It was funded by *Eurostat* and is maintained by *Statistics Netherlands*. It incorporates many cell suppression methods and is free to use.

Strategically, the impact of UWE's research represents a major milestone on the path to "information on demand", with the aim for a fundamental shift in the relationship between data providers and clients. Improved access to information enables business, government or other agencies to make more informed choices and leads to more efficient and effective planning and resource utilisation.

5. Sources to corroborate the impact

[text removed for publication]

Testimonial letters listed below are available from UWE, Bristol.

S1. Testimonial from Head of Collection and Editing Methods and Statistical Computing Unit, Office for National Statistics, Newport.

S2. Testimonial from former tau-Argus project manager (now retired), Statistics Netherlands. Corroborates that UWE research enabled Statistics Netherlands to protect large tables, and that nationals statistics agencies have been enabled to increase the amount of aggregated information published whilst protecting them from attacks to breach contributors' confidentiality.