Information Sciences 238 (2013) 22–32

Contents lists available at SciVerse ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A preprocessing optimization applied to the cell suppression problem in statistical disclosure control

Martin Serpell [a,*], Jim Smith [a], Alistair Clark [a], Andrea Staggemeier [b]

[a] University of the West of England, Frenchay Campus, Coldharbour Lane, Bristol BS16 1QY, United Kingdom
[b] Head of Center for Statistical and Analytical Intelligence, Office for National Statistics, Newport, United Kingdom

ARTICLE INFO

Article history:
Received 28 July 2010
Received in revised form 24 September 2012
Accepted 8 February 2013
Available online 13 March 2013

Keywords:
Information sharing
Statistical table
Cell suppression problem
Statistical disclosure control
Preprocessing optimization
Mathematical programming

ABSTRACT

As organizations start to publish the data that they collect, either internally or externally, in the form of statistical tables they need to consider the protection of the confidential information held in those tables. The algorithms used to protect the confidential information in these statistical tables are computationally expensive. However a simple preprocessing optimization applied prior to protection can save time, improve the resultant protection and on occasions enable the use of exact methods where otherwise heuristic methods would have been necessary. The theory behind this preprocessing optimization, how it can be applied and its effectiveness are described in this paper.

1. Introduction

How do we share information in the digital economy? Benefits can be obtained by pooling private data and then analyzing it, however this risks exposing private data. How to protect privacy in such circumstances is an active area of research [21,22]. This is a problem that has been addressed by National Statistics Agencies for some time but is now a problem that needs to be addressed by a wider variety of organizations. Much information that is made public is done so in the form of statistical tables as publishing the source data would break confidentiality. Each cell in the statistical table will typically contain an aggregate of some information, for example it could be the total cost of a given drug dispensed in a given area of a country. The statistical table will also have row and column totals that are referred to as marginals. Like National Statistics Agencies, many organizations are obliged to maintain the confidentiality of the information they hold. If confidentiality is compromised then organizations can lose the co-operation and goodwill of their data contributors and may even be subjected to legal action.

Confidentiality can be compromised if, for example, there is a cell in the published statistical table that is not an aggregate of data from many contributors but from just one or two. To illustrate this point in Figs. 1–3 the cell data is represented in the format $X_{(y)}$ where X is the sum of the data provided by y contributors. In Figs. 1–3 any cell with less than or equal to two contributors has been suppressed to ensure the confidentiality of those contributors. If the information in the cell can be traced back to source then confidentiality has been compromised. Typically in this situation the cell in question is left blank in the published statistical table and this cell is called a primary suppressed cell.

* Corresponding author.
E-mail address: Martin2.Serpell@uwe.ac.uk (M. Serpell).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.02.006

	1	2	3	4	5	6	Total
A	9 (1)	51 (5)	41 (4)	47 (5)	3 (1)	48 (5)	199 (21)
B	8 (2)	1 (1)	54 (7)	44 (5)	45 (2)	12 (2)	164 (19)
C	8 (11)	70 (8)	6 (2)	76 (8)	64 (7)	21 (2)	245 (38)
D	33 (6)	46 (7)	45 (6)	27 (6)	37 (6)	60 (8)	248 (39)
E	87 (8)	51 (6)	18 (6)	35 (5)	49 (7)	72 (4)	312 (36)
F	48 (7)	59 (6)	39 (5)	65 (9)	35 (6)	58 (9)	304 (42)
Total	193 (35)	278 (33)	203 (30)	294 (38)	233 (29)	271 (30)	1472 (195)

Fig. 1. An example 6×6 statistical table. The eight primary cells have been highlighted with shading. The number of contributors to each cell is given in brackets.

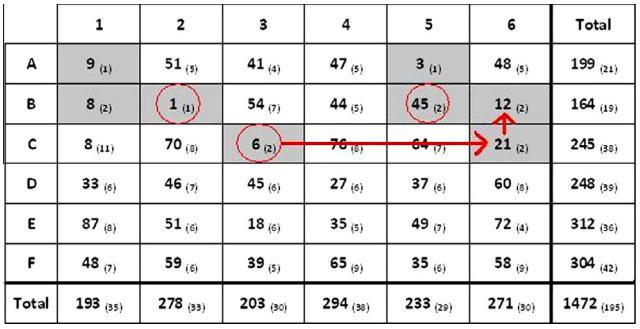

Fig. 2. An example 6×6 statistical table with the sequence in which suppressed cells are exposed using an unpicking algorithm. The eight primary cells have been highlighted with shading. The number of contributors to each cell is given in brackets.

	1	2	3	4	Total
A	100 (1)	100 (1)	100 (1)	100 (4)	400 (7)
B	100 (4)	100 (1)	100 (1)	100 (4)	400 (10)
C	100 (1)	100 (4)	100 (4)	100 (1)	400 (10)
D	100 (1)	100 (4)	100 (4)	100 (1)	400 (10)
Total	400 (7)	400 (10)	400 (10)	400 (10)	1600 (37)

Fig. 3. An example 4×4 statistical table, provided by Lawrence Cox, that cannot be unpicked. The nine primary cells have been highlighted with shading. The number of contributors to each cell is given in brackets.

It is the duty of the National Statistics Agency to determine which cells should be protected. For each primary suppressed cell the National Statistics Agency will also determine a range within which the value of the primary suppressed cells must not be able to be calculated. This range is bounded by the lower protection level, *lpl*, and the upper protection level, *upl*. The values assigned to the *lpl* and *upl* are not published. This alone however does not guarantee confidentiality as the cell value may possibly be calculated by subtracting the other cell values in the same row/column from the row/column total. Therefore to guarantee the confidentiality of the information being published in a statistical table it is necessary to suppress other cells in the statistical table. These other suppressed cells are known as secondary suppressed cells. The combination of primary and secondary suppressed cells are known as a suppression pattern. Finding the lowest cost suppression pattern is known as the cell suppression problem in statistical disclosure control, see [23,14] for more details.

The cell suppression problem is a member of the class of NP-hard problems when solving for optimality. In fact, the problem of finding a secondary suppression pattern is easy to be achieved, for example if all cells are suppressed this is a feasible pattern but clearly not optimal. It is when solving the cell suppression problem optimally that as the size of the table to be protected grows the number of possible solutions that need to be evaluated grows much quicker. For a table with n cells there are $2^n - 1$ possible suppression patterns. Because of the very large number of constraints that define the cell suppres-