Instituto Superior de Economia e Gestão
UNIVERSIDADE TÉCNICA DE LISBOA

Modelos e Algoritmos de Optimização Combinatória no Controlo de Divulgação Estatística

Filipa Duarte de Carvalho

ISEG – Universidade Técnica de Lisboa

CIO – Faculdade de Ciências da Universidade de Lisboa

TÉCNICA DA SUPRESSÃO

Consiste em suprimir os valores das células que representam risco de divulgação de informação confidencial - células sensíveis.

TABELA DE MACRODADOS

	Reg. A	Reg. B	Reg. C	Total Act.
Act. I	11	47	58	116
Act. II	1	**15**	33	49
Act. III	2	31	20	53
Total Reg.	14	93	111	218

Willenborg, L & de Waal, T., 1996

Célula sensível

SUPRESSÃO EM TABELAS DE MACRODADOS

	Reg. A	Reg. B	Reg. C	Total Act.
Act. I	11	47	58	116
Act. II	1	X	33	49
Ac.t III	2	31	20	53
Total Reg.	14	93	111	218

Willenborg, L & de Waal, T., 1996

Célula sensível → Supressão primária X

SUPRESSÃO EM TABELAS DE MACRODADOS

	Reg. A	Reg. B	Reg. C	Total Act.
Act. I	11	47	58	116
Act. II	1	X	33	49
Act. III	2	31	20	53
Total Reg.	14	93	111	218

Willenborg, L & de Waal, T., 1996

X = 49 - (1 + 33) = 15

SUPRESSÃO EM TABELAS DE MACRODADOS

	Reg. A	Reg. B	Reg. C	Total Act.
Act. I	11	47	58	116
Act. II	X	X	33	49
Act. III	X	X	20	53
Total Reg.	14	93	111	218

Willenborg, L & de Waal, T., 1996

Supressões secundárias X

SUPRESSÃO EM TABELAS DE MACRODADOS

	Reg. A	Reg. B	Reg. C	Total Act.
Act. I	11	47	58	116
Act. II	X	X	33	49
Act. III	X	X	20	53
Total Reg.	14	93	111	218

Willenborg, L & de Waal, T., 1996

Número de supressões secundárias = 3

Valor da informação perdida = 34

PROBLEMA DA SUPRESSÃO

Dado um conjunto de supressões primárias,

determinar um conjunto de supressões secundárias

de forma a proteger as supressões primárias

com perda mínima de informação não sensível.

PROBLEMA DA SUPRESSÃO

Dado um conjunto de supressões primárias,

determinar um conjunto de supressões secundárias

de forma a proteger as supressões primárias

com perda mínima de informação não sensível.

O problema da supressão é NP-difícil !

(Kelly *et al.*, 1992)

ABORDAGENS DE OPTIMIZAÇÃO COMBINATÓRIA

ALGORITMOS EXACTOS

(não são eficientes para problemas NP-difíceis)

fornecem soluções garantidamente óptimas

ABORDAGENS DE OPTIMIZAÇÃO COMBINATÓRIA

ALGORITMOS EXACTOS
(não são eficientes para problemas NP-difíceis)
fornecem soluções garantidamente óptimas

ALGORITMOS HEURÍSTICOS
fornecem boas soluções

ABORDAGENS DE OPTIMIZAÇÃO COMBINATÓRIA

ALGORITMOS EXACTOS
(não são eficientes para problemas NP-difíceis)
fornecem soluções garantidamente óptimas

ALGORITMOS HEURÍSTICOS
fornecem boas soluções

ALGORITMOS DE DETERMINAÇÃO DE MINORANTES
permitem avaliar a qualidade das soluções obtidas com os algoritmos heurísticos

1. PROTECÇÃO DE TABELAS GERAIS

2. PROTECÇÃO DE TABELAS NÃO NEGATIVAS

2.1. PROTECÇÃO EXACTA

2.2. PROTECÇÃO INTERVALAR

1. PROTECÇÃO DE TABELAS GERAIS

	C1	C2	C3	Total linha
L1	**X1**	-1	**X2**	3
L2	2	1	1	4
L3	**X3**	0	**X4**	2
Total coluna	2	0	7	9

1. PROTECÇÃO DE TABELAS GERAIS

	C1	C2	C3	Total linha
L1	**1**	-1	**3**	3
L2	2	1	1	4
L3	**-1**	0	**3**	2
Total coluna	2	0	7	9

1. PROTECÇÃO DE TABELAS GERAIS

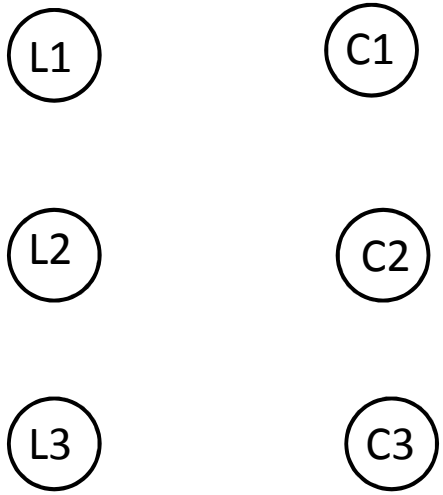
	C1	C2	C3	Total linha
L1	**1001**	-1	**-997**	3
L2	2	1	1	4
L3	**-1001**	0	**1003**	2
Total coluna	2	0	7	9

1. PROTECÇÃO DE TABELAS GERAIS

	C1	C2	C3	Total linha
L1	**1+M**	-1	**3-M**	3
L2	2	1	1	4
L3	**-1-M**	0	**3+M**	2
Total coluna	2	0	7	9

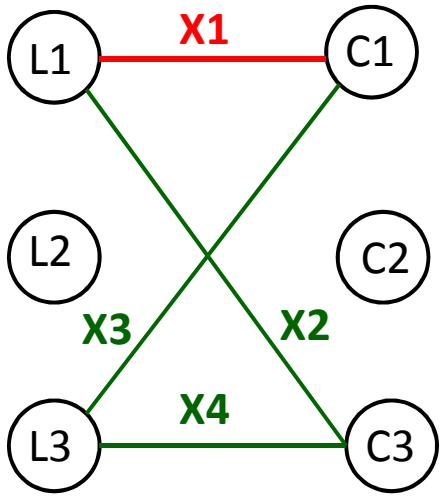
1. PROTECÇÃO DE TABELAS GERAIS

	C1	C2	C3	Total linha
L1	**X1**	-1	**X2**	3
L2	2	1	1	4
L3	**X3**	0	**X4**	2
Total coluna	2	0	7	9



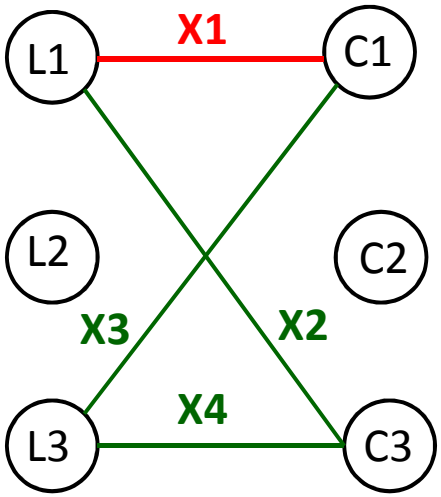
1. PROTECÇÃO DE TABELAS GERAIS

	C1	C2	C3	Total linha
L1	**X1**	-1	**X2**	3
L2	2	1	1	4
L3	**X3**	0	**X4**	2
Total coluna	2	0	7	9



1. PROTECÇÃO DE TABELAS GERAIS

	C1	C2	C3	Total linha
L1	**X1**	-1	**X2**	3
L2	2	1	1	4
L3	**X3**	0	**X4**	2
Total coluna	2	0	7	9



Uma supressão primária
está protegida
se e só se
pertence a um circuito de
supressões

1. PROTECÇÃO DE TABELAS GERAIS

$$\text{minimizar} \sum_{(r_i,c_j)\notin S_1} \beta_{r_i c_j} x_{r_i c_j}$$

para cada célula sensível k

$$y^k{}_{r_i c_j} + y^k{}_{c_j r_i} \leq x_{r_i c_j} \quad (r_i, c_j) \notin S_1$$

$$\sum_{r_i \in R} y^k{}_{c_{j_k} r_i} = 1 \qquad \sum_{c_j \in C} y^k{}_{c_j r_{i_k}} = 1$$

$$\sum_{c_j \in C} y^k{}_{c_j r_i} - \sum_{c_j \in C} y^k{}_{r_i c_j} = 0 \qquad r_i \neq r_{i_k}$$

$$\sum_{r_i \in R} y^k{}_{r_i c_j} - \sum_{r_i \in R} y^k{}_{c_j r_i} = 0 \qquad c_j \neq c_{j_k}$$

Variáveis = 0 ou 1

1. PROTECÇÃO DE TABELAS GERAIS

Algoritmo Lagrangeano combinado com uma heurística

1. PROTECÇÃO DE TABELAS GERAIS

Algoritmo Lagrangeano combinado com uma heurística

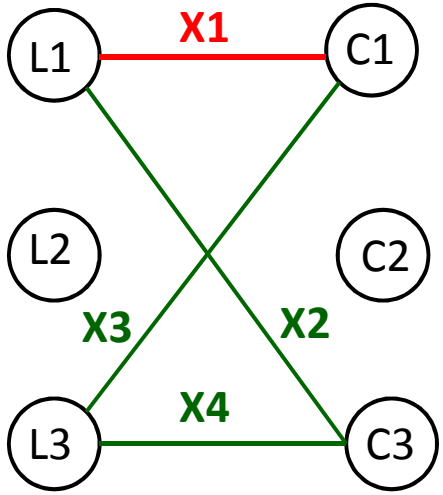
210 tabelas com dimensões 20×10 até 200×200

98% das tabelas protegidas com perda mínima de informação

Tempo computacional médio < 4 minutos

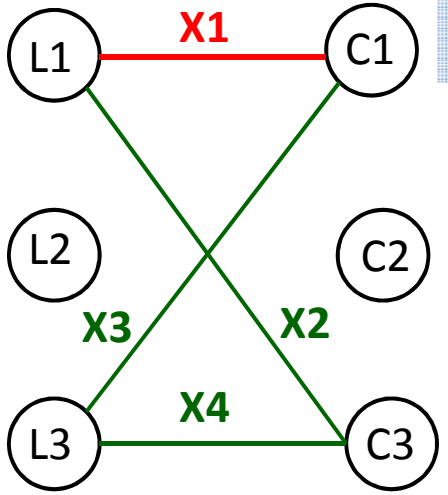
2.1. PROTECÇÃO EXACTA DE TABELAS NÃO NEGATIVAS

	C1	C2	C3	Total linha
L1	**X1**	3	**X2**	5
L2	1	4	1	6
L3	**X3**	5	**X4**	6
Total coluna	4	12	1	17



2.1. PROTECÇÃO EXACTA DE TABELAS NÃO NEGATIVAS

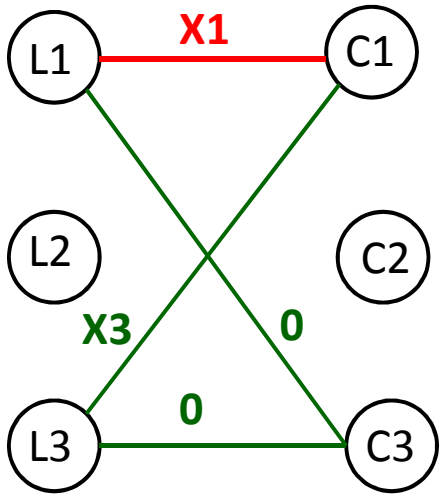
	C1	C2	C3	Total linha
L1	**X1**	3	**X2**	5
L2	1	4	1	6
L3	**X3**	5	**X4**	6
Total coluna	4	12	1	17



X2 + X4 = 0

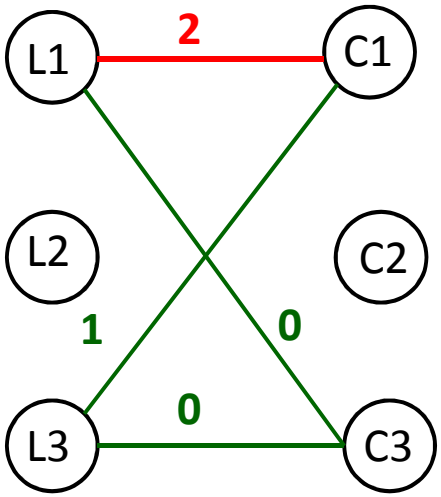
2.1. PROTECÇÃO EXACTA DE TABELAS NÃO NEGATIVAS

	C1	C2	C3	Total linha
L1	**X1**	3	**0**	5
L2	1	4	1	6
L3	**X3**	5	**0**	6
Total coluna	4	12	1	17



2.1. PROTECÇÃO EXACTA DE TABELAS NÃO NEGATIVAS

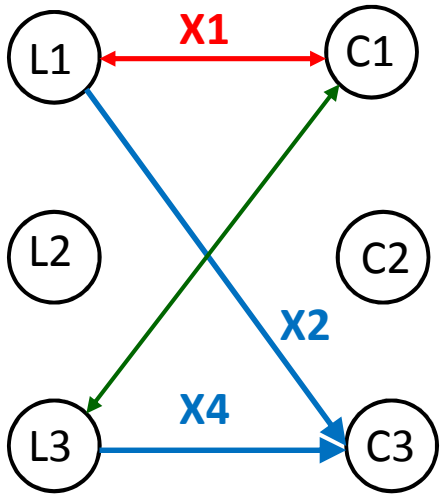
	C1	C2	C3	Total linha
L1	**2**	3	**0**	5
L2	1	4	1	6
L3	**1**	5	**0**	6
Total coluna	4	12	1	17



A supressão primária <u>não tem</u> protecção exacta

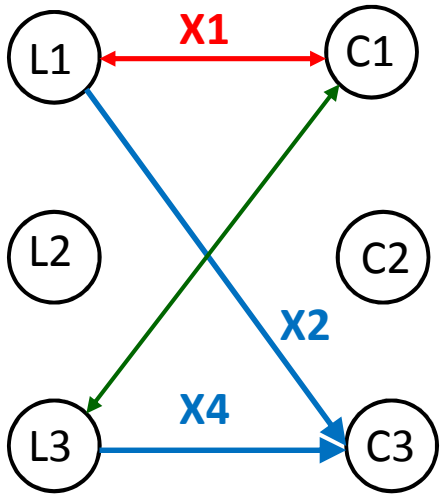
2.1. PROTECÇÃO EXACTA DE TABELAS NÃO NEGATIVAS

	C1	C2	C3	Total linha
L1	**X1**	3	**X2**	5
L2	1	4	1	6
L3	**X3**	5	**X4**	6
Total coluna	4	12	1	17



2.1. PROTECÇÃO EXACTA DE TABELAS NÃO NEGATIVAS

	C1	C2	C3	Total linha
L1	**X1**	3	**X2**	5
L2	1	4	1	6
L3	**X3**	5	**X4**	6
Total coluna	4	12	1	17



Uma supressão primária tem protecção exacta se e só se pertence a um circuito orientado de supressões

2.1. PROTECÇÃO EXACTA DE TABELAS NÃO NEGATIVAS

Algoritmo Lagrangeano combinado com uma heurística

170 tabelas com dimensões 50×10 até 100×100

95% das tabelas protegidas com perda mínima de informação

70% das tabelas protegidas com o número mínimo de supressões secundárias

Tempo computacional médio < 4 minutos

2.2. PROTECÇÃO INTERVALAR DE TABELAS NÃO NEGATIVAS

Definição

Uma célula sensível de valor a_k está protegida se os valores publicados não permitem deduzir nenhum valor para a célula no intervalo $[a_k - l_k, a_k + u_k],\ a_k - l_k \geq 0$.

2.2. PROTECÇÃO INTERVALAR DE TABELAS NÃO NEGATIVAS

	C1	C2	C3	C4	Tot. L
L1	1X0	20	35	X	158
L2	15	10	40	10	75
L3	1X	15	30	X	60
Tot. C	125	45	105	18	293

Intervalo de protecção : [85 , 115]

2.2. PROTECÇÃO INTERVALAR DE TABELAS NÃO NEGATIVAS

	C1	C2	C3	C4	Tot. L
L1	100	20	35	3	158
L2	15	10	40	10	75
L3	10	15	30	5	60
Tot. C	125	45	105	18	293

Intervalo de protecção : [85 , 115]

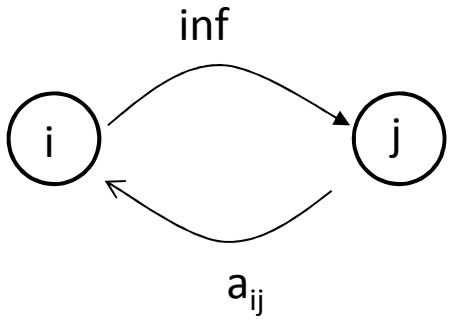
$$X \leq 158 - (20 + 35) = 103 < 115$$

A supressão primária não está protegida !

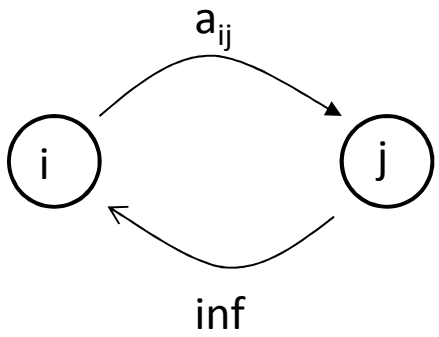
2.2. PROTECÇÃO INTERVALAR DE TABELAS NÃO NEGATIVAS

Rede orientada bipartida com <u>capacidades nos arcos</u>

Célula (i,j)



células internas
e célula do total



células dos subtotais

2.2. PROTECÇÃO INTERVALAR DE TABELAS NÃO NEGATIVAS

Algoritmo Exacto

Formalização de condições necessárias e suficientes de fluxo (Kelly *et al.*, 1992)

Formalização de condições necessárias de padrão

Formalização de condições necessárias de volume

2.2. PROTECÇÃO INTERVALAR DE TABELAS NÃO NEGATIVAS

Algoritmo Exacto

1410 tabelas com dimensões 10×10 até 500×500

95% das tabelas protegidas com perda mínima de informação

Tempo computacional médio < 4 minutos

REFERÊNCIAS PRINCIPAIS

FD Carvalho, MT Almeida. A Three-Phase Algorithm for the Cell Suppression Problem in Two-Dimensional Statistical Tables, *Journal of the Operational Research Society* 59: 556-562, **2008**.

MT Almeida, G Schütz, FD Carvalho. Cell suppression problem: A genetic-based approach. *Computers & Operations Research* 35: 1613-1623, **2008**.

MT Almeida, FD Carvalho. Exact Disclosure Prevention in two-dimensional Statistical Tables. *Computers & Operations Research* 32: 2919-2936, **2005**.

FD Carvalho, MT Almeida. Lower-bounding Procedures for the 2-Dimensional Cell Suppression Problem. *European Journal of Operational Research* 123(1):29-41, **2000**.

MT Almeida, FD Carvalho. Heuristic Methods for the Cell Suppression Problem in General Statistical Tables. Proceedings of the Conference Statistical Data Protection'98, Lisboa, Portugal, **1998**.

FD Carvalho, NP Dellaert, MS Osório. Statistical Disclosure in Two- Dimensional Tables: general tables. *Journal of the American Statistical Association* 89: 1547-1557, **1994**.

J Kelly, BL Golden, A Assad. Cell suppression: Disclosure protection for sensitive tabular data. *Networks* 22: 397-417, **1992**.

FD Carvalho, NP Dellaert, MS Osório. *Statistical Disclosure in Two Dimensional Tables: Positive Tables*. Report 9441/A. Econometric Institute. Erasmus University, Rotterdam, **1992**.

FD Carvalho. *O Problema da Supressão na Protecção de Informação Confidencial: Formalizações e Algoritmos*, Universidade Técnica de Lisboa , Tese de Doutoramento, **2002**.

FD Carvalho. *Optimização em Redes na Protecção de Informação em Tabelas Estatísticas Bidimensionais*, Universidade Técnica de Lisboa , Tese de Mestrado, **1995**.

FD Carvalho. *Statistical Disclosure in two Dimensional Tables*, Erasmus University, Roterdão, Holanda, Trabalho de Estágio, **1992**.