2000

confidentiality and statistical data protection techniques

konfidentzialtasuna eta datuak babesteko teknika estatistikoak

confidencialidad y técnicas estadísticas de protección de datos

LAWRENCE H. COX

40

nazioarteko estatistika mintegia euskadin

seminario internacional de estadística en euskadi

2000

confidentiality and statistical data protection techniques

konfidentzialtasuna eta datuak babesteko teknika estatistikoak

confidencialidad y técnicas estadísticas de protección de datos

LAWRENCE H. COX

40

Eustat

Lanketa / *Elaboración:*

Euskal Estatistika-Erakundea /
Instituto Vasco de Estadística (EUSTAT)

Argitalpena / *Edición:*

Euskal Estatistika-Erakundea /
Instituto Vasco de Estadística
Donostia-San Sebastián, 1 - 01010 Vitoria-Gasteiz

Ale-kopurua / *Tirada:*

500 **ale** / ejemplares

III-2000

Fotokonposaketarako tratamendu informatikoa:
Tratamiento informático de fotocomposición:

ITXAROPENA, S.A.
Araba kalea, 45 - Zarautz (Gipuzkoa)

Inprimaketa eta koadernaketa /
Impresión y encuadernación:

ITXAROPENA, S.A.
Araba kalea, 45 - Zarautz (Gipuzkoa)

ISBN: 84-7749-352-9

Lege-gordailua / *Depósito legal:* S.S. 279/2000

AURKEZPENA

Nazioarteko Estatistika Mintegia antolatzean, hainbat helburu bete nahi ditu EUSTAT-Euskal Estatistika Erakundeak:

– Unibertsitatearekiko eta, batez ere, Estatistika-Sailekiko lankidetza bultzatzea.
– Funtzionarioen, irakasleen, ikasleen eta estatistikaren alorrean interesatuta egon daitezkeen guztien lanbide-hobekuntza erraztea.
– Estatistika alorrean mundu mailan abangoardian dauden irakasle eta ikertzaile ospetsuak Euskadira ekartzea, horrek eragin ona izango baitu, zuzeneko harremanei eta esperientziak ezagutzeari dagokienez.

Jarduera osagarri gisa, eta interesatuta egon litzekeen ahalik eta pertsona eta erakunde gehienetara iristearren, ikastaro horietako txostenak argitaratzea erabaki dugu, beti ere txostengilearen jatorrizko hizkuntza errespetatuz; horrela, gai hori buruzko ezagutza gure herrian zabaltzen laguntzeko.

Vitoria-Gasteiz, 2000ko Apirila

LOURDES LLORENS ABANDO
EUSTATeko Zuzendari Nagusia

PRESENTATION

In promoting the International Statistical Seminars, EUSTAT-The Basque Statistics Institute wishes to achieve several aims:

– Encourage the collaboration with the universities, especially with their statistical deparments.
– Facilitate the professional recycling of civil servants, university teachers, students and whoever else may be interested in the statistical field.
– Bring to the Basque Country illustrious professors and investigators in the vanguard of statistical subjects, on a worldwide level, with the subsequent positive effect of encouraging direct relationships and sharing knowledge of experiences.

As a complementary activity and in order to reach as many interested people and institutions as possible, it has been decided to publish the papers of these courses, always respecting the original language of the author, to contribute in this way towards the growth of knowledge concerning this subject in our country.

Vitoria-Gasteiz, April 2000

LOURDES LLORRENS ABANDO
General Director of EUSTAT

PRESENTACION

Al promover los Seminarios Internacionales de Estadística, EUSTAT-Instituto Vasco de Estadística pretende cubrir varios objetivos:

– Fomentar la colaboración con la Universidad y en especial con los Departamentos de Estadística.
– Facilitar el reciclaje profesional de funcionarios, profesores, alumnos y cuantos puedan estar interesados en el campo estadístico.
– Traer a Euskadi a ilustres profesores e investigadores de vanguardia en materia estadística, a nivel mundial, con el consiguiente efecto positivo en cuanto a la relación directa y conocimiento de experiencias.

Como actuación complementaria y para llegar al mayor número posible de personas e Instituciones interesadas, se ha decidido publicar las ponencias de estos cursos, respetando en todo caso la lengua original del ponente, para contribuir así a acrecentar el conocimiento sobre esta materia en nuestro País.

Vitoria-Gasteiz, Abril 2000

LOURDES LLORENS ABANDO
Directora General de EUSTAT

BIOGRAFI OHARRAK

Lawrence Cox estatistikari matematikari seniorra da Estatu Batuetako Ingurugiro Babeserako Agentzian. Ingurugiro-zientziaren, -kudeaketaren eta informazioaren arazoei buruzko estatistika eta matematika arloko programa baten arduraduna da. Aurretik beste kargu hauek izan ditu: Zientzia Matematikoen Batzordeko zuzendaria Estatu Batuetako Zientzien Akademia Nazionalean eta estatistikari matematikari seniorra Estatu Batuetako Zentsu Bulegoan.

Cox Amerikar Estatistika Elkarteko kidea eta Nazioarteko Estatistika Erakundeko kide hautatua da. Amerikar Estatistika Elkarteko zuzendaritza-batzordeko eta National Computer Graphics Association-eko kidea izan da, eta Amerikar Estatistika Elkarteko bi atalen eta bi batzorderen buruz ere izan da.

Cox Matematikako doktorea da Brown-eko Unibertsitatetik, eta 100 ikerketa-artikulutik gora argitaratu ditu. Zentsu Bulegoan zegoela, SDLri (estatistiken zabalkunde-mugei) buruzko banakako eta erakunde mailako ikerketa-programa bati ekin zion, zeinak argitalpen-bilduma zabala eta astotarikoa eta zenbait SDL informatika-sistema sortarazi baitu. Horiek guztiak Estatu Batuetako zentsu eta inkesta nazionaletan erabiltzen dira. SDLri buruz, hainbat hitzaldi eta aholku eman ditu mundu zabalean.

BIOGRAPHICAL SKETCH

Lawrence Cox is Senior Mathematical Statistician for the U.S. Environmental Protection Agency. He is responsible for a program of mathematical and statistical research relevant to problems in environmental science, management and reporting. Previous positions include Director, Board on Mathematical Sciences, U.S. National Academy of Sciences, and Senior Mathematical Statistician, U.S. Bureau of the Census.

Cox is a Fellow of the American Statistical Association and an Elected Member of the International Statistical Institute. He has served on the Board of Directors of the American Statistical Association and of the National Computer Graphics Association, and has chaired two sections and two committees of the American Statistical Associarion.

Cox holds a Ph.D. in mathematics from Brown University and has published over 100 scholarly papers. Wuile at the Census Bureau, he initiated an individual and institutional program of research on statistical disclosure limitation (SDL) which has led to a large and varied collection of publications and several SDL computer systems used in U.S. national censuses and surveys. He has lectured and consulted extensively on SDL internationally.

NOTAS BIOGRAFICAS

Lawrence Cox es Estadístico Matemático Senior de la Agencia Estadounidense para la Protección Medioambiental. Es el responsable de un programa de investigación estadística y matemática relativo a los problemas de la ciencia medioambiental, gestión e información. Anteriormente ocupó los cargos de Director de la Junta de Ciencias Matemáticas, Academia Nacional de Ciencias Estadounidense y Estadístico Matemático de la Oficina Censal Estadounidense.

Cox es Miembro de la American Statistical Association y Miembro Electo del International Statistical Institute. Ha formado parte del Consejo de Dirección de la American Statistical Association y de la National Computer Graphics Association y ha presidido dos secciones y dos comités de la American Statistical Association.

Cox es Doctor en Matemáticas por la Universidad Brown y ha publicado más de 100 artículos de investigación. Estando en la Oficina Censal, inició un programa individual e institucional de investigación sobre SDL (Statistical Disclosure Limitation), el cual ha generado una amplia y variada colección de publicaciones y varios sistemas SDL para computadoras utilizados en censos nacionales y estudios en Estados Unidos. Ha dado multitud de conferencias y ha sido consultado sobre SDL internacionalmente.

CONTENTS

1. STATISTICAL DISCLOSURE AND DISCLOSURE LIMITATION

STATISTICAL DISCLOSURE

What is statistical disclosure?
Why is it a problem?

The topic "statistical confidentiality" and the technical area "statistical disclosure limitation" that deals with protecting statistical confidentiality have developed over time. Each has improved the other, and both have been affected by developments in computing technology and statistical methodology. While not formal definitions, broad consensus has developed surrounding the following concepts.

What is confidentiality preservation?

* holding close information of a personal or proprietary nature
pertaining to a respondent, and not revealing it
to an unauthorized third party

What is statistical confidentiality protection?

* preserving confidentiality in statistical data products

What is statistical disclosure?

* statistical disclosure occurs when the release of a
statistical data product enables a third party to
learn more about a respondent than the third
party had originally known (T. Dalenius)

Note: "*Respondent*" refers both to direct providers of data
(person, organization,business) and to the
"units of analysis" they represent
(families, corporations, groups)

Is confidentiality important?
Why should the data provider make efforts to preserve respondent confidentiality?

* required by law, regulation or policy

* ethical obligation: the social contract
* practical considerations
  - data accuracy
  - data completeness
  - developing trust

How is confidentiality threatened by release of statistical data?

* overt identification and disclosure of individual respondent data
* identification thru matching of attributes to another data file, leading to disclosure of individual Attributes
* association of a large percentage of an identifiable group with a characteristic *(group disclosure)*

Must confidentiality preservation be absolute?
What is its relative importance?

* the balance issue: right to privacy *vs. need to know*
* absolute confidentiality preservation is impossible: the release of any data divulges something about each respondent
* technology limits what can be done
  - technology to limit disclosure
  - technology to cause disclosure
* in principle:
  - minimum disclosure protection and data quality and completeness standards are not incompatible
  - a joint optimum can be reached
* in practice:
  - the balancing process is iterative
  - incompatibilities are resolved in favor of preserving confidentiality

What factors affect statistical disclosure?

* factors affecting the *likelihood* of Disclosure
  - number of variables

  - level(s) of data aggregation or Presentation
  - accuracy/quality of data
  - sampling rate(s)
  - knowledge about survey participation
  - distribution of characteristics
  - time

* factors affecting the *risk* of Disclosure

  - likelihood of disclosure
  - number of confidential variables
  - sensitivity of confidential data
  - time- target of disclosure
    - # targeted respondent
    - # arbitrary respondent: *fishing expedition*
    - # group disclosure
  - existence/quality of matching files
  - motivation/abilities of intruder
  - cost to achieve disclosure
  - ease to access/manipulate data

STATISTICAL DISCLOSURE LIMITATION (SDL)

What is most commonly done to limit statistical disclosure?

Restricted Access

Restrict who gets data and what data they get. These are primarily administrative solutions.

- sworn special agents
- restricted use agreements
- restrict access to specified data sets or summaries
- restricted data centers
- review/approve analytical outputs

Restricted Data

Restrict the data released (to the public) by limiting the quantity and scope of data release and/or by statistical modification of the data *(statistical disclosure limitation).*

* *sample* the data

  - population file is drawn from a sample survey

    - *subsample* the population file
* abbreviate the data
    - remove direct identifiers
    - reduce the number of variables
    - remove *salient* records and/or records from salient respondents
    - *suppress* item detail
    - *topcode* sensitive items
* aggregate the data
    - *collapse* geographic identifiers
    - collapse data categories
* *switch* data: 1990 U.S. Decennial Census
* release fabricated data

What techniques are available to limit statistical disclosure?

* remove the problem: respondent *waivers*
* anticipate the problem: data release *checklists*
* limit data dissemination
    - restricted access
    - restricted use
    - *encrypted* microdata
    - let the computer decide: statistical data base query systems
* data abbreviation
    - eliminate variables from the released data file
    - eliminate respondents from the released data file
        # eliminate high risk records
        # release a sample
    - suppress selected item detail
    - *truncate* distributions: top (or bottom) code item detail
    - release different file *extracts* to different data users
* data aggregation or grouping
    - coarsen data
        # collapse data categories/detail
        # replace continuous data by categories
    - *microaverage* responses
    - release data summaries
        # tabulations
        # regression equations
        # variance/covariance matrices
* data modification
    - *round* item data (random or Controlled)
    - *perturb* item data (random and/or controlled)
    - replace item data by *imputations*
* data fabrication

- statistical matching
- data *swapping*
- data switching

New approaches to disclosure limitation in microdata

* *supersample* the data file
- sample the (population) data file with replacement
- reweight the new file
- release or subsample the new file

* data fabrication/reconstruction
- *(multiple)* imputation
- *multi-way raking* (iterative proportional fitting)

* statistical *data base query* systems
- static
- dynamic

* use of *contextual* data

* alternative forms of data release
- *interval* data
- maps and graphics

* combine use of respondent waivers and non-disclosure agreements

* probability based measures of disclosure risk combined with information based measures of data utility

* disclosure checklists (ICDAG)

Equation balancing in two-dimensional tables

Equation *balancing* is moving units between cells of a table while maintaining the additivity of the table and the nonnegativity of its entries

For example, to move **10** units between the ***italicized*** cells of the table

20^{+}	10	20	***10***$^{-}$	20	80
10^{-}	10	***20***$^{+}$	5	15	60
40	10	***10***$^{-}$	***20***$^{+}$	10	90
5^{-}	5	15	***10***$^{+}$	5	40
75	35	65	45	50	270

simply add and subtract **5** or **10** units to/from cell values as indicated by +/- signs

The resulting table is

30	10	20	***0***	20	80
5	10	***25***	5	15	60
40	10	***5***	***25***	10	90
0	5	15	***15***	5	40
75	35	65	45	50	270

Both tables are feasible solutions to the table with suppressions

D	10	20	**D**	20	80
D	10	**D**	5	15	60
40	10	**D**	**D**	10	90
D	5	15	**D**	5	40
75	35	65	45	50	270

This illustrates the connection between equation balancing and cell suppression in tables

N.B.: The same connection can be made with data perturbation and rounding methods!!!!

EQUATION BALANCING AND MATHEMATICAL NETWORKS

These topics are pursued later in the course:

* mathematical networks are a natural way to represent two-dimensional tables
* mathematical networks provide a means to perform equation balancing in two-dimensional tables
* mathematical networks enjoy mathematical properties that make their use desirable
    # networks are extremely efficient computationally
    # linear optimization is easily performed over networks
    # clever use of cost *functions* and *capacity constraints* enable network-based cell suppression models

2. STATISTICAL DISCLOSURE LIMITATION FOR TABULAR FREQUENCY DATA. ROUNDING

Controlled Random Rounding

ORIGINAL TABLE

37	3	30	6	4	**80**
1	16	23	5	15	**60**
30	15	8	27	10	**90**
7	1	4	7	21	**40**
75	**35**	**65**	**45**	**50**	**270**

Problem: Small counts (1, 2, 3, or 4) can cause disclosure

Solution: Round all counts to base 5

Constraints:

* multiples of 5 remain fixed
* nonmultiples round to one of the two adjacent multiples of 5
* rounded table is additive
* rounding procedure is random (unbiased)

Method: Controlled Random Rounding using “stepping stones”

Algorithm

* begin with any nonmultiple of the rounding base (B = 5)
  e.g., (1, 2) cell
* create an *alternating cycle*
  involving only nonmultiples
  e.g., (1,2), (1,4), (3,4), (3,3),
  (4,3), (4,2)
* calculate d_+ = maximum amount that can
  be added to (1,2), subtracted from
  (1,4), added to (3,4),......,

subtracted from (4,2) without
violating adjacency constraint:
$d_+ = 1$

* calculate d_- = maximum amount that can
be subtracted from (1,2), added to
(1,4), etc., etc.: $d_- = 2$

* calculate $p_+ = d_-/(d_+ + d_-)$
and $p_- = d_+/(d_+ + d_-)$
$p_+ = 1/3$, $p_- = 2/3$

* randomly select a direction (+ or -)
along the cycle according to p_+, p_-

* perform the selected balanced
Adjustment

Repeat until there are no nonmultiples

N.B.: At least one nonmultiple is transformed to a multiple at each iteration. Therefore, the procedure converges.

Reference: Cox, L. (1987), “A Constructive Procedure for Unbiased Controlled Rounding”, *Journal of the American Statistical Association* **82**, 520-524.

Illustration

37	3	30	6	4	**80**
1	16	23	5	15	**60**
30	15	8	27	10	**90**
7	1	4	7	21	**40**
75	**35**	**65**	**45**	**50**	**270**

Select, say, the - direction, and obtain:

37	**1**	30	**8**	4	**80**
1	16	23	5	15	**60**
30	15	**10**	**25**	10	**90**
7	**3**	**2**	7	21	**40**
75	**35**	**65**	**45**	**50**	**270**

Next,

- select, say, (2,2)
- form the cycle (2,2), (2,3), (4,3),
  (4,5), (1,5), (1,2)
- $d_+ = d_- = 1; p_+ = p_- = ½$
- select, say, - direction

37	2	30	8	3	**80**
1	15	24	5	15	**60**
30	15	10	25	10	**90**
7	3	1	7	22	**40**
75	**35**	**65**	**45**	**50**	**270**

In 3 more iterations, we obtain the final unbiased controlled rounding of the original table:

35	5	30	5	5	**80**
0	15	25	5	15	**60**
30	15	10	25	10	**90**
10	0	0	10	20	**40**
75	**35**	**65**	**45**	**50**	**270**

Controlled Rounded Table

37	3	30	6	4	**80**
1	16	23	5	15	**60**
30	15	8	27	10	**90**
7	1	4	7	21	**40**
75	**35**	**65**	**45**	**50**	**270**

Original Table

Controlled rounding and perturbation

Both perturbation and rounding are based on the same transportation or network model. This model assures the balancing.

Reference: Cox, L., J. Fagan, B. Greenberg and R. Hemmig (1986), “Research at the Census Bureau into Disclosure Avoidance Techniques for Tabular Data”, *Proceedings of the Survey Research Methods Section, American Statistical Association*, Alexandria, VA, 388-393.

To get the probabilities right, use the method illustrated by the preceding example.

Reference: Cox, L. (1987), “A Constructive Procedure for Unbiased Controlled Rounding”, *Journal of the American Statistical Association* **82**, 520-524.

Introduction to transportation theory

Problem

* commodities are to be shipped between m *sources* and n *destinations*
* source i has *supply* of S_i units
* destination j has *demand* for D_j units
* $\sum_i S_i = \sum_j D_j = T$
* *flow* along each source-destination pair (i,j) is denoted x_{ij}; $x_{ij} \geq 0$
* unit cost for (i,j) flow is c_{ij}

Further restrictions

The xij may be capacitated: $l_{ij} \geq x_{ij} \geq u_{ij}$

Objective

Assign nonnegative flows xij so that

$$\sum_{i,j} c_{ij}x_{ij}$$ is minimized

Method

Transportation or *network simplex* algorithm

Algorithm and illustration

x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	**10**
x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	**5**
x_{31}	x_{32}	x_{33}	x_{34}	x_{35}	**5**
x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	**10**
5	**5**	**10**	**5**	**5**	**30**

Minimize: $c(x) = x_{21} + x_{42} + x_{43}$

Method

First, find a *basic feasible solution:*

i.e., a linearly independent set of nonnegative x_{ij} consistent with the table structure

* first maximize x_{11}:

$x_{11} = 5$

# this forces other $x_{i1} = 0$

* proceed similarly with any other x_{ij}

whose value has not already been fixed

This will result in nonnegative values for a set of linearly independent x_{ij} (the *basic variables*) and other *(nonbasic)* variables which are set to 0 (or to lower bounds if capacitated)

After the first (x_{11}) iteration:

5	x_{12}	x_{13}	x_{14}	x_{15}	**10**
0	x_{22}	x_{23}	x_{24}	x_{25}	**5**
0	x_{32}	x_{33}	x_{34}	x_{35}	**5**
0	x_{42}	x_{43}	x_{44}	x_{45}	**10**
5	**5**	**10**	**5**	**5**	**30**

At the next iteration, say, x_{22}:

5	0	x_{13}	x_{14}	x_{15}	**10**
0	5	0	0	0	**5**
0	0	x_{33}	x_{34}	x_{35}	**5**
0	0	x_{43}	x_{44}	x_{45}	**10**
5	**5**	**10**	**5**	**5**	**30**

At the next iteration, say, x_{43}:

5	0	0	x_{14}	x_{15}	**10**
0	5	0	0	0	**5**
0	0	0	x_{34}	x_{35}	**5**
0	0	10	0	0	**10**
5	**5**	**10**	**5**	**5**	**30**

In 2 more iterations (x14 and x35), we obtain an initial basic feasible solution:

5	0	0	**5**	0	**10**
0	**5**	0	0	0	**5**
0	0	0	0	**5**	**5**
0	0	**10**	0	0	**10**
5	**5**	**10**	**5**	**5**	**30**

At this stage, either specialized transportation optimization algorithms or general linear programming can be used to produce an optimal solution

Optimal solution

0	**5**	0	**0**	**5**	**10**
0	0	**5**	0	0	**5**
0	0	**5**	0	**0**	**5**
5	0	0	**5**	0	**10**
5	**5**	**10**	**5**	**5**	**30**

Min $\{x_{21} + x_{42} + x_{43}\} = 0$

Reference: Causey, B., L. Cox and L. Ernst (1985), “Applications of Transportation Theory to Statistical Problems”, *Journal of the American Statistical Association* **80**, 903-909.

Connection with Controlled Rounding

37	3	30	6	4	**80**
1	16	23	5	15	**60**
30	15	8	27	10	**90**
7	1	4	7	21	**40**
75	**35**	**65**	**45**	**50**	**270**

Original Table

Subtract lower multiple of base B=5 from all internal entries and adjust totals:

2	3	0	1	4	**10**
1	1	3	0	0	**5**
0	0	3	2	0	**5**
2	1	4	2	1	**10**
5	**5**	**10**	**5**	**5**	**30**

Controlled rounding of original table is equivalent to solving transportation problem subject to: $x_{ij} = 0$ if original entry = 0; $0 \leq x_{ij} \leq 5$ otherwise

Optimal controlled rounding is possible using to any linear objective function. Familiar objective functions are "linearized".

Reference: Cox, L. and L. Ernst (1982), "Controlled Rounding", *INFOR* **20**, 423-432.

3. STATISTICAL DISCLOSURE LIMITATION FOR TABULAR MAGNITUDE DATA

3.1. COMPLEMENTARY CELL SUPPRESSION

Recall the general principles. Confidentiality protection is:

Ethical statistical practice

* social *contract* between statistical
organization and respondent
* often required by law or government regulation

Sound statistical practice

Sound statistical practice
* maintain high levels of response
* preserve data completeness and accuracy

"In return for providing information of a private nature, the respondent is assured that the statistical organization will hold this information ***confidential****".*

Protecting Confidentiality in Tabular Data

1. Define disclosure

* intuitive/historical notions
* quantitative definition
* measurement of disclosure

2. Organize the aggregation structure

* aggregation reduces disclosure; however,
* manipulation of aggregation equations
  can result in disclosure

3. Limit disclosure

* suppress *disclosure cells*
* suppress additional *complementary suppressions*
  until no published aggregate is a disclosure
* verify protection *(disclosure audit)*

Mathematical Methods for SDL in Tabular Data

Problem	Method
Define/measure disclosure	Sensitivity measures
Organize aggregations	Mathematical lattices Mathematical networks General linear programming
Limit disclosure	Combinatorial algorithms Mathematical networks General linear programming Graph theory Stochastic optimization Integer linear programming

DEFINING AND MEASURING STATISTICAL DISCLOSURE

Notation

X denotes a statistical cell. Respondents contributing to **X** are denoted j = 1, 2,...., LAST. Respondent *contributions* are: $x_1 \geq x_2 \geq ....\geq x_{LAST} \geq 0$.

The *cell value* is: **V(X)** = $x_1 + x_2 +....+ X_{LAST}$.
The *tail* is : $x_{m+} = x_m +....+ X_{LAST}$

Standard disclosure rules

(n,k) - dominance rule (U.S. Census Bureau):

X is a *disclosure cell* if its n largest respondents contribute more than k% of the cell value:
$x_1 +....+ x_n > (k/100)V(X)$

$((100 - k)/100)(x_1 +....+ x_n) - (k/100)x_{(n+1)}+ > 0$

$\mathbf{S_{n,k}(X) = (x_1 +....+ x_n) - (k/(100-k))x_{(n+1)}+ > 0}$

p-percent rule:

X is a *disclosure cell* if the second largest respondent can use the cell value to estimate x_1 to within p%

$V(X) - x_2 < ((100 + p)/100)x_1$

$(p/100)x_1 - x_3+ > 0$

$\mathbf{S_{p\%}(X) = x_1 - (100/p)x_3+ > 0}$

pq ambiguity rule (Statistics Canada):

X is a *disclosure cell* if the second largest respondent can estimate x_1 to within p%, given that it can estimate any respondent to within q%

$x_1 - (100/p)x_3+ > 0$

But second largest knows

$x_{3+} > L = (q/100)x_{3+}$

So

$x_1 - (100/p)(q/100)x_{3+} > 0$

$\mathbf{S_{pq}(X) = x_1- (q/p)x_{3+} > 0}$

NOTE: When q = 100, $S_{pq}(X) = S_{p\%}(X)$.

Combination rules:

$$S(X) = S_1(X) + S_2(X)$$

$$S(X) = \text{Max}\{S_1(X), S_2(X)\}$$

These are examples of *linear sensitivity measures*

Reference: Cox, L. (1981), “Linear Sensitivity Measures in Statistical Disclosure Control”, *Journal of Statistical Planning and Inference* **5**, 153-164.

LINEAR SENSITIVITY MEASURES

A linear sensitivity measure is:

$$S(X) = \sum_{1}^{\text{LAST}} \omega_i x_i \qquad (\omega_1 = 1)$$

Subadditivity: If

$$\omega_1 \geq \omega_2 \geq .... \geq \omega_{LAST}$$

then S(X) is subadditive:

$$S(X \cup Y) \leq S(X) + S(Y)$$

for disjoint cells **X** and **Y**. This assures the disjoint union of two nondisclosure cells remains a nondisclosure cell.

Measuring Disclosure

S(X) measures the amount of additional suppression needed to disclosure-limit X:

$$P(X) = S(X) / |\omega_{LAST}|$$

If **X** is a nondisclosure cell, then P(X) measures the inherent protection **X** provides to its respondents.

Examples

In each example, **X** is given by: $x_1 = 70$, $x_2 = 15$, $x_3 = 5$, $x_{4+} = 10$

(3,80)-dominance rule:

$S_{3,80}(X) = x_1 + x_2 + x_3 - (80/20)x_{4+}$
$= 70 + 15 + 5 - (4)10$
$= 50 > 0$
X is a disclosure cell under the (3, 80) -dominance rule
X requires 50/4 = 12.5 units of additional protection

20-percent rule:

$S_{20\%}(X) = x_1 - (100/20)x_{3+} = 70 - (5)15 = -5 < 0$
X is a nondisclosure cell under the 20-percent rule

20%-50% ambiguity rule:

$S_{20\%\text{-}50\%}(X) = x_1 - (50/20)x_{3+}$
$= 70 - (2.5)15$
$= 32.5 > 0$
X is a disclosure cell under the 20%-50% ambiguity rule
X requires 32.5/2.5 = 13 units of additional protection

COMPLEMENTARY CELL SUPPRESSION

Example

Table 1

20	10	20	10	20	80
10	10	***20***	5	15	60
40	10	10	***20***	10	90
5	5	15	***10***	5	40
75	35	65	45	50	270

Cells in ***italics*** are disclosure cells.
Assume each of the *4* disclosure cells requires disclosure protection = *50%* of cell value.

Combinatorial Method (INTRA)

Constructs complementary suppression patterns for a two-dimensional table that limit disclosure along rows and columns, using the minimum number of suppressions necessary.

Selects one such minimal pattern involving the least total value suppressed.

Constructs a suppression pattern for the *entire table* in one operation

Treats 3-dimensional tables as stacked 2-dimensional tables that are processed in an organized sequential manner

Can fail to detect disclosure resulting from combination of row and column equations. Such “misses” are detected and corrected by an *audit* procedure

Implemented in 1977 & 1982 U.S. Economic Censuses

Reference: Cox, L. (1980), “Suppression Methodology and Statistical Disclosure Control”, *Journal of the American Statistical Association* **75**, 377-385.

Example

P = primary disclosure cell
C = complementary disclosure cell

P	10	20	**C**	20	80
C	10	**P**	5	15	60
40	10	**C**	**P**	10	90
C	5	15	**P**	5	40
75	35	65	45	50	270

In practice, all suppressed cells would be replaced by a unique symbol **D** as an additional safeguard against disclosure

This pattern is **ideal**:
* it provides full disclosure protection
* it suppresses the fewest number of cells possible *(4)*
* it suppresses the least total value possible *(35)*

In general, these optima are not achieved simultaneously

EXERCISE

D	D	D	0	10
0	D	D	0	9
D	0	0	D	8
D	0	0	D	7
12	8	7	7	34

Table with Suppressions
INTRA Condition Satisfied
But One Cell Disclosed Exactly

General Linear Programming Method (Statistics Canada)

Organizes the aggregation structure as one *system* of linear equations

Provides disclosure protection to disclosure cells *sequentially*, beginning with the cell requiring the most protection

Optimizes a *logarithmic* function of cell value attempting to balance minimum number of cells suppressed with minimum total value suppressed

Is *self-auditing* within the entire aggregation system

Reference: Robertson, D. (1993), “Cell Suppression at Statistics Canada”, *Proceedings of the Annual Research Conference, U.S. Bureau of the Census*, Washington, DC: Department of Commerce, 107-131.

Network Method (U.S. Bureau of the Census)

Represents a two-dimensional table as a *mathematical network*, which can be solved very quickly

Can be *extended* from a single table to sets of tables related *hierarchically*, such as all SIC tables for a geographic area

Is *self-auditing* for one table or hierarchically related set of tables

Provides disclosure protection to disclosure cells *sequentially*, beginning with the one needing the most protection; treats three-dimensional tables as *stacks* of two-dimensional tables

Implemented in the 1987, 1992 & 1997 U.S. Economic Censuses

Example

P_i = primary disclosure cell protected at iteration **i**
C_i = complementary suppression selected at iteration ***i***

$\mathbf{P_1}$	$\mathbf{C_1}$	20	10	20	80
$\mathbf{C_1}$	$\mathbf{C_1}$	$\mathbf{P_2}$	5	15	60
40	10	$\mathbf{C_2}$	**P**	10	90
$\mathbf{C_2}$	$\mathbf{C_2}$	15	**P**	5	40
75	35	65	45	50	270

This pattern involves ***6*** complementary suppressions of total value ***50*** units

Although the purpose of the 2nd iteration is to protect $\mathbf{P}_2$ (only), the 2 remaining disclosure cells in column 4 also receive full protection at that stage. This is not uncommon and is one reason why the processing sequence is from cells needing most protection to cells leading least protection

This example can be used to illustrate both the General Linear Programming and Network methods, under appropriate simplifying assumptions

MATHEMATICAL NETWORKS, EQUATION BALANCING AND CELL SUPPRESSION

Equation Balancing and Cell Suppression

Equation balancing is moving units between cells of a table while maintaining the additivity of the table and the nonnegativity of its entries

For example, to move **10** units between the ***italicized*** cells of the table

20^{+}	10	20	***10***$^{-}$	20	80
10^{-}	10	***20***$^{+}$	5	15	60
40	10	***10***$^{-}$	***20***$^{+}$	10	90
5^{-}	5	15	***10***$^{+}$	5	40
75	35	65	45	50	270

simply add and subtract **5** or **10** units to/from cell values as indicated by +/- signs

The resulting table is

30	10	20	***0***	20	80
5	10	***25***	5	15	60
40	10	***5***	***25***	10	90
0	5	15	***15***	5	40
75	35	65	45	50	270

Both tables are ***feasible solutions*** to the table with suppressions

D	10	20	**D**	20	80
D	10	**D**	5	15	60
40	10	**D**	**D**	10	90
D	5	15	**D**	5	40
75	35	65	45	50	270

This illustrates the connection between equation balancing and cell suppression in tables

N.B.: The same connection can be made with data perturbation and rounding methods!!!!

Equation Balancing and Mathematical Networks

The following will be developed in the next section:

* mathematical networks are a natural way to represent two-dimensional tables
* mathematical networks provide a means to perform equation balancing in two-dimensional tables

Mathematical networks enjoy mathematical properties that make their use desirable

* networks are extremely efficient computationally
* linear optimization is easily performed over networks
* clever use of *cost functions* and *capacity constraints* enable network-based cell suppression models

Reference: Cox, L. (1995), “Network Models for Complementary Cell Suppression”, *Journal of the American Statistical Association* **90**, 1453-1462.

Extending Network Methods Beyond a Single 2-way Table

Hierarchies of tables

* network-based methods can be extended to hierarchies of 2-way tables

Tables with subtotals

* tables with subtotal constraints along either rows or columns—but not both—are equivalent to hierarchies oftables and network methods extend to these problems
* network methods do not admit theoretical extensions to tables with both row and column subtotal constraints, but heuristic solutions based on networks are possible

Reference: Cox, L. and J. George (1989), "Controlled Rounding for Tables with Subtotals", *Annals of Operations Research* **20**, 141-157.

Higher dimensional tables

* network methods do not admit theoretical extensions to higher dimensions, but heuristic solutions based on networks have been implemented

Limitations of Complementary Cell Suppression Methods

Shared limitations of U.S. Bureau of the Census and Statistics Canada methodologies for complementary cell suppression:

* cannot minimize number of cells suppressed
* seek to minimize total value suppressed but in fact do not

This is evident by examination of their *objective functions*. (Formulas below represent the type of objective function used for purposes of illustration)

USBC: $\mathbf{c(x)} = \sum_{\mathbf{i,j}} \mathbf{a_{ij}x_{ij}}$

STATCAN: $\mathbf{c(x)} = \sum_{\mathbf{i,j}} \mathbf{(\log(1+a_{ij}))x_{ij}}$

$\mathbf{a_{ij}}$ denotes cell value. $\mathbf{x_{ij}}$ denotes the *amount of protection* complementary cell (i, j) is providing. This causes the problem.

* to minimize total value suppressed, need x_{ij} = 0 or 1
* to minimize number of cells suppressed,

need x_{ij} = 0 or 1 and $\mathbf{c(x)} = \sum_{\mathbf{i,j}} \mathbf{x_{ij}}$

3.2. SOLVING THE COMPLEMENTARY CELL SUPPRESSION PROBLEM USING NETWORK OPTIMIZATION

Mathematical Networks

A ***mathematical network N*** consists of

- * ***nodes***, represented as points **i, j** (and,here, corresponding to rows and columns of the 2-way table **A**)
- * ***arcs***, represented as (directed) arrows between ordered pairs of nodes, and represented here as (**i, j, +)** (resp., **(i, j, -)**) for flows from node i to node j (resp., from j to i)

Non-negative quantities $\mathbf{x}_{\mathbf{ij+}}$, $\mathbf{x}_{\mathbf{ij-}}$ ***(network flows)*** are assigned along arcs, subject to:

- * ***node requirements*** $\mathbf{r}_{\mathbf{k}}$**:** prescribed values for *net flow* (= total out-flow - total in-flow) at each node
- * ***arc capacities*** $\mathbf{u}_{\mathbf{ij+}}$**,** $\mathbf{u}_{\mathbf{ij-}}$**:** upper limits on flows $\mathbf{x}_{\mathbf{ij+}}$, $\mathbf{x}_{\mathbf{ij-}}$

Arc costs per unit flow are denoted $\mathbf{c}_{\mathbf{ij+}}$, $\mathbf{c}_{\mathbf{ij-}}$

x = column vector of arc flows

u = column vector of arc capacities

c = row vector of arc costs

R = column vector of node requirements

B = ***node-arc incidence matrix*** of ***N:*** one row for each node and one column for each arc

$\mathbf{b}_{\mathbf{k,l}}$ = +1 if node for row **k** of **B** is from-node of arc for column **l** of **B**

$\mathbf{b}_{\mathbf{k,l}}$ = -1 if node is to-node of corresponding arc

$\mathbf{b}_{\mathbf{k,l}}$ = 0 otherwise

A network optimization ***(N, c)*** defines a *linear program:*

$$\mathbf{min\ cx:\ \ Bx = R,\ 0 \leq x \leq u}$$

Integrality Property of Networks: If node requirements **R** and capacities **u** are integer, so is any optimal solution **x**

An ***alternating cycle Y*** of ***N*** is a cyclic sequence of arcs for which:

- * successive arcs in opposite directions
- * to-node of each arc equals from-node of its successor

It is possible to add any value **q** to the flow along each positive (resp., negative)arc of **Y** and correspondingly subtract **q** from the flow along each negative (resp., positive) arc of **Y** without violating **BX = R**

Careful choice of **q** ensures $\mathbf{0} \le \mathbf{x} \le \mathbf{u}$:

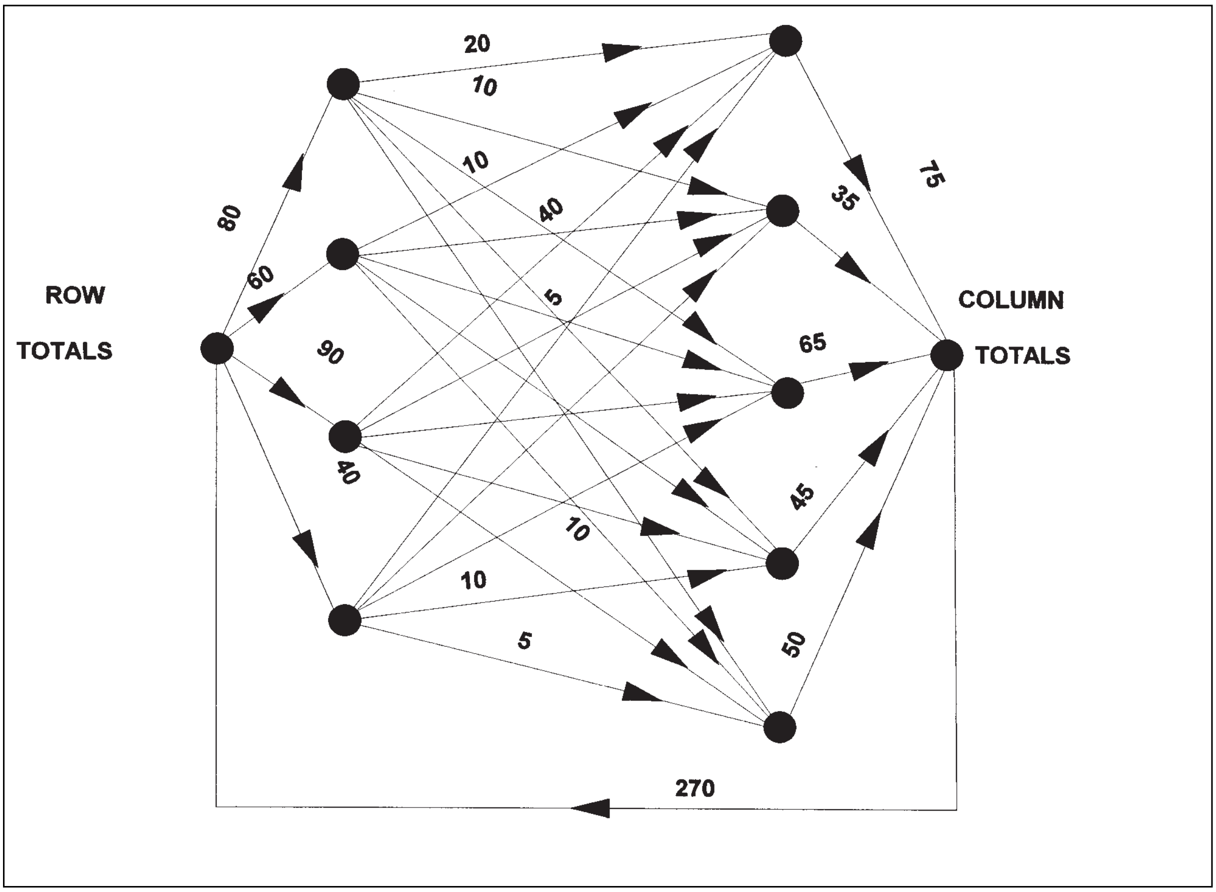
$$\mathbf{0 \le q \le g(Y) = \min\{a_{ij}: (i, j) \in Y\}}$$

In this manner, networks may be used for *equation balancing*.

Networks provide a natural mechanism to represent 2-way tables

20 **(D)**	10	20	10	20	**80**
10	10	20 **(D)**	5	15	**60**
40	10	10	20 **(D)**	10	**90**
5	5	15	10 **(D)**	5	**40**
75	**35**	**65**	**45**	**50**	**270**

Table 1



Network for Table 1

3.3. NETWORKS, EQUATION BALANCING AND COMPLEMENTARY SUPPRESSION: TECHNICAL DETAILS

Networks and equation balancing

N can represent the *deviations* from current values:

$\mathbf{x}_{ij+}$ = amount to be added to $\mathbf{a}_{ij}$
$\mathbf{x}_{ij-}$ = amount to be subtracted from $\mathbf{a}_{ij}$

Costs $\mathbf{c}_{ij*} > \mathbf{0}$ ensure *complementarity*,
i.e., $\mathbf{x}_{ij+}\mathbf{x}_{ij-} = \mathbf{0}$

Arc capacities $\mathbf{u}_{ij-} \leq \mathbf{a}_{ij}$ ensure non-negativity

Equation balancing and complementary cell suppression

Disclosure cell **(I, J)** with protection limit $\mathbf{p}_{IJ}$ is ***protected*** if there exists an alternating cycle **Y** comprising only suppressed cells that contains **(I, J)** and satisfies bold $\mathbf{g(Y)} \geq \mathbf{p}_{IJ}$

Network model for complementary suppression in a 2-way table

(I, J) denotes the *target* disclosure cell **S** denotes the set of already-suppressed cells

NODES m+n+2 nodes, corresponding to internal and totals rows and columns of **A**

ARCS 2(m+1)(n+1) arcs: positive arcs $\mathbf{x}_{ij+}$, negative arcs $\mathbf{x}_{ij-}$ corresponding to all internal and totals cells of A

NODE REQUIREMENTS **R = 0**

ARC CAPACITIES $\mathbf{u}_{IJ+} = \mathbf{1}$, $\mathbf{u}_{IJ-} = \mathbf{0}$. All other $\mathbf{u}_{ij*} = \mathbf{1}$, **except**: in a positive table **A**, $\mathbf{a}_{ij} = \mathbf{0}$ implies $\mathbf{u}_{ij*} = \mathbf{0}$

This Defines the Network *N*

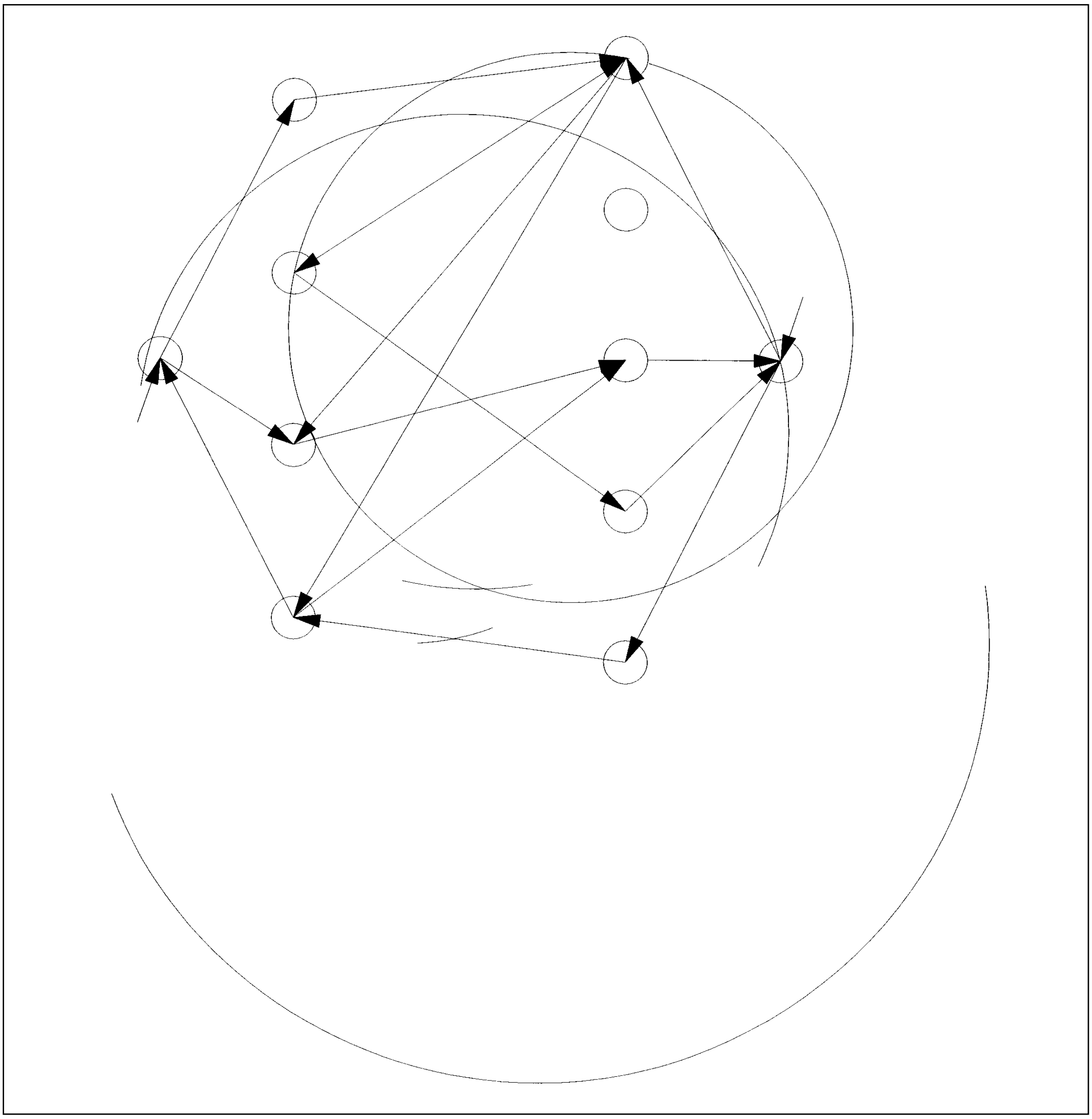
ARC COSTS $\mathbf{c_{IJ+} = -\infty; c_{IJ-} = 1}$

For (i, j) NOTIN S ~ (I, J), $\mathbf{c_{ij*} = 1}$

For (i, j) NOTIN S, arc costs obey $\mathbf{c_{ij*}} \geq \mathbf{\#S}$

* min total suppressed: $\mathbf{c_{ij*} = \#S + a_{ij}}$
* min number of suppressions: $\mathbf{c_{ij*} = \#S + 1}$

This Defines Network Optimization(s) *(N, c)*



Network for Complementary Cell Suppression: (I, J) = (1, 1)
(Only Some Arcs Drawn)

How *(N, c)* solves complementary suppression

(1) Integrality of Networks **(+)** {0, 1}arc
Capacities $\Longrightarrow$ optimal **x** satisfy $\mathbf{x_{ij*} = 0}$ **or 1**

(2) $\mathbf{c_{IJ+} << 0}$ **(+)** $\mathbf{R = 0}$ $\Longrightarrow$ optimal solution
includes alternating cycle containing **(I, J)**

(3) Other $\mathbf{c_{ij*} > 0}$ **(+)** $\mathbf{u_{IJ-} = 0}$ $\Longrightarrow$ no trivial
cycles $\mathbf{(x_{ij+} = x_{ij-} = 1)}$

(1) + (2) + (3) $\Longrightarrow$ ***(N, c)*** **creates a cell suppression pattern containing (I, J) by means of the following rule**

Complementary suppression rule

Suppress *cell* ***(i, j)*** *if* x_{ij+} ***or*** x_{ij-} ***= 1*** *in the optimal solution of* ***(N, c)***

* method provides disclosure protection to one disclosure cell at a time
* method applied iteratively in order of decreasing protection Required
* at each iteration, a single ***protection cycle*** is created containing no subcycles or ***superfluous suppressions***
* if **(I, J)** does not receive full protection, another iteration is performed

Multiple-cell complementary suppression

Single-cell complementary suppression requires at least one optimization for each suppressed cell (I, J). The complementary suppression pattern for the entire table (or hierarchy of tables) equals the union of the individual patterns, minus any superfluous complementary suppressions (which must be detected separately)

This procedure is prone to produce suppression patterns that are suboptimal for the table as a whole

A method that provides disclosure protection to all suppressed cells in a single iterative step is is desirable–***multiple-cell complementary suppression***

The existence of efficient multiple-cell methods is clouded by *NP hardness* results. However, a partial solution under the minimum number of additional suppressions criterion exists which is formulated here as a network optimization problem

The method is based on the following problem

Problem: Given a single two-way table A and a set of suppressions S, select a minimal set of complementary suppressions so that each row and column containing suppression(s) contains at least two suppressions

Cox (1980) provided an optimal multiple-cell solution to this Problem, which is weaker than the minimum number of suppressions problem because a solution to the Problem may fail to be contained in an alternating cycle

Theorem (Cox 1980): Let A be a general two-way table with suppressions under the minimum number of additional suppressions criterion. Let **m'** (respectively, **m"**) denote the number of rows of A containing suppressions (resp., requiring an additional suppression), and define **n'** (resp., **n"**) similarly. Without loss of generality, assume m" ≥ n" and m" ≥ 1. If max {m', n'}= 1, then the Problem can be solved by three additional suppressions. Otherwise, **m"** additional suppressions suffice

There are two degenerate cases:

* **max {m', n'}= 1:** use ***(N, c)***
* **n" = 0** and **n' = 1** (all suppressions are along a single column

U): use ***(N, c)***, except $c_{iJ} = -(c_0 + 1)$ for *(i,J)* ∈ *U*

The principal cases, **min {m', n'}>1**, are solved using the network optimization ***(M, e)***:

NODES Node set of **M** equals node set of *N*

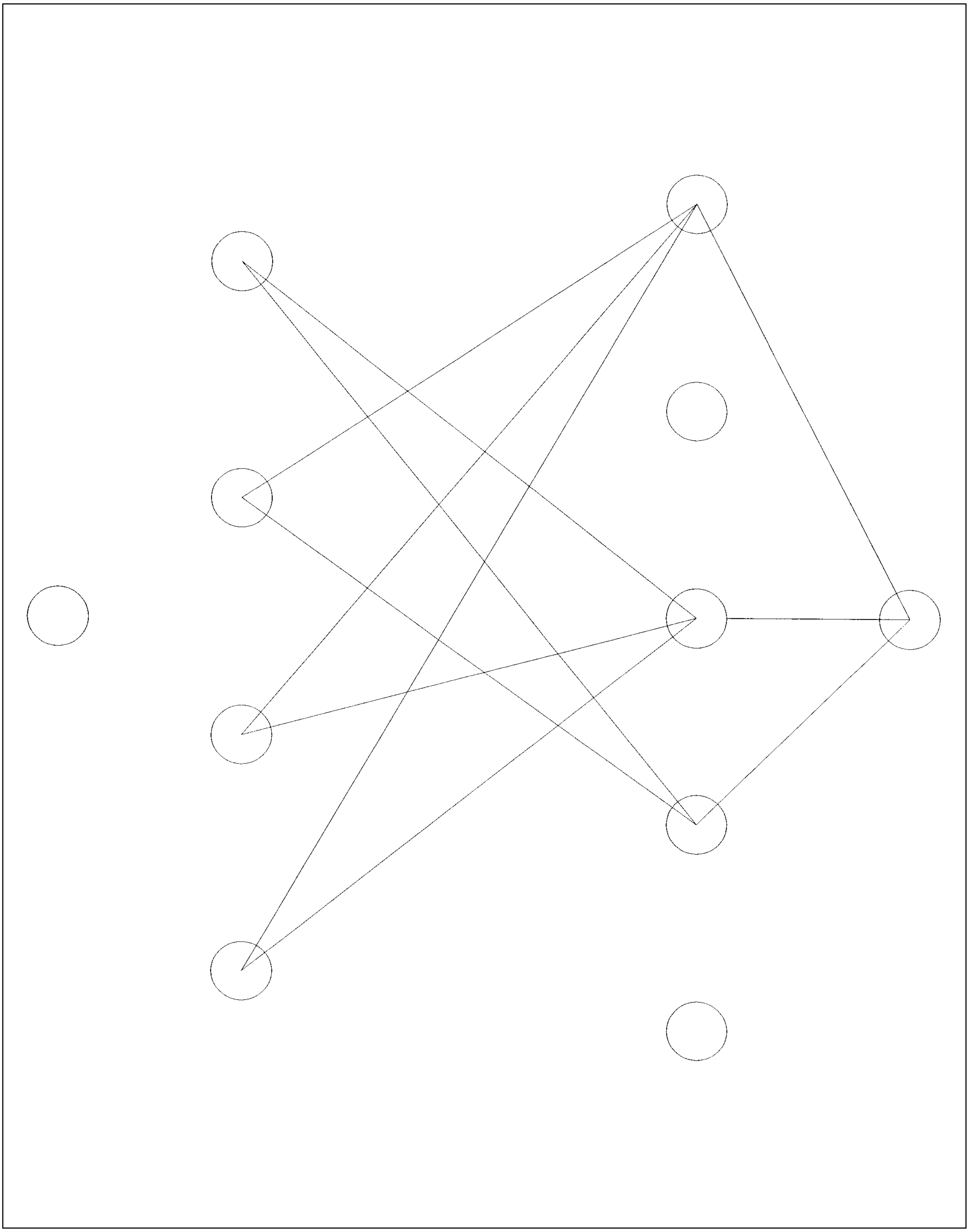
NODE REQUIREMENTS Node requirements are zero, except: the requirement for the first grand total node equals **m"**, that for the second grand total node equals **-(m" - n")**, and that for each column node requiring additional suppression equals **-1**

ARCS All arcs are positive. One arc from the first grand total node to each row node **i** requiring additional suppression (flows denoted $\mathbf{x_{i,n+1}}$), one arc from the first grand total node to each column node j requiring additional suppression (flows denoted $\mathbf{z_j}$), one arc from each row node i requiring suppression to each column node containing suppressions (flows denoted $\mathbf{x_{ij}}$), and one arc from the first grand total node to the second grand total node (flow denoted $\mathbf{x_{m+1,n+1}}$)

ARC CAPACITIES $\mathbf{u_{m+1,n+1} = m" - n"; u_{ij} = 0}$ **if**

(i, j) ∈ *S*; capacities equal **1** otherwise

COSTS e Cost on arcs from the first grand total node to column nodes requiring additional suppression and to the second grand total node equals **+1**; all other arc costs equal zero



Network for Multiple Cell Suppression

20 **(D)**	10	20	10 **(C)**	20	80
10 **(C)**	10	20 **(D)**	5	15	60
40	10	10 **(C)**	20 **(D)**	10	90
5 **(C)**	5	15	10 **(D)**	5	40
75	35	65	45	50	270

Optimal Cell Suppression Pattern
for Table 1

4. STATISTICAL DISCLOSURE LIMITATION FOR MICRODATA

STATISTICAL DISCLOSURE IN MICRODATA

What factors affect statistical disclosure?

* factors affecting the *likelihood* of disclosure

- number of variables
- level(s) of data aggregation or presentation
- accuracy/quality of data
- sampling rate(s)
- knowledge regarding survey participation
- distribution of characteristics
- time

* factors affecting the *risk* of disclosure

- likelihood of disclosure
- number of confidential variables
- sensitivity of confidential data
- time
- target of disclosure
    - # targeted respondent
    - # arbitrary respondent: *fishing expedition*
    - # group disclosure
- existence and quality of matching files
- motivation/abilities of intruder
- cost to achieve disclosure

What is most commonly done to limit statistical disclosure?

* *restrict* data dissemination
* *sample* the data
  - population file is drawn from a sample survey
  - *subsample* the population file
* abbreviate the data
  - remove direct identifiers
  - reduce the number of variables
  - remove *salient* records and/or records from salient respondents
  - *suppress* item detail
  - *topcode* sensitive items
* aggregate the data
  - *collapse* geographic identifiers
  - collapse data categories
* *switch* or blank-and-impute data

What techniques are available to limit statistical disclosure?

* disappear the problem: respondent *waivers*
* limit data dissemination
  - restricted access
  - restricted use
  - *encrypted* microdata
  - let the computer decide:
    statistical data base query systems
* data abbreviation
  - eliminate variables from the released data file
  - eliminate respondents from the released data file
    # eliminate high risk records
    # release a sample
  - suppress selected item detail
  - *truncate* distributions: top (or bottom) code item detail
  - release different file *extracts* to different data users
* data aggregation or grouping
  - coarsen data
    # collapse data categories/detail
    # replace continuous data by categories
  - *microaverage* responses
  - release data summaries
    # tabulations
    # regression equations
    # variance/covariance matrices

* data modification
    - *round* item data (random or controlled)
    - *perturb* item data (random or controlled)
    - replace item data by *imputations*
* data fabrication
    - statistical matching
    - data *swapping*
    - data switching

MATRIX MASKING METHODS FOR MICRODATA

The statistical literature contains many methods for disclosure limitation in microdata. However, their use by statistical agencies and understanding of their properties and effects has been limited. For purposes of furthering education, research, and use of these methods, and facilitating their evaluation, comparison, implementation and quality assurance, it would be desirable to formulate them within a single framework. A framework called ***matrix masking***—based on ordinary matrix arithmetic—is presented, and explicit matrix mask formulations are given for the principal microdata disclosure limitation methods in current use. This enables improved understanding and implementation of these methods by statistical agencies and other practitioners.

The Problem

Many methods for disclosure limitation in microdata **(MDL)** have been proposed. However,

* few methods have been fully developed
* fewer have been implemented, tested and used
* fewer still are in use by statistical agencies
* selection of methods by agencies is unsystematic
* effects of methods on data usefulness has not been examined
* different methods have not been compared
* quality-assured software for methods is not available for wide distribution

Matrix masks are proposed as a means to

* simplify the development, revision, quality assurance, and transportability *(sharing)* of MDL software
* provide a mechansism to compare different MDL methods
* provide a framework for empirical and analytical studies of the effects of MDL methods on data usefulness
* provide a common language in which to discuss and develop MDL methods

Matrix Masks

A microdata file containing **p** attribute values for each of **n** (respondent-level) data records is represented as an **nxp** matrix **X** with entries $\mathbf{x_{ij}}$

Unless stated otherwise, **X** contains no missing values

A ***matrix mask*** **(A, B, C)** is a transformation of **X** of the form:

$\tilde{\mathbf{X}}$ **= AXB + C, with A, B ≠ 0**, whose sums and products involve ordinary matrix addition and multiplication

As **A** operates across the rows of **X, A** is called a ***record transforming mask***

As **B** operates down the columns of **X, B** is an **attribute transforming mask**

C is a ***displacing mask*** as the entries of X are additively displaced by amounts given by the entries of **C**

An ***elementary matrix mask*** of **X** is a matrix mask of the form **AX**, **XB**, or **X + C**.

A, B, C are not necessarily fixed, e.g., to add random noise to attributes, **C** is a random matrix

A, B, C can depend upon **X**

* to displace **X** by additive random noise proportional to size, draw cij randomly from $\mathbf{N(0, (kx_{ij})^2)}$, for constant **k**
* for **A = X', M = AX** is sufficient for ordinary least squares regression

Matrix masks are **iterations** of elementary matrix masks

Notation

I denotes the identity matrix
Z denotes the zero matrix
J the matrix of all ones
$\mathbf{Ui_j}$ denotes the matrix all of whose entries equal **0**, except $\mathbf{u_{ij} = 1}$
I is a square matrix; **Z**, **J** and $\mathbf{U_{ij}}$ need not be square

When used as a pre-(post-) multiplier

* $\mathbf{U_{ij}}$ retains the values of only one row (column) of the matrix it multiplies
* **J** produces the sum of the values along the column (row) of the matrix it multiples

Dimensions of submatrices vary between and within individual formulations here and will be specified for clarity

REPRESENTATION OF MDL METHODS AS ELEMENTARY MATRIX MASKS

Removing and Selecting Microdata

Attribute suppression of the **k**$^{\mathbf{th}}$ attribute is represented as an attribute transforming

Mask bold $\tilde{\boldsymbol{X}} = \boldsymbol{XB}$; **B** is the **px(p-1)** matrix

$$\boldsymbol{B = Supp\ (k) =} \left| \begin{array}{ccc} \boldsymbol{I} & & \boldsymbol{Z} \\ & \boldsymbol{Z} & \\ \boldsymbol{Z} & & \boldsymbol{I} \end{array} \right|$$

upper I-matrix is **(k1)x(k1)**
lower I-matrix is **(p-k)x(p-k)**
central Z-matrix is **1x(p1)**

X is **nx(p-1)**

Suppression of several attributes is represented as a product of **B**-matrices

Supp(k)Supp(j) first suppresses the **k**$^{\mathbf{th}}$ attribute of **nxp** matrix **X**, then suppresses the **j**$^{\mathbf{th}}$ attribute of resulting **nx(p-1)** matrix **XSupp(k)**

Supp(k) is **px(p-1)**
Supp(j) is **(p-1)x(p-2)**

Record deletion of the **h**$^{\mathbf{th}}$ record

* analogous to attribute suppression
* = suppression of columns from the transposed **X**-matrix
* deletion of **h**$^{\mathbf{th}}$ record is represented by the
  record transforming mask $\tilde{\boldsymbol{X}} \boldsymbol{= AX}$
* A is **(n-1)xn** matrix identical to
  Supp(h), except
  - central **Z**-matrix is **(n-1)x1**
  - upper and lower **I**-matrices are
    (h1)x(h1) and **(n-h)x(n-h)**

This **A**-matrix is denoted by **Del(h)**

$$\boldsymbol{Del\ (h) =} \left| \begin{array}{ccc} \boldsymbol{I} & & \boldsymbol{Z} \\ & \boldsymbol{Z} & \\ \boldsymbol{Z} & & \boldsymbol{I} \end{array} \right|$$

To **delete multiple records**

- to delete **h**th and **i**th records, **i > h**, use **Del(i1)Del(h)**
- to **systematically delete** every **h**th record **(n = hr)**, use the **A**-matrix comprising **r** block **(h-1)xn** matrices **Del(h)** arranged vertically
- this generalizes to nonsystematic removal of multiple records

Record sampling

- if viewed as the complement of record deletion, the problem is solved
- to **systematically sample** every **h**th record, use the **rxn A**-matrix whose **q**th row is the **1xn U**-matrix $U_{1,qh}$
- to draw an arbitrary sample of size **s** comprising records indexed by set **S = {**s_v**: v=1,...,s}**, use **sxn A**-matrix **Sam(X, S)**, each row of which is a **1xn U**-matrix $\mathbf{U_{1,s_v}}$

Aggregating and Grouping Microdata

Attribute aggregation

To replace **j**th attribute by sum of **j**th and **k**th **(j < k)**, use **px(p-1) B**-matrix

$$\boldsymbol{Agg\ (j,k)} = \left|\begin{array}{ccc} \boldsymbol{I} & & \boldsymbol{Z} \\ & \boldsymbol{U_{1j}} & \\ \boldsymbol{Z} & & \boldsymbol{I} \end{array}\right|$$

upper **I**-matrix is **(k1)x(k1)**
lower **I**-matrix is **(p-k)x(p-k)**
central **U**-matrix $\mathbf{U_{1j}}$ is **1x(p1)**

Aggregation-deletion over multiple attributes is a product of **Add(j, k)**-matrices

- $\mathbf{B_1}$ aggregates 2 attributes to a subtotal that replaces first; deletes second
- iteratively apply $\mathbf{B_2,..., B_{c1}}$ until done

Aggregation-replacement of **j**th attribute without deleting **k**th: use **pxp B**-matrix

$$\mathbf{Add(j,k) = I + U_{kj}}$$

Include more summands **v** by adding more $\mathbf{U_{vj}}$

To create totals attribute without replacement, use **px(p+1) B**-matrix

$$\mathbf{B} = \left| \mathbf{I} \,\middle|\, \mathbf{U_{j1} + U_{k1}} \right|$$

Collapsing categories

* represent categorical variable of c mutually exclusive categories by **c** columns of **X**
* presence (absence) of trait is recorded as **1 (0)**
* grouping the **c** attribute categories to form one combined category is simply aggregation across the c attributes, replacing first attribute by the aggregate, and deleting all others
* this is represented as a product of **B**-matrices as previously

Microaggregation sums attribute values across microrecords in predetermined groupings

* records to be microaggregated are consecutive; group sizes $\mathbf{n_1,...., n_s; n = n_1 + n_2 +....+ n_s}$
* use diagonal block **nxn A**-matrix; main diagonal comprised of an ordered block $\mathbf{n_v xnv}$ **J**-matrices, **v = 1,...., s**
* microaggregates replace original values

To replace each group by one microaggregated record, use $\mathbf{1xn_v}$ **J**-matrices

To construct ***microaverages***, replace each **J**-matrix by a $\mathbf{(1/n_v)J}$

Scrambling Record Order

* use a stochastic **A**-matrix
* given reordering of the rows (records) (a **permutation P** of the row numbers **{1, ..., N}**)
* if row **h** is moved to row position **i (P(i) = h)**, then **i**th row of **A** is **1xn U**-matrix $\mathbf{U_{1h}}$

A is denoted by **Reo(P)**

Reo(P) facilitates **data swapping**

Rounding and Perturbing Microdata

Data rounding and additive data perturbation can be represented as displacing masks

* for each $\mathbf{x_{ij}}$, the displacement $\mathbf{c_{ij}}$ to be applied to $\mathbf{x_{ij}}$ is computed according to the rounding or perturbation algorithm, with $\mathbf{c_{ij} = 0}$ for those values not subject to change

* $\mathbf{\tilde{X} = X + C}$ is the matrix of rounded (perturbed) values

Attribute Topcoding

Attribute topcoding replaces all values of the **j**th attribute above a predetermined (large) value $\mathbf{T_j}$ with **Tj**

Topcoding is also known as ***truncation*** of the distribution

Given $\mathbf{x_{ij} = f_{ij}T_j + r_{ij}}$; $\mathbf{f}_{ij}$ the integer quotient, $0 \leq \mathbf{r_{ij}} < Tj$ the remainder; compute

$$\mathbf{t_{ij} = (Max \{r_{ij}, (T_j +1)^{f_{ij}}-1\}) \; mod \; (T_j +1)}$$

To topcode **X**, apply the displacing mask

$$\mathbf{Tco(X) = (t_{ij}\} - x_{ij})}$$

Representation of Data Masks as Matrix Masks

The matrix masks previously were applied to the full matrix **X**

It is desirable to apply masks selectively to subsets of records (rows) and attributes (columns) of **X** ***(subset selection)***

An important application is ***blurring***– selective microaveraging

The ability to apply all methods developed here to selected subsets is accomplished via matrix masks that **extract** and **restore** arbitrary submatrices

Method

* Apply an extraction masks **Ign(Q, R)** defined in terms of selected records **Q** and attributes **R**, to extract the submatrix ***X***
* Apply the matrix mask **M** mask corresponding to the desired operation to ***X*** using the methods developed previously
* Apply a restoring mask **Res(Q, R)** to the created matrix to restore the ignored values of **X**

The **ignoring mask** is **Ign(Q, R) = AXB**

- **A** is the **nxn** matrix $\boldsymbol{A = \sum_{i^{epsilon}Q} U_{ii}}$
- **B** is the **pxp** matrix $\boldsymbol{B = \sum_{j^{epsilon}R} U_{jj}}$

A leaves values in selected rows unchanged, replacing other values by **0**

B has similar effect on columns

To preserve dimensions of ***X***, deletion operations are modified to replace deleted values by **0**

The final masked matrix is

$$\boldsymbol{\tilde{X} = M (Ign (Q, R)) + X - Ign (Q, R)}$$

Alternative Formulations

The **U**-matrices provide a powerful tool for representing and combining microdata masks

$$\mathbf{Supp(k)} = \sum_{j<k} \mathbf{U_{jj}} + \sum_{j>k} \mathbf{U_{j,j-1}} \qquad \mathbf{px(p-1)}$$

$$\mathbf{Del(h)} = \sum_{i<h} \mathbf{U_{ii}} + \sum_{i>h} \mathbf{U_{i-1,i}} \qquad \mathbf{(n-1)xn}$$

$$\mathbf{Agg(j, k) = Supp(k) + U_{kj}, j<k} \qquad \mathbf{px(p-1)}$$

$$\mathbf{= Supp(k) + U_{k,j-1}, j>k}$$

$$\mathbf{Reo(P)} = \sum_{i=1}^{n} \mathbf{U_{P(i),i}} \qquad \mathbf{nxn}$$

5. NEW PROBLEMS AND RESEARCH AREAS IN STATISTICAL DISCLOSURE LIMITATION

Small area and small population data

* Small Area Data needs
- fine-grained geographic detail
- high signal to noise ratio
- low missing data rate
- observations from joint distributions
- longitudinal or repeat-sample observations
- respondent-level data *(microdata)*
- ability to *link* two files

* confidentiality problems associated with these needs
- high identification risk for individual respondents
- potentially high (differential) sampling rate(s)
- potential for exact disclosure
- increased likelihood of identification and/or increased sensitivity of data
- ethical/legal aspects of data reuse and file linkage often unclear

Are Small Populations only defined by small geographic areas?

** individual identifiers are*

(combinations of) characteristics with high identification power

- social security number *(SSN)*
- race + sex + date of birth in a small community
- address + age

* individual identifiers are composed of *key* variables
* geography is typically a reliable key variable

However,
* IRS, SSA, banks, credit (card) companies, and employers could use SSN an individual identifier
* health insurers could use an uncommon disease + other data as an individual identifier
* USDA could use land size, type and use as an individual Identifier
* DOE could use energy utilization data as an individual identifier

So,
* confidentiality problems for Small Area Data are essentially
the same as those for *Small Population Data*
- studies involving individuals with rare diseases
- econometric studies involving sparsely or unevenly populated industry groups
- studies involving sparsely or unevenly populated occupation groups
- studies involving the wealthy
- studies involving individuals with high exposure to environmental pollutants
- studies focused on distributional tails

Except that,
* geographic identifiers are available for use by almost anyone
* one can at least map geographically referenced data, but must rely on more "direct" means of data dissemination for Small Population Data

New directions in SDL for microdata

* *supersample* the data file
- sample the (population) data file with replacement
- reweight the new file
- release or subsample the new file

* data fabrication/reconstruction
- *(multiple)* imputation of confidential data
- *multi-way raking* (iterative proportional fitting)

* statistical *data base query* systems
- static
- dynamic

* use of *contextual* data
* alternative forms of data release
- *interval* data
- maps and graphics

* combine use of respondent waivers and data user non-disclosure agreements
* probability based measures of disclosure risk combined with information based measures of data usefulness

Emerging SDL methods for microdata

* recent trends
  - multiple imputation (Rubin 1993)
  - log-linear models for categorical data (Fienberg and Makov 1997)
* problems
  - model selection confines all Relationships/inferences
  - identification/disclosure thru
    # high dimensional resolution
    # distributional tails

Precisely where inference from sample surveys is least reliable

* alternative
  - *resample* microdata with replacement
  - minor perturbation, etc., to avoid exact duplicates
  - some topcoding, reweighting
* advantages
  - new-sample "sample uniques" not necessarily = original-sample uniques
  - original-sample uniques not nec. new-sample "sample uniques"
  - distributional center unaffected

Public use statistical data bases

Standard Approaches

- query size restrictions
- round/perturb query responses
- perturb underlying microdata
- use of atomic queries

An excursion through 3-dimensional tables with a view towards public use statistical data bases

SDL in Multi-Dimensional Frequency Tabulations

3-D Controlled Rounding

* Ernst (1989) showed that controlled roundings do not always exist in 3-D

* however, Ernst (1989) also
  - relaxed the adjacency constraint to within two multiples of the rounding base)
  - gave an exact algorithm for such “relaxed” rounding
* Ernst (1989) method:
  - stack the planes
  - bottom plane: zero-restricted controlled rounding
  - next and successive internal planes:
    # subtract sum of rounded values below from sum of original values up to current plane
    # apply zero-restricted CR
  - totals plane:
    # entries equal totals of roundings below
    # get usual CR automatically
* this method deserves consideration

Frechet Bounds

In a two-way table:

* LUB for each a_{ij} is maximum of its row and column sums r_i and c_j *(usual upper bound)*
* GLB is the Frechet Bound:max $\{0, r_i + c_j - t\}$(t = grand total)
* Frechet and usual bounds are *exact*

Counter examples

D	**D**	**0**	**D**	***9***
0	**D**	**D**	**0**	***3***
D	**0**	**D**	**0**	***6***
0	**D**	**0**	**D**	***4***
4	***10***	***3***	***5***	***22***

Example 0: Frechet and Usual Bounds Fail Under Zero-Restrictions

				1
				1
				0
				0
1	1	0	0	2

				1
				0
				1
				0
1	0	1	0	2

				0
				1
				0
				1
0	1	0	1	2

				0
				0
				1
				1
0	0	1	1	2

0	0	0	2	2
0	0	2	0	2
0	2	0	0	2
2	0	0	0	2
2	2	2	2	8

Example 1a: Infeasible 3-D Table

		0
		1
1	0	1

		1
		0
0	1	1

1	0	1
0	1	1
1	1	2

Example 1a: Infeasible 3-D Table

								2
								2
								0
								0
								2
								2
								0
								0
2	2	0	0	2	2	0	0	8

								2
								0
								2
								0
								2
								0
								2
								0
2	0	2	0	2	0	2	0	8

								0
								2
								0
								2
								0
								2
								0
								2
0	2	0	2	0	2	0	2	8

								0
								0
								2
								2
								0
								0
								2
								2
0	0	2	2	0	0	2	2	8

0	0	0	2	0	0	0	2	4
0	0	2	0	0	0	2	0	4
0	2	0	0	0	2	0	0	4
2	0	0	0	2	0	0	0	4
0	0	0	2	0	0	0	2	4
0	0	2	0	0	0	2	0	4
0	2	0	0	0	2	0	0	4
2	0	0	0	2	0	0	0	4
4	4	4	4	4	4	4	4	32

Example 2:
Infeasible Frechet Consistent
3-D Table

			3
			1
			1
3	1	1	5

			1
			3
			1
1	3	1	5

			1
			1
			3
1	1	3	5

1	1	3	5
3	1	1	5
1	3	1	5
5	5	5	15

Example 2a:
Infeasible Frechet Consistent
3-D Table

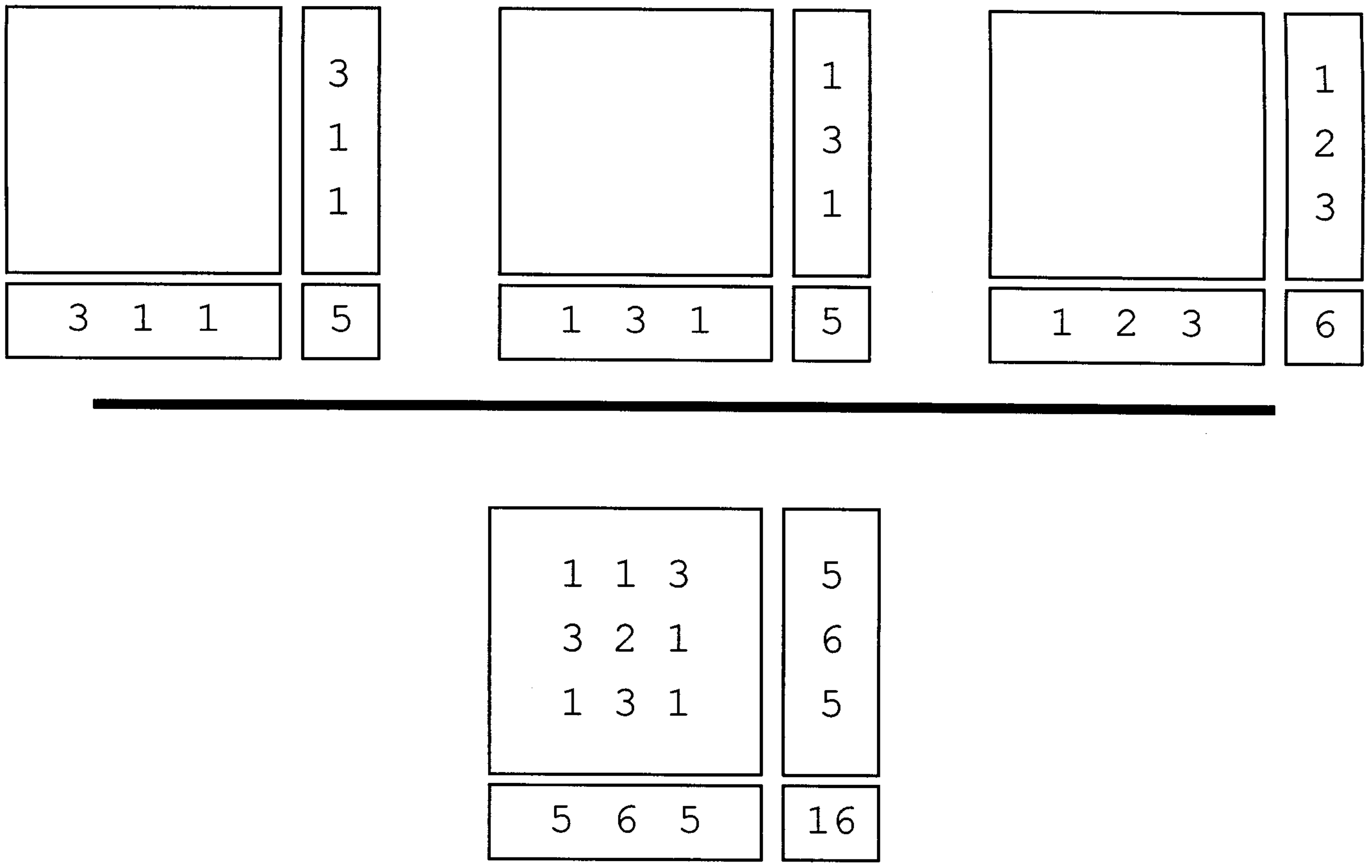
			3
			1
			1
3	1	1	5

			1
			3
			1
1	3	1	5

			1
			1
			3
1	1	3	5

1	2	2	5
2	1	2	5
2	2	1	5
5	5	5	15

Example 3: Feasible (unique)
3-D Table with Inexact Upper Bounds



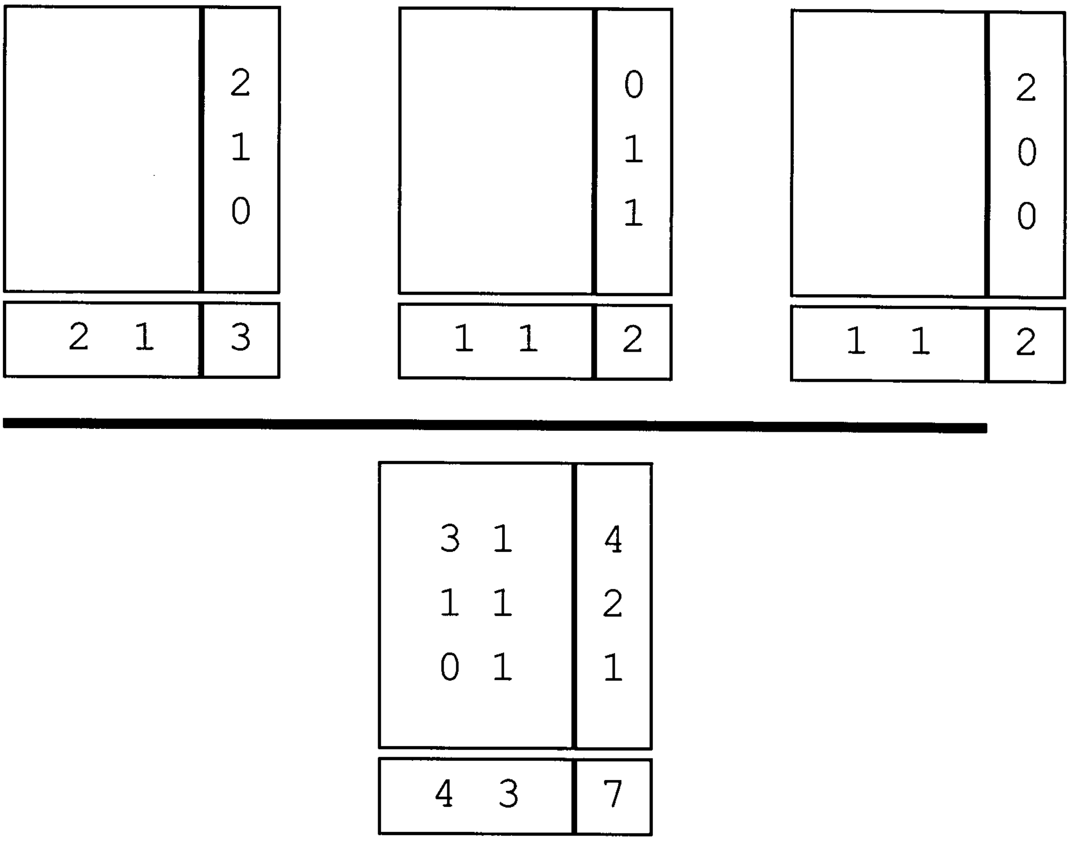
			3
			1
			1
3	1	1	5

			1
			3
			1
1	3	1	5

			1
			2
			3
1	2	3	6

1	1	3	5
3	2	1	6
1	3	1	5
5	6	5	16

Example 3a: Feasible (unique)
3-D Table with Inexact Upper Bounds



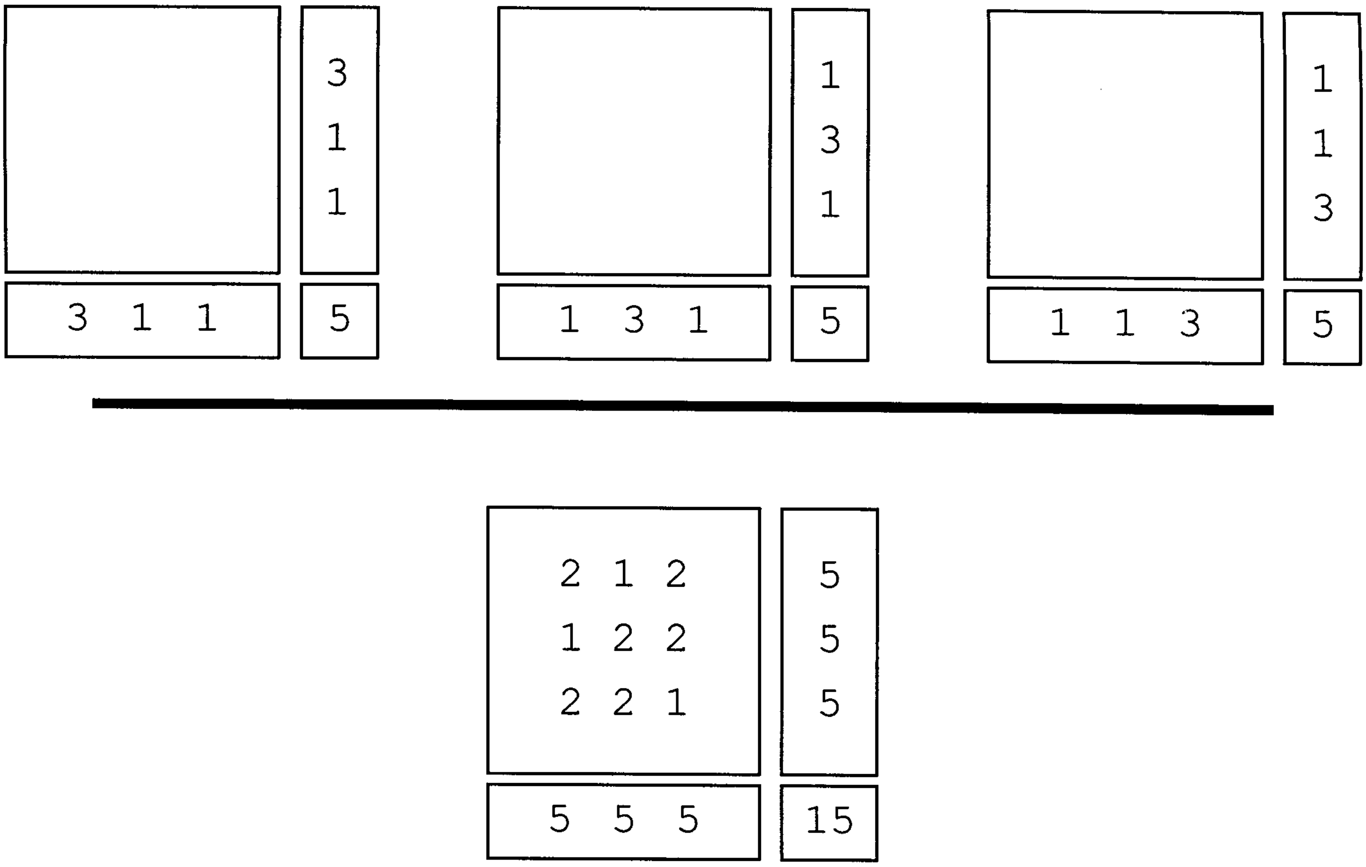
		2
		1
		0
2	1	3

		0
		1
		1
1	1	2

		2
		0
		0
1	1	2

3	1	4
1	1	2
0	1	1
4	3	7

Example 4: Feasible (unique) 3-D Table with Inexact Frechet Bounds



Example 5: Feasible 3-D Table with 3 df and 4 Integer Solutions

$$\begin{array}{|ccc|}\hline 0 & * & * \\ * & 0 & * \\ * & * & 0 \\ \hline\end{array}\qquad\begin{array}{|ccc|}\hline * & * & 0 \\ * & 0 & * \\ 0 & * & * \\ \hline\end{array}\qquad\begin{array}{|ccc|}\hline * & 0 & * \\ 0 & 0 & 0 \\ * & 0 & * \\ \hline\end{array}$$

Example 6: 3x3x3 Table with a Unique Cover But No Circuit

$$\begin{array}{|ccc|}\hline + & - & 0 \\ 0 & - & + \\ - & + & - \\ \hline\end{array}\qquad\begin{array}{|ccc|}\hline - & + & 0 \\ + & 0 & - \\ 0 & - & + \\ \hline\end{array}\qquad\begin{array}{|ccc|}\hline 0 & 0 & 0 \\ - & + & 0 \\ + & - & 0 \\ \hline\end{array}$$

Example 7: A Unique Odd Circuit

REFERENCES

Causey, B., L. Cox and L. Ernst (1985), “Applications of Transportation Theory to Statistical Problems”, *Journal of the American Statistical Association* **80**, 903-909.

Cox, L. (1980), “Suppression Methodology and Statistical Disclosure Control”, *Journal of the American Statistical Association* **75**, 377-385.

Cox, L. (1981), “Linear Sensitivity Measures in Statistical Disclosure Control”, *Journal of Statistical Planning and Inference* **5**, 153-164.

Cox, L. (1987), “A Constructive Procedure for Unbiased Controlled Rounding”, *Journal of the American Statistical Association* **82**, 520-524.

Cox, L. (1995), “Network Models for Complementary Cell Suppression”, *Journal of the American Statistical Association* **90**, 1453-1462.

Cox, L. and L. Ernst (1982), “Controlled Rounding”. *INFOR* 20, 423-432.

Cox, L., J. Fagan, B. Greenberg and R. Hemmig (1986), “Research at the Census Bureau into Disclosure Avoidance Techniques for Tabular Data”, *Proceedings of the Survey Research Methods Section, American Statistical Association*, Alexandria, VA, 388-393.

Cox, L. and J. George (1989), “Controlled Rounding for Tables with Subtotals”, *Annals of Operations Research* **20**, 141-157.

Federal Committee on Statistical Methodology (1994), **Report on Disclosure Limitation Methodology**, Statistical Policy Working Paper 22, Washington, DC: Office of Management and Budget.

Gusfield, D. (1988), “A Graph Theoretic Approach to Statistical Data Security”, *SIAM Journal on Computing* **17**, 552-571.

Kelly, J., B. Golden and A. Assad (1992), “Cell Suppression: Disclosure Protection for Sensitive Tabular Data”, *NETWORKS* **22**, 397-417.

Robertson, D. (1993), “Cell Suppression at Statistics Canada”, *Proceedings of the Annual Research Conference, U.S. Bureau of the Census*, Washington, DC: Department of Commerce, 107-131.

APPENDIX

Excerpts From a Disclosure Checklist

The U.S. Interagency Confidentiality and Data Access Group(ICDAG) has representatives from various U.S. statistical agencies. A major effort of ICDAG over recent years has been to develop a generic *checklist* of items for an agency to consider when preparing to release tabular data or a microdata file containing confidential information. Some excerpts from the Checklist relative to microdata release follow.

CHECKLIST ON DISCLOSURE POTENTIAL OF PROPOSED DATA RELEASES

Introduction

Federal statistical agencies and their contractors often collect data from persons, businesses, or other entities under a pledge of confidentiality. Before disseminating the results as either public-use **microdata files**[1] or tables, these agencies should apply statistical methods to protect the confidentiality of the information they collect. A review and evaluation of the statistical disclosure limitation techniques used by Federal statistical agencies can be found in the Federal Committee on Statistical Methodology's 1994 report, *Report on Statistical Disclosure Limitation Methodology* (Statistical Policy Working Paper [SPWP] # 22). In addition, SPWP # 22 contains a set of 12 recommendations to improve disclosure limitation practices.

One of the recommendations in SPWP # 22 is that agencies should centralize their review of disclosure-limited data products. In discussing this recommendation, SPWP # 22 suggests that if the number of programs is small, such a review could be handled by one individual; alternatively, if an agency has multiple or large programs, a review panel, team, or board might be needed. In this document, the term **Disclosure Review Board** is used to refer to the formally or informally designated unit or individual that handles such review. The attached document, "Checklist on Disclosure Potential of Proposed Data Releases" (called **Checklist**), is one tool that can assist agencies in reviewing disclosure-limited data products. Completed Checklists should be submitted to the Disclosure Review Board for review.

Most agency data products are intended for **public use**, with no restrictions on eligibility and intended use. Products that meet the criteria for public release may not have sufficient detail to satisfy the analytical requirements of all users. Consequently, some agencies have developed **restricted access** procedures for making more detailed microdata files and tables available to some users, subject to conditions of eligibility, location of use, purpose of use, security procedures, and other features associated with access to the data. *This Checklist is intended primarily for use in the development of public-use data products.* Some of the disclosure limitation procedures described in the Checklist may be of value in preparing data pro-

[1] A **microdata file** consists of records at the respondent level. Each record contains values of variables for a person, household, establishment, or other unit.

ducts for restricted access; however, the procedures may have to be relaxed to some degree to meet users' analytical requirements. The Interagency Confidentiality and Data Access Group (ICDAG) plans to develop additional documents (perhaps including another checklist) for use in developing arrangements for restricted releases of microdata files and tables. Pending availability of these documents, agencies may wish to consult a 1993 article by Jabine which summarizes restricted access procedures in use at that time.

The Checklist consists of a series of questions that are designed to assist an agency's Disclosure Review Board to determine the suitability of releasing either public-use microdata files or tables from data collected from individuals and/or organizations under an assurance of confidentiality. Section 4 pertains to microdata files that contain information from individuals or establishments, while Sections 5 and 6 refer to tabular data from individuals and establishments, respectively. This Checklist is based on one used at the U.S. Bureau of the Census. In creating its Checklist, ICDAG has liberally borrowed descriptions and definitions from SPWP # 22.

Uses of the Checklist

The Checklist was developed with following in mind:

- It should be completed by a person who has appropriate statistical knowledge and is familiar with the microdata file or tabular material in question (i.e., branch chief, survey manager, statistician, or programmer). While this implies a considerable familiarity with survey and statistical terminology, those without such background will nonetheless be able to understand much of what it intends to accomplish. (Those who need a "primer" on statistical disclosure limitation methods should see Chapter 2 of SPWP # 22. Other references can be found in Section 6 of this Checklist.)

- Responses to questions in the Checklist are not intended to supply all of the information required by a Disclosure Review Board before a microdata file or table is released to the public. Some additional questions may need to be answered and/or given special consideration. Nonetheless, if files and tabular material are reviewed with the aid of the Checklist early enough, the need for time-consuming and costly re-programming of the data to be released can be avoided. This allows additional time for coordination with collaborators and/or other potential users.

In addition to helping an agency's Disclosure Review Board determine the disclosure potential of proposed data releases, the Checklist has other uses:

- It can serve an important educational function for program staff who complete the Checklist.

- It can provide documentation when an agency is considering release of related data files and tabulations.

- It can be very useful in defending legal challenges to an agency's decision to withhold certain tabular data or restrict data contained on a public-use file.

The Checklist reflects the current standards of the Census Bureau and the National Center for Health Statistics for the release of public-use data. The Checklist is not a static document but a "work in progress" that will be changed, refined, and modified as new approaches and techniques are developed. With appropriate modifications, the Checklist can be adapted by Federal agencies and other organizations and used by them to review materials of varying levels of confidentiality. ICDAG encourages agencies to modify this document to suit their particular needs.

Brief Overview of Contents

- **Section 2. Cover Sheet:** This asks for basic information about the proposed data release.

- **Section 3: Microdata Files**

  Most microdata files contain data collected from persons or households (referred to as **demographic data**). *Some questions in this section may not be applicable for establishment-based files.*

  A major part of this section of the Checklist focuses on geographic information because it is the key factor in permitting inadvertent identification. In a demographic survey, few respondents could likely be identified within a single State, but more respondents — especially those with rare and visible reported characteristics — could be identified within a county or other geographic area with 100,000 or fewer persons.

  The risk of inadvertent disclosure is higher with a publicly released data set that has both detailed geographic variables and a detailed, extensive set of survey variables. The risk is also often a function of the quality and quantity of "auxiliary" information (data from sources external to the data being released). This auxiliary information may be difficult to assess for its disclosure risk. "Coarsening" a data set by dropping survey variables, collapsing response categories for other variables, and/or introducing statistical perturbation, called "noise", to the data are techniques that may reduce the risk of inadvertent disclosure (Kim and Winkler, 1995).

  For surveys of establishments, the issues are generally different because such entities are often selected from very skewed populations. For example, in the U.S., there are very few hospitals with 1,000 or more beds, and inadvertent disclosure in a survey of hospitals might be possible using detail on the number of beds and geographic information as large as a Census region.

- **Section 4: Tabular Data from Persons or Households (“Demographic Data”)**

  This section pertains to tables based on data collected from persons or households under a pledge of confidentiality. Tables can be of two types. Tables of **frequency count data** show the number in the population with certain characteristics or, equivalently, the percent of the population with certain characteristics. Tables of **magnitude data** present the aggregate of a “quantity of interest” over all units in the cell. Equivalently, the data may be presented as an average by dividing the aggregate by the number of units in the cell. Demographic data are typically reported as frequency count data.

  Section 4 of this Checklist should always be completed if the tabulations are based on a complete count or an enumeration of the target population. Its use should also be considered when:

  - the tabulations identify small geographic areas, e.g., areas with populations less than 100,000, or
  - a large sampling fraction was used, as in the case of the decennial census long-form sample, or
  - the tables have a large number of dimensions or cells, or
  - the tables cover especially sensitive topics.

- **Section 5: Tabular Data from Establishments or Other Types of Organizations**

  This section pertains to tabular data collected from organizations under a pledge of confidentiality. As with demographic data, tables can be of two types. Tables of **frequency count data** contain the number of units in a cell. Tables of **magnitude data** present the aggregate of a “quantity of interest” over all units in the cell. Thus, a table of the number of establishments within the manufacturing sector by industrial classification group is an example of the former, whereas a table that presents the total value of shipments for the same cells is an example of the latter. Different statistical disclosure limitation methods can be used depending on the type of data being presented, although, for practical purposes, entirely rigorous definitions are not necessary.

SECTION 2: COVER SHEET

SURVEY TITLE: ______________________________
DATE: ______________________________
Project Manager's Name: ______________________________
Division and/or Branch: ______________________________
Phone: ______________________________

1. **Is this survey sponsored/co-sponsored by another agency?**

   Yes. *Please list name(s) of agency(agencies).* ______________________

   ■ **No.**

2. **What type of data are you releasing?**

   ■ **Public-use microdata file.** *Please attach the proposed layout and content of the microdata file.*

   ■ **Tables.**

3. **When were the data collected?** ______________________

4. **Does(Do) the reference period(s) of the data collection differ from the actual date of collection?**

   ■ **Yes.** *Please give reference period(s).* ______________________

   ■ **No.**

5. **What is the periodicity of the proposed data release?**

   ■ **This is a preliminary release.**

   ■ **This is a one-time release of a public-use microdata file from a one-time collection of data.**

   ■ **This is a release of a special tabulation.**

   ■ **This is one in a series of releases (either microdata file or tables) with substantially the same content.** *Please specify the interval at which future products will be released or prior products have been released.* ______________________

   ■ **This is the re-release of an approved product, with the addition of supplemental or previously unreleased data.** *Please give the date the original product was submitted.*

   ____

   (NOTE. If this is a re-release of a previously approved product, then only complete those Checklist questions for which the answers are now different.)

6. **Will there be other data release(s) (either microdata files or tables) from this survey?**

   ■ **Yes.**

   ■ **No.**

Section 3. Microdata Files

3.1 Geographic Information on the File

List all geographic identifiers to be released and the minimum population of each identifier

General Rule: All geographic areas identified ***must*** have at least 100,000 persons in the sampled area.

3.1.1. Have you chosen to adopt the above rule or another?

■ Yes, will use the rule of 100,000.

■ No, will use other rule. Please specify and provide rationale

3.1.2. Records in many data bases are sequenced so that the first cases are in the lower numbered PSU or county that is first in alphabetic order.

3.1.3. Identify other geographically-related variables (e.g., center city, non-center city, metropolitan area, non-metropolitan area) on the file.

3.1.4. Sampling information may also provide some geographic indicators. For example, certain sampling weights may distinguish between self-representing and nonself-representing PSU's or identify types of areas intentionally oversampled. Also, codes for "second stage units", "hit number", etc., may be related to geography.
(a) List all sampling information — including that for variance estimation — that will be deleted for confidentiality reasons or subsampling plans to make weights less identifying
(b) List all other sampling information that you think might have geographic significance, but could not decide if it should be deleted

3.2 File Contents Presenting an Unusual Risk of Individual Disclosure

The disclosure criteria for public-use microdata require a review of each file to determine if any of the proposed contents present an unusual risk of individual disclosure. The Disclosure Review Board has identified several measures that can reduce the possibility of identifying an individual through the characteristics available on a file. The measures are discussed below, and relevant information pertaining to the proposed file is requested to assist the Disclosure Review Board in its review.

3.2.1. Names, addresses, and other unique numeric identifiers such as Social Security, Medicare, or Medicaid numbers ***must*** be removed from the file.

3.2.2. High income is a visible characteristic of individuals or households and is considered to be a sensitive item of information. Therefore, each income figure on the file, whether for households, persons, or families, including total income and its individual components, should be **topcoded**.

There are no hard and fast rules for determining which cutoffs to use in topcoding. Decisions should be based on examination of the structure of the distribution, in combination with other key variables like race, gender, etc. For example, one rule used at the Census Bureau is to topcode at least the top 1/2% of the non-zero values. Note that the strict use of the same criterion could result in changing the cutoff from year to year, which would make things very difficult for data users. One suggested solution would be to change the cutoff only when there has been a substantial change in the upper tail of the distribution. Before making such a change it is important to take into account how the proposed change will affect time series analyses.

Certain special cases require more thought when rules for topcoding are being developed. For example, consider a variable with a high proportion of zero values for most of the population (such as welfare income). As the proportion of non-zero values decreases, it may be desirable to topcode in such a manner that a higher proportion of them are above the cutoff. Be aware that a data base containing rare and unusual details on race and ethnicity may be a problem, unless there is little geographic detail. In addition, data bases that contain “unusual” subgroups may need special attention (for instance, high-income persons who pay no taxes). In developing topcode rules, it might be prudent to discuss alternatives with the Disclosure Review Board well in advance of the final submission for approval to release a file.

(a) Please describe the topcoding rule that is used. If you have different rules for different income variables, please give details.

(b) Do all income topcodes satisfy the appropriate rule(s)?

3.2.3. In addition to income, certain other characteristics may make an individual more visible than others. Some examples include: unusual occupation (as revealed by coding to 3 digits); unusual health condition (e.g., as shown in highly detailed International Classification of Disease codes); very high age; value or purchase price of own property; rent or amount of mortgage. Depending on the geographic detail shown on the file, consideration should be given to topcoding (and/or collapsing) these items when they are represented as interval or ordinal variables. One rule of thumb suggested by the Census Bureau’s Disclosure Review Board is that these topcode categories include at least _ of 1 percent of the total universe (persons/households) represented on the file (weighted counts).

In a few cases, where variables apply only to very small populations, the Disclosure Review Board may consider topcoding categories, including approximately 3 to 5 percent of the appropriate subpopulation. Approved topcodes at the Census Bureau include:

- Age — 85 years old and over. (Approximately 1.2% of all persons in the 1990 census.)
- Value of property — $500,000 or more. (Approximately 0.7% of all units, not just owneroccupied units in the 1990 census.)
- Gross Rent (including utilities) — $1,000 or more. Approximately 1.2% of all units, not just renter-occupied units in the 1990 census.)

- Payments on mortgages — $1,000/month (Approximately 3.0% of all mortgage holders on the 1984 Survey of Income and Program Participation file.)

  In addition, some variables may require bottom-coding, such as year of birth before 1914 or large negative value for income.

3.2.8. Describe any proposed information to be released for the bottom- or topcoded data items (for example, means or medians of the coded values):

3.2.9. Depending on the amount of geographic detail on the file, there are other characteristics that may make a person highly visible. These typically are represented as categorical or nonordinal variables and, therefore, cannot be topcoded . Some examples include the following: codes indicating Foreign or Indian Tribal language spoken; detailed racial identification such as Eskimo, Aleut, Guamian, or Samoan; detailed ethnic origins; codes for place of prior residence; codes for tenure in the area (“Always”, “Lifetime”). In these cases, the amount of detail on the file may have to be collapsed into larger categories.

3.2.10. Contextual or Ecologic Variables

Contextual or ecologic variables are those that describe some aspect of an area, such as a State, county, census tract, or block group; percent or frequency of the area’s population employed, foreign born, receiving public assistance; number of health facilities; number and specialty of physicians; local government expenditures; measures of air quality; etc.

3.3 Disclosure Risks with Administrative and Other External Data

Efforts must be made to reduce the potential for matching microdata on this file to data on external files because external files usually contain names and addresses and, thus, can be used to identify survey respondents. Such matching may be possible if the survey contains highly specific characteristics also found on mailing lists or administrative records maintained by other agencies or organizations. For example, the inclusion of vehicle make, model, and year in conjunction with specific geographic identifiers is unacceptable because these items can be matched to automobile registration lists that contain names and addresses. These items probably could be left on the file if they were recoded into broad categories. In addition to the external files mentioned above, other potential source of such files include: manufacturer’s list of purchasers of particular major durable goods (for example, airplanes); voter registration lists in some states; Federal, State, or local tax records; criminal justice system records; state hunting and fishing license registers; and membership rosters of certain trade associations.

Disclosure risk is also high if the sampling frame for a survey comes from a source outside the agency or if the file contains information obtained from other agencies. In such cases, the agency that provided the sampling frame or the auxiliary information may be able to match survey records to its original records, particularly if survey records include data from the originating agency’s files: e.g., amount of program benefit received, date of entry into program.

3.3.2.3. ...if longitudinal data are being collected; i.e., if the data for the same respondents/units will be collected for several different reference periods. Primary concern relates to time series of data items potentially matchable to outside records; e.g., income tax or employment records.

3.3.2.4. ...if highly specific geography is included on the file, such as State, Metropolitan statistical area, etc.

3.3.2.5. ...if data collected from multiple persons in a household are linked on the released file. Disclosure risks associated with linking of household members are well-known. For example, households can be identified because of significant difference in spouses' ages, atypical number and ages of children, a "unique" multi-racial composition of the household, etc. — not to mention the fact that one household member, by self-identification, could look up other members' reported information.

3.3.6. Cross-Tabulations To Identify Unique Sets of Characteristics

3.4. The Addition of Statistical Perturbation (or "Noise")

The addition of statistical perturbation, called "noise", is another statistical disclosure limitation technique. Essentially, "noise" is defined as the addition of small amounts of random variation to quantitative data. There are several methods that can be used to add noise to data.

3.4.1. Was any noise added to the data?

3.4.2. What procedure(s) was(were) used to add noise to the data? Please give specifics for that procedure (i.e., percent of records affected, distribution of noise, etc.). Some possibilities include the following:

- random noise
- record swapping
- rank swapping
- blanking and imputation

3.4.7. Was any attempt made to match back the noise-added data to the original file?