Publicacions més rellevants de la línia de recerca:
Control de la revelació estadística

Referència: Castro, J. Minimum-distance controlled perturbation methods for large-scale tabular data protection. *European Journal of Operational Research*, **171** (2006), pp. 39–52.

Abstract: National Statistical Agencies routinely release large amounts of tabular information. Prior to dissemination, tabular data needs to be processed to avoid the disclosure of individual confidential information. One widely used class of methods is based on the modification of the table cells values. However, previous approaches were not able to preserve the values of the marginal cells and the additivity relations for a general table of any dimension, size and structure. This void was recently filled by the controlled tabular adjustment and one of its variants, the quadratic minimum-distance controlled perturbation method. Although independently developed, both approaches rely on the same strategy: given a set of tables to be protected, they find the minimum-distance values to the original cells that make the released information safe. Controlled tabular adjustment uses the L1 distance; the quadratic minimum-distance variant considers L2. This work presents both approaches within an unified framework, and includes a new variant based on L1. Among other benefits, the unified framework permits the simple comparison of the three distances, and a single general result about their disclosure risk. The three distances are evaluated with the unique standard library for tabular data protection currently available. Some of the complex instances were contributed by National Statistical Agencies, and, therefore, are good representatives of theirs real needs. Unlike alternative methods, the three distances were able to solve all the instances, requiring only few seconds for each of them on a personal computer using a general purpose solver. The results show that this class of methods are an effective and promising tool for the protection of large volumes of tabular data. All the linear and quadratic problems solved in the paper are delivered to the optimization community in MPS format.

Referència: Castro, J. A shortest paths heuristic for statistical disclosure control in positive tables. *INFORMS Journal on Computing*, **19(4)** (2007), pp. 520–533.

Abstract: National Statistical Agencies (NSAs) routinely release large amounts of tabular information. Prior to dissemination, tabular data need to be processed to avoid the disclosure of individual confidential information. Cell suppression is one of the most widely used techniques by NSAs. Optimal procedures for cell suppression are computationally xpensive with large real-world data, and heuristic procedures are used in practice. Most heuristics for positive tables (i.e,

cell values are non-negative) rely on the solution of minimum cost network flows subproblems. A very efficient heuristic based on shortest paths was already developed in the past, but it was only appropriate for general tables (i.e., cell values can be either positive or negative), whereas in practice most tables are positive. The method presented in this work sensibly combines and improves previous approaches, overcoming some of their drawbacks: it is designed for positive tables and only requires the solution of shortest path subproblems—therefore being much more efficient than other network flows heuristics. We report an extensive computational experience in the solution of randomly generated and real-world instances, comparing the heuristic with alternative procedures. The results show that the method, currently included in a software package for statistical data protection, fits NSAs needs: it is extremely efficient and provides good solutions.

Referència: Castro, J. Quadratic interior-point methods in statistical disclosure control. *Computational Management Science*, **2(2)** (2005), pp. 107–121.

Abstract: The safe dissemination of statistical tabular data is one of the main concerns of National Statistical Institutes (NSIs). Although each cell of the tables is made up of the aggregated information of several individuals, the statistical confidentiality can be violated. NSIs must guarantee that no individual information can be derived from the released tables. One widely used type of methods to reduce the disclosure risk is based on the perturbation of the cell values. We consider a new controlled perturbation method which, given a set of tables to be protected, finds the closest safe ones—thus reducing the information loss while preserving confidentiality. This approach means solving a quadratic optimization problem with a much larger number of variables than constraints. Real instances can provide problems with millions of variables. We show that interior-point methods are an effective choice for that model, and, also, that specialized algorithms which exploit the problem structure can be faster than state-of-the art general solvers. Computational results are presented for instances of up to 1000000 variables.