Report on DIMACS* Working Group on Data De-Identification, Combinatorial Optimization, Graph Theory, and the Stat/OR Interface

Date of Working Group Meeting: November 9, 2004

Working Group Organizers:

Larry Cox, CDC
Brenda Latka, DIMACS
Fred Roberts, DIMACS

Report Author:
Martin Milanič
RUTCOR, Rutgers Center for Operations Research
Rutgers University

Date of Report: December 2, 2005

*DIMACS was founded as a National Science Foundation Science and Technology Center. It is a joint project of Rutgers University, Princeton University, AT&T Labs-Research, Bell Labs, NEC Laboratories America, and Telcordia Technologies, with affiliated partners Avaya Labs, Georgia Institute of Technology, HP Labs, IBM Research, Microsoft Research, Rensselaer Polytechnic Institute, and Stevens Institute of Technology.

1 Working Group Focus

This meeting was a subgroup meeting of the DIMACS Working Group on Privacy / Confidentiality of Health Data. It was aimed to explore problems in combinatorial optimization, graph theory, and the interface between statistics and operations research that arise from issues of data privacy and, more specifically, data de-identification. These problems have become very important in applications such as health data privacy, government statistical data, and counter-terrorism. The emphasis was on identifying and working on problems of discrete optimization and on identifying and exploring relevant algorithms.

Specific problems of interest discussed/examined from the OR perspective included combinatorial structure of the feasible region defined by a partially specified multi-dimensional table; generating extremal points and statistical samples from a feasible region defined by a system of multi-dimensional tabular constraints; and (near)-optimization of (nonlinear) statistical functions over a system of tabular constraints. These problems recently have been approached algebraically—from the standpoint of the theory of Gröbner bases—but the intended focus of this working group meeting was on combinatorial and mathematical programming approaches and their computability.

The meeting was organized around four presentations given by Dr. Lawrence H. Cox from the Office of Research and Methodology, National Center for Health Statistics, Center for Disease Control and Prevention. The introductory talk as well as the three additional ones, which focused on more specific problems, were all followed by a discussion among the participants.

There were 20 participants in the working group: Aaron Archer from AT&T Labs, Adam Buchsbaum from AT&T Research, Tamra Carpenter from Telcordia, Chandra Chekuri from Bell Labs, Larry Cox from CDC, Peter Hammer from Rutgers University, Howard Karloff from AT&T Research, Jim Kelly from OptTek Systems, Leonid Khachiyan from Rutgers University, Jim Landwehr from Avaya, Brenda Latka from Rutgers University, Vadim Lozin from Rutgers University, David Madigan from Rutgers University, Colin Mallows from Avaya, Fred Roberts from Rutgers University, Mike Saks from Rutgers University, Iraj Saniee from Bell Labs, Bruce Shepard from Bell Labs, Larry Shepp from Rutgers University, and Yehuda Vardi from Rutgers University.

2 Summary of Presentations

2.1 Introductory talk: Data De-Identification and Properties and Problems Related to m-Dimensional Tables

Dr. Cox presented the confidentiality problem which arises in the process of de-identification of data in statistical tables, and various methods and corresponding mathematical models which are used to deal with this problem.

First, he discussed the Complementary Cell Suppression Problem and a closed form optimization model for it. The problem was viewed as a large mixed integer linear programming problem, which gives the optimal result whenever it is computable. To facilitate the presentation, the speaker showed the main ideas of the model on a few simple 3×3 examples. However, as he pointed out, this problem is NP-hard even for the one-dimensional case. Dr. Cox also presented some heuristic methods for solving this problem. These methods turn out to be particularly efficient in the 2-dimensional case, where the network structure of the problem and the total unimodularity of the coefficient matrix are utilized.

Several proposals regarding the above mentioned model were offered by the audience. For example, a question was raised about the possibility of suppressing not only the cell values, but some of the marginal values as well. Dr. Cox responded that this can be done within he general model. Also, Dr. James P. Kelly from OptTek Systems introduced the so called "sliding protection", i.e., the idea of fixing only the lengths of the protection intervals, not their midpoint.

Next, Dr. Cox presented the Controlled Tabular Adjustment method. The corresponding mathematical model is again a mixed integer linear programming problem, but this time the number of integer (binary) variables is reasonably small. As he pointed out, this method can be efficiently extended to preserve important statistics like mean, variance, correlation, covariance and regression.

Another method outlined by Dr. Cox was the Controlled Rounding Problem (CRP), which in two dimensions again efficiently exploits the network structure. The speaker also mentioned its statistical counterpart, the Unbiased Controlled Rounding. Unfortunately, as he emphasized, CRP could not be efficiently extended to higher dimensions. He illustrated this fact with an example of a three dimensional table where the controlled rounding fails.

He continued the talk by presenting the optimal solutions to a particular two-dimensional input produced by each of the above discussed models, plus

one additional model, analogous to the rounding: the so called Perturbation Technique. Very briefly, he mentioned the public use statistical data base query systems, and the way the Gröbner basis decomposition may be used in integer optimization.

Dr. Cox concluded this introductory talk by exhibiting some obstacles which may arise when one tries to extend the model from two dimensions to three.

2.2 Problem 1: Generating Representative Samples of Non-negative Integer-Feasible Solutions to Partially Specified m-Dimensional Tables

To justify generation of random samples of feasible solutions, Dr. Cox first pointed out that the objective value of the problems arising in statistics usually is not so clear as in a business context, and that just finding a feasible solution to the constraints is often satisfactory. He also addressed a question on the issue of negative tabular data raised by Dr. Fred Roberts from DIMACS.

He continued his talk by showing how the "confidentiality" methods are connected to the network structure (in two dimensions and in another class of tables), and that each of the methods can be viewed and computed in terms of integer-feasible moves within the table. To formalize this discussion, he introduced the notion of Markov basis and explained that one can generate a Markov basis with the help of Gröbner bases. However, as he mentioned, the complexity of the Buchberger algorithm used to compute a Gröbner basis is doubly exponential, so this is not a practical solution.

Next, Dr. Cox showed that the basic moves are generated by simple (0,1,-1) moves in two dimensions, but that this fails in higher-dimensional tables. He presented the Diaconis-Sturmfeld algorithm which generates a random sample of integer-feasible solutions to a partially specified table. In general, these are not computable. He provided a classification for "tables of network type" and for such tables presented his own algorithm, based on networks, and thus computable. He concluded his talk with an example of a thin table having precisely two integer-feasible solutions.

In the discussion that followed, questions were raised and comments given such as:

- Dr. Jim Kelly from OptTek Systems asked: In the cell suppression problem, would it be reasonable to replace the bounds of the protection intervals with a more general distribution of the possible cell values?

- Dr. Leonid Khachiyan from Rutgers University observed that generating random samples from the set of extreme points of the polytope is much harder than sampling from the polytope (where feasible real, not necessarily integer, solutions are allowed).

- Dr. Khachiyan also mentioned the fact that there are polynomial-time randomized algorithms for nearly uniform sampling from within any polytope P. For instance, if P is R-rounded (and any polytope in n dimensions can be n-rounded), then the following "hit-and-run" method mixes in time bounded by a polynomial in R and n [**?**, **?**]:

  Pick a starting point x_0 in P and draw a random line L through x_0 in the affine hull of P. Pick at random a point x_1 from the line segment $L \cap P$. Replace x_0 by x_1 and iterate.

2.3 Problem 2: Use of Linear Programming and Search in Lieu of Support of Integer Programming

Dr. Cox provided an overview of confidentiality issues in higher dimensional statistical databases of counts, i.e., their properties and problems which arise while one tries to achieve statistical disclosure limitation in multi-dimensional tables. He illustrated with examples how controlled rounding, controlled random perturbation and assurance of feasibility given consistent integer marginals may all fail in three dimensions.

Another problem of interest outlined by Dr. Cox was the disclosure audit problem. In particular, he emphasized its computational intractability in higher dimensions where, except in specialized cases, any of the sufficient conditions that would guarantee its tractability via LP may fail. However, the situation is manageable in two dimensions, where simple Fréchet and Bonferroni formulae are available.

Dr. Cox pointed out the interesting interdisciplinary interplay between graph theory and statistics (especially applications to log-linear models and data security) arising from a class of decomposable graphical models. He also cited some theoretical results about the unboundedness of integer programming gaps and gaps in integer-feasible values, and emphasized the importance of a recent paper of De Loera and Onn in which they showed that a very special class of tables contains all the complexity of the seemingly much larger class of rational polytopes.

He concluded the talk by briefly summarizing the main issues that had been discussed. Again, he pointed out the De Loera-Onn result which al-

lows one to explore only the narrow range of $b \times c \times 3$ tables, and reminded the audience of the special tabular structure of the coefficient matrix which might be exploited in search of efficient algorithms.

The talk was followed by an interesting discussion, the main points of which are summarized here.

- There were several comments on the computational complexity of the problems:
  Dr. Howard Karloff from AT&T Research and Dr. Michael Saks from Rutgers University expressed the thought that NP-hardness occurs in four-dimensional tables, mentioning a reduction from the three-dimensional matching problem. Dr. Kelly replied by referencing his paper in which he showed that the controlled rounding problem in three dimensions is NP-hard.

- The main concerns raised by Dr. Cox were two. Again, he emphasized the special structure of the coefficient matrix. Second, although the De Loera-Onn result is of a highly theoretical nature, he suggested that any algorithmic ideas or improvements for three dimensions may bring hope for future developments for the higher dimensional cases as well.

- Dr. Cox answered affirmatively to Dr. Roberts' question of whether tables of more than three dimensions arise in practice. However, it is practically impossible to present tables of more than four dimensions in a nice visual manner.

- Finally, he pointed out that the entries of the coefficient matrix in row-reduced, echelon form belong to the ring $\mathbf{Z}[\frac{1}{2}, \ldots, \frac{1}{m-1}]$, where m is the dimension of the table, and proposed to explore how this fact can possibly be utilized. For example, can this property help one to determine the coefficients of a (rounding) cutting plane?

2.4 Problem 3: Identifying and Exploiting Convenient Mathematical Structures in m-Dimensional Tables

Before Dr. Cox began his final talk, the disclosure audit problem for a special case of an $n \times n \times 3$ table was discussed. Dr. Cox raised questions about what kinds of techniques may prove useful in computing the best lower and upper bounds on the entries, and whether a graph theoretical approach (in-

cluding hypergraphs) may help. Again, he reviewed the problems that arise in the three-dimensional case. A lively discussion followed, questions and comments mainly involving (yet unknown) computational complexity of the feasibility problem: *Determine whether an integer table with given marginal sums exists.* A relaxed problem where the table has at most k suppressed cells in a line (i.e., row or column) was also considered. The main observation was that the problem is easy for $k = 1$, but it becomes intractable for $k = 2$. Also, it was noticed that the dimension of the feasible polytope cannot exceed the number of suppressed cells.

In his final talk, Dr. Cox summarized the main results of the previous presentations. He also presented his recent idea of utilizing the LP duality in connection with the transpose of the original coefficient matrix, especially if most dimensions of the corresponding contingency table are small. He illustrated with an example how this approach may reduce a $2 \times 2 \times 2$ table into a restricted two-dimensional table. Finally, the speaker addressed the problem of generating all the marginal and internal cell values for the purpose of statistical database query systems. The total number of values may be enormous even for thin m-dimensional tables.

After the talk a lively discussion ensued. Here is the summary of the discussed issues:

- Dr. Khachiyan discussed the complexity issue of the feasibility problem, the main question being: Is the Feasibility Problem NP-hard?
- Dr. Khachiyan also suggested the use of Lagrangian relaxation approximation schemes (e.g., exponential potential reduction methods) for cell optimization in 3-dimensional contingency tables.
- Dr. Chandra Chekuri from Bell Labs suggested that the maximal gap among integer feasible values could perhaps be bounded above by a maximal gap among rational feasible values, plus a function of the dimension.
- His colleague from Bell Labs, Dr. Bruce Shepherd, pointed out that unlike for the Cell Suppression Problem (with the Fischetti-Salazar model) there is no explicit LP or MILP model for the feasibility problem. However, this should be of no surprise since the problem might be NP-hard, Dr. Kelly replied.

- Can we be satisfied with approximate solutions, since the exact ones are too hard to obtain? Namely, as Dr. Khachiyan pointed out, generating an approximate solution can be done much faster than solving $2n$ linear programming problems. Dr. Cox confirmed that a small relative error (e.g., an accuracy of 95%) is allowable.

- Dr. Khachiyan also proposed a game-theoretical view for this problem, in which one could use the minmax theorem. The projection of the feasible polytope onto the space of suppressed variables should not be thin in any direction. Hence, we may consider the game where we ask an adversary for a direction, and then compute the width of the polytope in this direction via LP. The goal is, of course, to maximize the minimal width of the underlying polytope.

- Questions regarding Markov bases were also raised, e.g., when does the Markov basis of a given 0-1 matrix consist just of simple (0,1,-1) vectors? What do you do if the Markov basis is simple, but not every integer solution in a sequence from one integer-feasible solution to another is feasible? The connection of these questions with the generation of perfect matchings was also discussed.

3 Open Problems and Research Challenges

The working group developed a variety of ideas that will lead to future investigations. The following is a list of open problems and possible future research challenges:

1. **Polyhedral Questions**

   The main objective is to understand better the polyhedral structure of this class of objects. In general, we are given integers $n_1, n_2, \ldots n_m, d_1, \ldots, d_k$ and some $n_1 \times n_2 \times \cdots \times n_m$ matrix A with a subset F of fixed entries. We are interested in the polytope

   $$P = P(A, F, n_1, \ldots, n_m, d_1, \ldots, d_k)\,,$$

   the convex hull of integer $n_1 \times n_2 \times \cdots \times n_m$ matrices X whose marginal totals are given by the d_i's and such that for each $\alpha \in F$, $X_\alpha = A_\alpha$.

   The main questions raised were:

(a) Are there sensible classes of cutting planes for these polytopes? Given the special structure of the coefficient matrix of the tabular system, can one say anything about the coefficients of the cutting planes?

(b) The "Markov Basis Question": Can one understand for which instances the Markov Basis consists of simple (0,1,-1) vectors? (This is part of the question of trying to sample over the polytope P.)

(c) Some additional remarks from Dr. Bruce Shepherd:

In the 2-dimensional case, the most natural class of simple directions is obtained from cycles (i.e., matrices with exactly 2 nonzeroes per row and column). It is plausible to expect that these generalize in the m-dimensional case to matrices with exactly 2 nonzeroes along any axis. Given two matrices in P, one is presumably looking at m-regular directed graphs on the entries. These graphs are encoding the "differences" between the two matrices. By regular is meant that each node either has degree 0, or it has outdegree m or indegree m.

2. **Algorithmic Questions**

(a) The question which seemed to generate the most discussion is that of trying to solve the optimization problems:

$$(*) \quad [\min X_\alpha, \max X_\alpha]$$

subject to X being in the polytope P.

(b) An even more basic question was asked. Given an instance, is the polytope P nonempty? In particular, if the subset F of fixed entries is empty, this is exactly the Feasibility Problem.

(c) The special case where $m = 3, n_1 = n_2 = n$ and $n_3 = 3$ was already considered interesting.

(d) One can view question 1a) as being an attempt to get a handle on these algorithmic questions.

(e) Another version of the question arises when we also have some subset S of entries (outside F) which are sensitive. For each such $\alpha \in S$, we may be given target intervals $[l_\alpha, u_\alpha]$. The idea is that we want to determine if there is a feasible matrix in P for which the min/max intervals $(*)$ contain these target intervals.

(f) k-Approximate versions of this problem were also suggested. For example, given an instance of such a problem, can we determine if there is a matrix for which some k-factor of the interval includes the target intervals?

(g) Alternatively, we may relax even the marginal constraint, so that we are happy with any matrix in some $P(A, F, n_1, \ldots, n_m, d'_1, \ldots d'_k)$ where the d'_i's are within some tolerance of the original d_i's.

(h) Knowing more about these approximate versions may also prove useful to LP based solution techniques.

References

[1] L.H. Cox and L.R. Ernst, *Controlled rounding*, INFOR 20, 1982, 423-432.

[2] B.D. Causey, L.H. Cox and L.R. Ernst, *Applications of transportation theory to statistical problems*, Journal of the American Statistical Association 80, 1985, 903-909.

[3] L.H. Cox and J.A. George, *Controlled rounding for tables with subtotals*, Annals of Operations Research 20, 1989, 141-157.

[4] D. Gusfield, *A graph theoretic approach to statistical data security*, SIAM J. Comput. 17, 1988, no. 3, 552–571.

[5] L.H. Cox, *Network models for complementary cell suppression*, Journal of the American Statistical Association 90, 1995, 1453-1462.

[6] M. Fischetti and J.J. Salazar-Gonzalez, *Models and algorithms for the 2-dimensional cell suppression problem in statistical disclosure control*, Math. Program. 84, 1999, 283-312.

[7] L. Lovász, *Hit and run mixes fast*, Math. Program. 86, 1999, no. 3, Ser. A, 443–461. Available online at http://research.microsoft.com/users/lovasz/papers.htm

[8] L. Lovász and S. Vempala, *Hit and run is fast and fun*, Tech Report MSR-TR-2003-05, available online at http://research.microsoft.com/users/lovasz/papers.htm

4 Acknowledgements

The author would like to thank Dr. Bruce Shepherd for the mathematical formulation of questions of interest and Dr. Brenda Latka and Dr. Fred Roberts for their valuable comments. The author and the DIMACS Center acknowledge the support of the National Science Foundation under grant number CCR 03-14161 to Rutgers University.