# Statistical Disclosure Limitation: Releasing Useful Data for Statistical Analysis

## Stephen E. Fienberg

**Department of Statistics**

**Center for Automated Learning & Discovery**

**Center for Computer & Communications Security**

**Carnegie Mellon University**

**Pittsburgh, PA, U.S.A.**

1

# Restricted Access vs. Releasing Restricted Data

- **Restricted Access:**
  - Special Sworn Employees.
  - Licensed Researchers.
  - External Sites.
  - Firewalls.
  - Query Control.

- **Releasing Restricted Data:**
  - Confidentiality motivates possible transformation of data before release.
  - Assess risk of disclosure and harm.
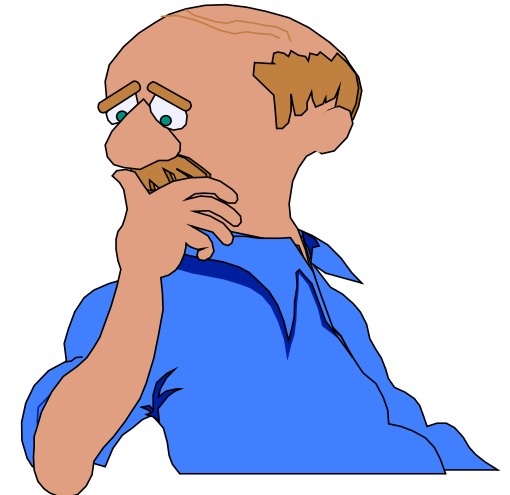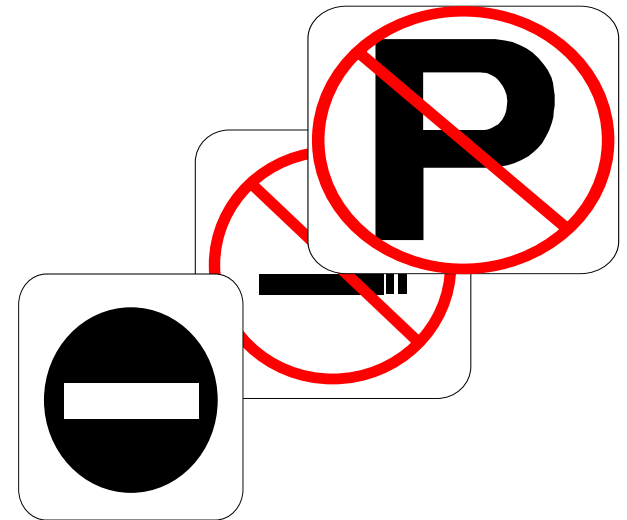
# Statistical Disclosure Limitation

- **What is goal of disclosure limitation?**
    - "Protecting" confidentiality.
    - Providing access to statistical data:
        - Statistical users want more than to retrieve a few numbers.
        - They want data useful for statistical analysis.

- **Statistical disclosure limitation needs to assess tradeoff between preserving confidentiality and usefulness of released data, especially for inferential purposes.**

# What Makes Released Data Statistically Useful?

- **Inferences should be the same as if we had original data.**
  - Reversing the disclosure protection mechanism, not for individual identification, but for inferences about parameters in statistical models (may require likelihood function for disclosure procedure).

- **Sufficient variables to allow for proper multivariate analyses.**

- **Ability to assess goodness of fit of models.**

# Examples of DL Methods

- **DL methods with problematic inferences:**
  - Cell suppression and related "interval" methods.
  - Data swapping without reported parameters.
  - Adding unreported amounts of noise.
  - Argus.
- **DL methods allowing for proper inferences:**
  - Post-randomization for key variables–PRAM.
  - Multiple imputation approaches.
  - Reporting data summaries (sufficient statistics) allowing for inferences AND assessment of fit.

File   Edit   View   Favorites   Tools   Help

Back   Search

www.eia.doe.gov
**Energy Information Administration**

# Avoiding Statistical "Swiss Cheese"

(Data released in May 2002; next release May 2003)

**Table 4. U.S. Uranium Mine Production and Number of Mines and Sources, 1992-2001**

| Mining Method | 1992 | 1993 | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 |
|---|---|---|---|---|---|---|---|---|---|---|
| Underground | | | | | | | | | | |
| (thousand pounds $U_3O_8$) | W | 0 | 0 | 0 | W | W | W | W | W | 0 |
| Openpit | | | | | | | | | | |
| (thousand pounds $U_3O_8$) | W | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| In Situ Leaching | | | | | | | | | | |
| (thousand pounds $U_3O_8$) | W | W | 2,448 | 3,372 | 4,379 | 4,084 | 3,721 | 3,830 | 2,995 | W |
| Other[a] | | | | | | | | | | |
| (thousand pounds $U_3O_8$) | 986 | 2,050 | 78 | 156 | 326 | 626 | 1,062 | 718 | 128 | W |
| **Total Mine Production** | | | | | | | | | | |
| (thousand pounds $U_3O_8$) | 986 | 2,050 | 2,526 | 3,528 | 4,705 | 4,710 | 4,782 | 4,548 | 3,123 | 2,647 |
| **Number of Mines Operated** | | | | | | | | | | |
| Underground | 4 | 0 | 0 | 0 | 1 | 1 | 4 | 3 | 1 | 0 |
| Openpit | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| In Situ Leaching | 4 | 5 | 5 | 5 | 6 | 7 | 6 | 6 | 4 | 3 |
| Other Sources[b] | 8 | 7 | 7 | 7 | 6 | 6 | 5 | 5 | 5 | 4 |
| **Total Mines and Sources** | 17 | 12 | 12 | 12 | 13 | 14 | 15 | 14 | 10 | 7 |

[a]For 1992, "Other" includes production from underground, openpit, and in situ leach mines and uranium bearing water from mine

Done                                                                                                     My Computer

Start   Microsoft PowerPoint - ...   Search Results   Help and Support Center   Uranium Industry A...   3:50 PM

file:///C:/Documents%20and%20Settings/fienberg/Desktop/BTS%20-%20Commodity%20Flow%20Survey%20-   Go   Search

Home   Bookmarks   Free AOL & U...   Instant Message   WebMail   Contact   People   Yellow Pages   Download   Find Sites   Channels

Top   Up   First   Previous   Next   Last   Document   More

**B T S**       Contents       Search       References       Contacts

**Bureau of Transportation Statistics**

# Commodity Flow Survey

CFS | Detailed Description | Methods & Limitations | Reports & Products | Future Plans | Applications

### State-to-State Commodity Flows: 1997

The following tables summarize data published by the U.S. Bureau of the Census in the individual state reports for the 1997 Commodity Flow Survey. See those reports for definitions of terms and data limitations.

**VALUE ($ mil) of Shipments by State of Origin/Destination.**

| Origin in row below, Destination in column to right | US | AL | AK | AZ | AR | CA | CO | CT | DE | DC | FL | GA | HI | ID | IL | IN | IA | KS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| US | 6943988 | 102491 | 12610 | 96362 | 71690 | 777276 | 88178 | 70339 | 20763 | 9191 | 305943 | 242954 | 18208 | 21626 | 349291 | 178649 | 91335 | 68301 |
| AL | 101547 | 34369 | 5 | 274 | 1312 | 2734 | S | 310 | 73 | 17 | 5363 | 8751 | S | 109 | 2456 | 1742 | 413 | 492 |
| AK | 6653 | S | 5376 | S | S | S | S | S | S | S | S | S | S | S | 1 | S | S | S |
| AZ | 86256 | S | 22 | 32386 | 287 | 14616 | 738 | S | 10 | 6 | 1613 | 670 | S | S | S | 826 | 262 | 253 |
| AR | 71670 | 784 | 7 | 766 | 25256 | 2919 | 458 | 315 | 63 | 21 | 1149 | 1510 | 3 | 63 | 2636 | 1113 | 610 | 779 |
| CA | 802192 | 3000 | 764 | 20425 | 3730 | 489246 | 8803 | 2630 | 713 | 526 | 17755 | 10893 | 3729 | 2018 | 13073 | 4511 | 1581 | 3637 |

Start       Microsoft PowerPoint - [...]       BTS - Commodity Flo...       11:35 PM

# Overview

- **Background and some fundamental abstractions for disclosure limitation.**

- **Methods for tables of counts:**
  - **Results on bounds for table entries.**
  - **Uses of Markov bases for exact distributions and perturbation of tables.**
  - <span style="color:darkred">**Links to log-linear models, and related statistical theory and methods.**</span>

- <span style="color:darkred">**Some general principles for developing new methods.**</span>

# R-U Confidentiality Map
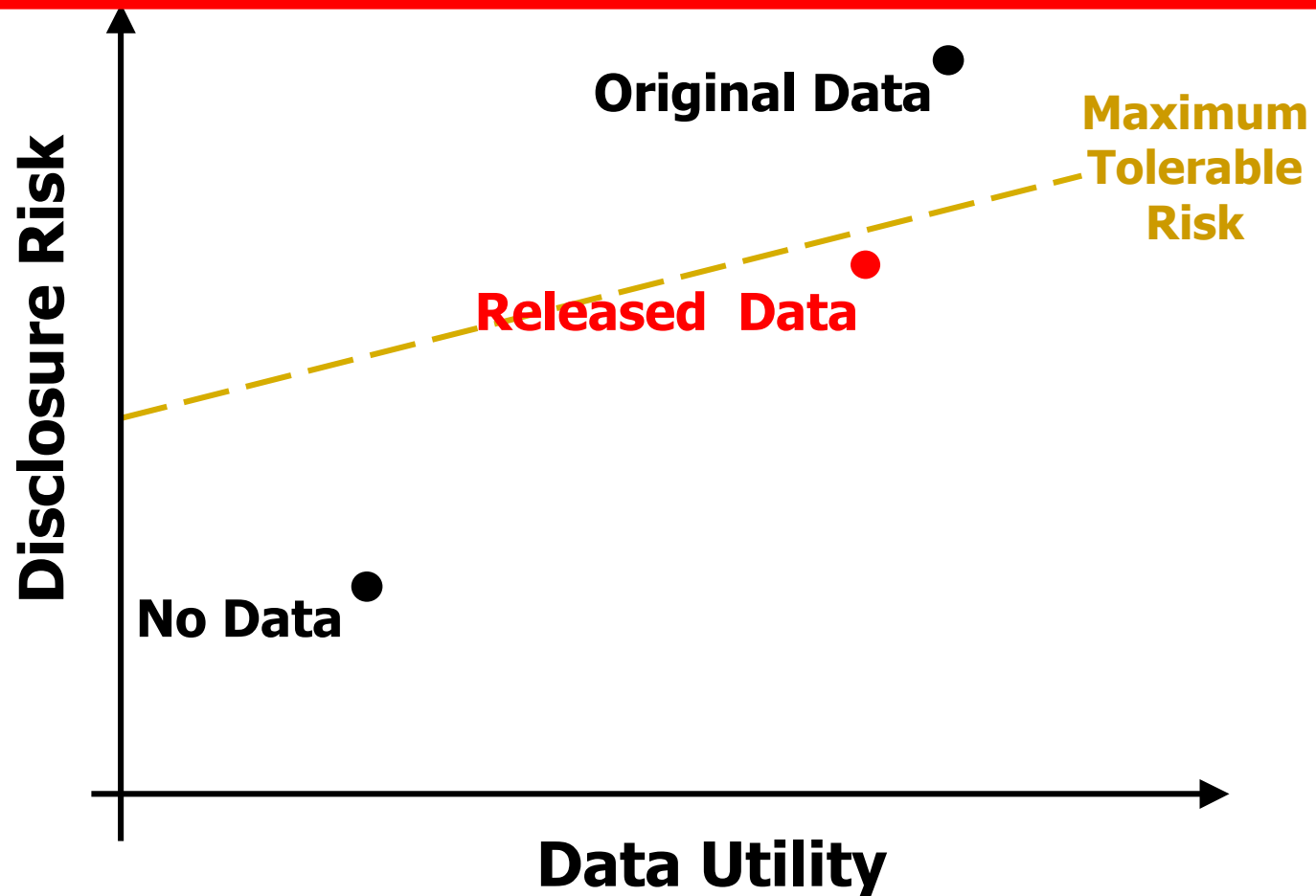


(Duncan, et al. 2001)

# NISS Prototype Query System

- For *k*-way table of counts.
- *Queries*: Requests for marginal tables.
- *Responses*: Yes--release; No; (and perhaps "Simulate" and then release).
- As released margins cumulate we have increased information about table entries.
- Margins need to be consistent ==> possible simulated releases get highly constrained.

# Confidentiality Concern

- **Uniqueness in population table $\Leftrightarrow$ cell count of "1".**
  - Uniqueness allows intruder to match characteristics in table with other data bases **that include same variables** to learn confidential information.
  - **Assuming data are reported without error!**

- **Identity versus attribute disclosure.**

- **Sample vs. population tables:**
  - **Identifying who is in CPS and other sample surveys.**

# Fundamental Abstractions

- **Query space, Q, with partial ordering:**
  - Elements can be marginal tables, conditionals, *k*-groupings, regressions, or other data summaries.

  - *Released set:* R(*t*), and implied *Unreleasable set:* U(*t*).

  - *Releasable frontier:* maximal elements of R(*t*).

  - *Unreleasable frontier:* minimal elements of U(*t*).

- *Risk* and *Utility* defined on subsets of Q.

  - *Risk Measure*: identifiability of small cell counts.

  - *Utility*: reconstructing table using log-linear models.

  - Release rules must balance risk and utility:
    - R-U Confidentiality map.
    - General Bayesian decision-theoretic approach.

# Why Marginals?

- **Simple summaries corresponding to subsets of variables.**

- **Traditional mode of reporting for statistical agencies and others.**

- **Useful in statistical modeling: Role of log-linear models.**

- **Collapsing categories of categorical variables uses similar DL methods and statistical theory.**

# Example 1: 2000 Census

- **U.S. decennial census "long form"**
  - 1 in 6 sample of households nationwide.
  - 53 questions, many with multiple categories.
  - <span style="color:darkred">**Data measured with substantial error!**</span>
  - <span style="color:darkred">**Data reported after application of data swapping!**</span>
- **Geography**
  - 50 states; 3,000 counties; 4 million "blocks".
  - Release of detailed geography yields uniqueness in sample and at some level in population.
- *American Factfinder*  releases various 3-way tables at different levels of geography.

14

Back | Forward | Stop | Refresh | Home | Search | Favorites | History | Mail | Print | Edit | Discuss

Address  A:\AFF.HTM

Links »

# U.S. Census Bureau

**American FactFinder**

Main | Search | Feedback | FAQs | Help

*Your source for population, housing, economic and geographic data.*

## Search

- ⦿ keyword
- ○ place name

[   ] **Go**

Enter a **street address** to find Census 2000 data

| Site Tour |
| What's in FactFinder |
| Confidentiality |
| What's New |
| Products |
| Reference Maps |
| Thematic Maps |
| Data Sets |

Censo 2000 Puerto Rico en español

**Kids'** Corner

Browser Note...

### Start with Basic Facts

⦿ Tables          ○ Maps

**Show me**   [ Population and Housing (QT)  ▼ ]

**for**   [ a State  ▼ ]

[ --select a state--  ▼ ]   **Go**

Where is my state?

Census 2000 Release Schedule   |   American Community Survey Data

### Items of Interest                                        more...

**Census 2000 Supplementary Survey**
US and state estimates on income, poverty, education, and other topics are available in this new product release

**Summary File 1**
Tables on age, sex, households, families, and housing for smaller areas are now being released on a state-by-state basis

United States Census 2000

**Housing Unit Counts**
Housing unit counts for states, counties, places, and more

**Demographic Profiles**
Age, sex, race, and Hispanic/Latino counts, and information on housing, households and families

**Rankings, Comparisons, and Summaries**
Rankings and 1990-2000 population change for the United States, states, counties, metropolitan areas, and large places

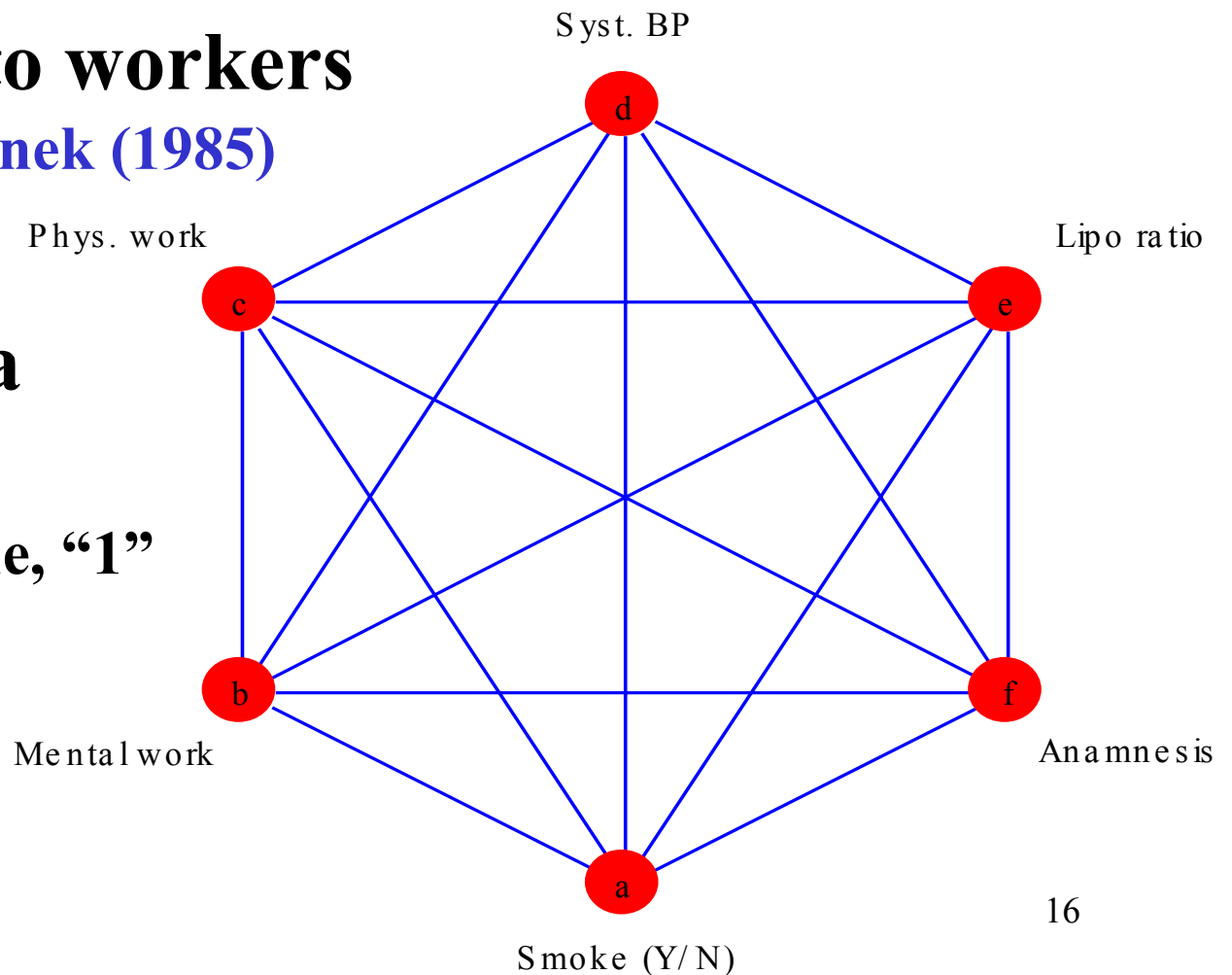▶ United States total population: **281,421,906**  (April 1, 2000)

My Computer

# Example 2: Risk Factors for Coronary Heart Disease

- **1841 Czech auto workers**
  **Edwards and Havanek (1985)**

- **$2^6$ table**

- **population data**
  - **"0" cell**
  - **population unique, "1"**
  - **2 cells with "2"**

Syst. BP

Lipo ratio

Phys. work

Anamnesis

Mental work

Smoke (Y/N)

16

# Example 2: The Data

| F | E | D | C | B no | no yes | yes no | yes |
|---|---|---|---|---|---|---|---|
|   |   |   | A | no | yes | no | yes |
| neg | < 3 | < 140 | no | 44 | 40 | 112 | 67 |
|   |   |   | yes | 129 | 145 | 12 | 23 |
|   |   | ≥ 140 | no | 35 | 12 | 80 | 33 |
|   |   |   | yes | 109 | 67 | 7 | 9 |
|   | ≥ 3 | < 140 | no | 23 | 32 | 70 | 66 |
|   |   |   | yes | 50 | 80 | 7 | 13 |
|   |   | ≥ 140 | no | 24 | 25 | 73 | 57 |
|   |   |   | yes | 51 | 63 | 7 | 16 |
| pos | < 3 | < 140 | no | 5 | 7 | 21 | 9 |
|   |   |   | yes | 9 | 17 | 1 | 4 |
|   |   | ≥ 140 | no | 4 | 3 | 11 | 8 |
|   |   |   | yes | 14 | 17 | 5 | 2 |
|   | ≥ 3 | < 140 | no | 7 | 3 | 14 | 14 |
|   |   |   | yes | 9 | 16 | 2 | 3 |
|   |   | ≥ 140 | no | 4 | 0 | 13 | 11 |
|   |   |   | yes | 5 | 14 | 4 | 4 |

# Example 3: NLTCS

- **National Long Term Care Survey**
  - **20-40 demographic/background items.**
  - **30-50 items on disability status, ADLs and IADLs, most binary but some polytomous.**
  - **Linked Medicare files.**
  - **5 waves: 1982, 1984, 1989, 1994, 1999.**

- **We've been working with $2^{16}$ table, collapsed across several waves of survey, with $n$=21,574.**

**Erosheva (2002)**

**Dobra, Erosheva, & Fienberg (2003)**

# Two-Way Fréchet Bounds

- **For 2×2 tables of counts$\{n_{ij}\}$ given the marginal totals $\{n_{1+},n_{2+}\}$ and $\{n_{+1},n_{+2}\}$:**

$$
\begin{array}{cc|c}
n_{11} & n_{12} & n_{1+} \\
n_{21} & n_{22} & n_{2+} \\
\hline
n_{+1} & n_{+2} & n
\end{array}
$$

$$\min(n_{i+},n_{+j}) \geq n_{ij} \geq \max(n_{i+} + n_{+j} - n, 0)$$

- **Interested in multi-way generalizations involving higher-order, overlapping margins.**

# Bounds for Multi-Way Tables

- ***k*-way table of non-negative counts, *k* ≥ 3.**
  - Release set of marginal totals, possibly overlapping.
  - *Goal*: Compute bounds for cell entries.
  - LP and IP approaches are NP-hard.

- **Our strategy has been to:**
  - Develop efficient methods for several special cases.
  - Exploit linkage to statistical theory where possible.
  - Use general, less efficient methods for residual cases.

- **Direct generalizations to tables with non-integer, non-negative entries.**

# Role of Log-linear Models?

- **For 2×2 case, lower bound is evocative of MLE for estimated expected value under independence:**
$$\hat{m}_{ij} = n_{i+} n_{+j} / n.$$

  - Bounds correspond to log-linearized version.
  - Margins are *minimal sufficient statistics (MSS)*.

- **In *3*-way table of counts, $\{n_{ijk}\}$, we model logs of expectations $\{E(n_{ijk})=m_{ijk}\}$:**

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)}$$

- *MSS* are margins corresponding to highest order terms: $\{n_{ij+}\}$, $\{n_{i+k}\}$, $\{n_{+jk}\}$.

# Graphical & Decomposable Log-linear Models

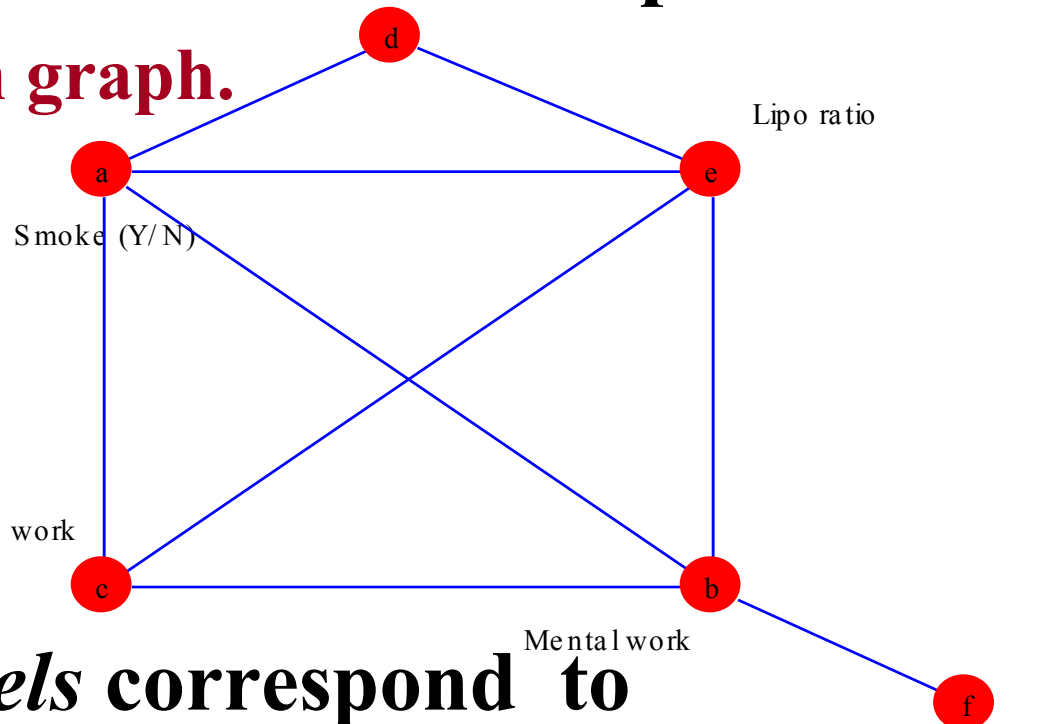- *Graphical models:* **defined by simultaneous conditional independence relationships**
  - **Absence of edges in graph.**

**Example 2:**

**Czech autoworkers**

**Graph has 3 cliques:**

**[ADE][ABCE][BF]**

- *Decomposable models* **correspond to triangulated graphs.**

Syst. BP

d

Lipo ratio

a

e

Smoke (Y/N)

Phys. work

c

b

Mental work

f

Ana

# MLEs for Decomposable Log-linear Models

- **For decomposable models, expected cell values are explicit function of margins, corresponding to MSSs (*cliques* in graph):**
  - For conditional independence in 3-way table:

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)}$$

$$m_{ijk} = \frac{m_{ij+} m_{i+k}}{m_{i++}}$$

- **Substitute observed margins for expected in explicit formula to get MLEs.**

# Multi-way Bounds

- **For decomposable log-linear models:**

$$\text{Expected Value} = \frac{\prod MSSs}{\prod Separators}$$

- *Theorem:* **When released margins correspond to those of a decomposable model:**

  - *Upper bound:* **minimum of relevant margins.**
  - *Lower bound:* **maximum of zero, or sum of relevant margins minus separators.**
  - **Bounds are sharp.**

**Fienberg and Dobra (2000)**

24

# Multi-Way Bounds (cont.)

- *Example*: Given margins in *k*-way table that correspond to (*k*-1)-fold conditional independence given variable 1:

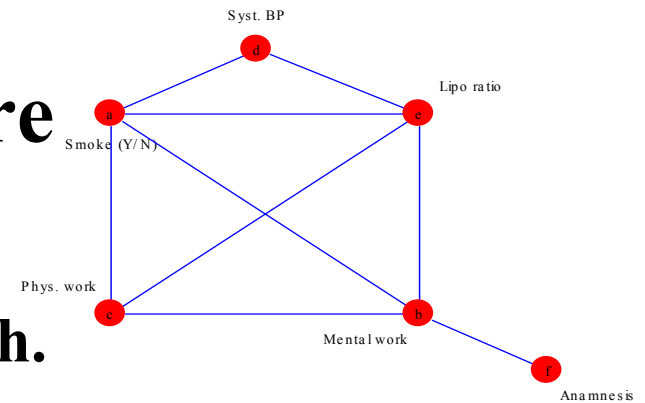$$\{n_{i_1 i_2 + \ldots +}\} \ \{n_{i_1 + i_3 \ldots +}\} \ldots \{n_{i_1 + \ldots + i_k}\}$$

- Then bounds are

$$\min\{n_{i_1 i_2 + \ldots + +}, n_{i_1 + i_3 \ldots + +}, \ldots, n_{i_1 + \ldots + i_k}\} \geq n_{i_1 i_2 i_3 \ldots i_k}$$

$$\geq \max\{n_{i_1 i_2 + \ldots + +} + n_{i_1 + i_3 \ldots + +} + \ldots + n_{i_1 + \ldots + i_k} - n_{i_3 + + \ldots + +}(k - 2), 0\}$$

# Ex. 2:  Czech Autoworkers

- **Suppose released margins are**

  **[ADE][ABCE][BF] :**

  - Correspond to decomposable graph.
  - Cell containing population unique has bounds [0, 25].
  - Cells with entry of "2" have bounds: [0,20] and [0,38].
  - Lower bounds are all "0".

- **"Safe" to release these margins; low risk of disclosure.**

# Bounds for [BF][ABCE][ADE]

| F | E | D | C | B no | | B yes | |
|---|---|---|---|---|---|---|---|
| | | | A | no | yes | no | yes |
| neg | < 3 | < 140 | no | [0,88] | [0,62] | [0,224] | [0,117] |
| | | | yes | [0,261] | [0,246] | [0,25] | [0,38] |
| | | ≥ 140 | no | [0,88] | [0,62] | [0,224] | [0,117] |
| | | | yes | [0,261] | [0,151] | [0,25] | [0,38] |
| | ≥ 3 | < 140 | no | [0,58] | [0,60] | [0,170] | [0,148] |
| | | | yes | [0,115] | [0,173] | [0,20] | [0,36] |
| | | ≥ 140 | no | [0,58] | [0,60] | [0,170] | [0,148] |
| | | | yes | [0,115] | [0,173] | [0,20] | [0,36] |
| pos | < 3 | < 140 | no | [0,88] | [0,62] | [0,126] | [0,117] |
| | | | yes | [0,134] | [0,134] | [0,25] | [0,38] |
| | | ≥ 140 | no | [0,88] | [0,62] | [0,126] | [0,117] |
| | | | yes | [0,134] | [0,134] | [0,25] | [0,38] |
| | ≥ 3 | < 140 | no | [0,58] | [0,60] | [0,126] | [0,126] |
| | | | yes | [0,115] | [0,134] | [0,20] | [0,36] |
| | | ≥ 140 | no | [0,58] | [0,60] | [0,126] | [0,126] |
| | | | yes | [0,115] | [0,134] | [0,20] | [0,36] |

Table 1 - Bounds for Autoworkers data given the marginals [BF], [ABCE], [ADE].      27

# Example 2 (cont.)

- **Among all 32,000+ decomposable models, the tightest possible bounds for three target cells are: (0,3), (0,6), (0,3).**
  - **31 models with these bounds! All involve [ACDEF].**
  - **Another 30 models have bounds that differ by 5 or less (*critical width*) and these involve [ABCDE].**
  - **Method used to search for "optimal" decomposable release also identifies [ABDEF] as potentially problematic.**
- **Allows proper statistical test of fit for most interesting models.**

# More on Bounds

- **Extension for log-linear models and margins corresponding to reducible graphs.**

- **For $2^k$ tables with ($k$-1) dimensional margins fixed (need one extra bound here and it comes from log-linear model theory: existence of MLEs).**

  - **Extend to general $k$-way case by looking at all possible collapsed $2^k$ tables.**

- **General "shuttle" algorithm in Dobra (2002) works for all cases but computationally intensive:**

  - **Also generates most special cases with limited extra computation.**

# Example 2: Release of All 5-way Margins

- **Approach for $2 \times 2 \times 2$ generalizes to $2^k$ table given (k-1)-way margins.**

- **In $2^6$ table, if we release all 5-way margins:**

  - Almost identical upper and lower values; they all differ by 1.

  - Only 2 feasible tables with these margins!

- **UNSAFE!**

# Example 2: Making Proper Statistical Inferences

- **In Example 2, we know we can't release [ABCDE] and [ACDEF].**

- **Suppose we deem release of everything else to be safe, i.e., we release [ACDE] [ABCDF][ABCEF][BCDEF][ABDEF] and we announce that users can make correct inference from release.**

- **What can user and intruder do?**

# Example 2: Making Proper Statistical Inferences (cont.)

- **Includes among models that can be fitted our "favorite"one: [ADE][ABCE][BF].**

- **Can do proper log-linear inferences using MLE and variation of chi-square tests based on expected values from model linked to released marginals.**

- **Announcement that releases can be used for proper inference will not materially reduce space of possible tables for intruder's inferences.**

# Example 3: NLTCS

- **$2^{16}$ table of ADL/IADLs with 65,536 cells:**
  - 62,384 zero entries; 1,729 cells with count of "1" and 499 cells with count of "2".
  - *n*=21,574.
  - Largest cell count: 3,853—no disabilities.

- **Used simulated annealing algorithm to search all decomposable models for "decomposable" model on frontier with**

  **max[upper bound – lower bound] >3.**

- **Acting *as if* these were *population* data.**

33

# NLTCS Search Results

- **Decomposable frontier model:**

  **{[1,2,3,4,5,7,12], [1,2,3,6,7,12], [2,3,4,5,7,8],**

  **[1,2,4,5,7,11], [2,3,4,5,7,13], [3,4,5,7,9,13],**

  **[2,3,4,5,13,14], [2,4,5,10,13,14], [1,2,3,4,5,15],**

  **[2,3,4,5,8,16]}.**

- **Has one 7-way and eight 6-way marginals.**

# Sparseness in NLTCS Data

- **Sparseness of table in this example extends to margins we might want to release, e.g., $2^{10}$ table of ADLs and $2^6$ table of IADLs:**

  - We need to alter margins to allow for release.

- **Perturbation of table subject to marginal constraints for already-released margins:**

  - Part of framework for NISS prototype.

# Perturbation Maintaining Marginal Totals

|  | $w_1$ | $w_2$ | $w_3$ | $w_4$ |
|---|---|---|---|---|
| $v_1$ | +1 | 0 | −1 | 0 |
| $v_2$ | −1 | 0 | +1 | 0 |
| $v_3$ | 0 | 0 | 0 | 0 |
| $v_4$ | 0 | 0 | 0 | 0 |

- **Perturbation distributions given marginals require Markov basis for perturbation moves.**

# Exact Distribution of Table Given Marginals

- **Exact probability distribution for log-linear model given its MSS marginals:**

$$\sigma(\mathbf{n}) = \frac{\prod_{i \in I} \dfrac{1}{n(i)!}}{\sum_{m \in S(c)} \left( \prod_{i \in I} \dfrac{1}{m(i)!} \right)}$$

  – Can generate distribution using **Diaconis-Sturmfels (1998)** MCMC approach using Markov basis.

  **Fienberg, Makov, Meyer, Steele (2002)**

# Markov Basis "Moves"

- ## Simple moves:
  - Based on standard linear contrasts involving 1's, 0's, and -1's for embedded $2^l$ subtables.
  - For example, in 2×2×2 table, there is 1 move of form:

| 1 | -1 | | -1 | 1 |
|---|----|----|----|---|
| -1 | 1 | | 1 | -1 |

- ## "Non-simple" moves:
  - Require combination of simple moves to reach extremal tables in convex polytope.

# Perturbation for Protection

- **Perturbation preserving marginals involves a parallel set of results to those for bounds:**

  - Markov basis elements for decomposable case requires only "simple" moves. **(Dobra, 2002)**

  - Efficient generation of Markov basis for reducible case. **(Dobra and Sullivent, 2002)**

  - Simplifications for $2^k$ tables ("binomials").

  - Rooted in ideas from likelihood theory for log-linear models and computational algebra of toric ideals.

# Some Ongoing Research

- **Queries in form of combinations of marginals and conditionals.**

- **Inferences from marginal releases.**

- **What information does the intruder really have?**

- **Record linkage and matching.**

- **Simplified cyclic perturbation distributions.**

# Some General Principles for Developing DL Methods

- **All data are informative for intruder including, non-release or suppression.**

- **Need to define and understand potential statistical uses of data in advance:**

  - **Leads to useful reportable summaries.**

- **Methods should allow for reversibility for inference purposes:**

  - **Missing data should be "ignorable" for inferences.**
  - **Assessing goodness of fit is important.**

# Where Will Tools Come From?

- Statistical methods and theory and modern datamining methods.

- Optimization approaches from OR.

- New mathematics, e.g., computational algebraic geometry.

# Summary

- **Presented some fundamental abstractions for disclosure limitation.**

- **Illustrated what I refer to as statistical approach to DL using tables of counts.**

  - New theoretical links among disclosure limitation, statistical theory, and computational algebraic geometry.

- **Articulates some general principles for developing DL methods.**

# The End

- **Most papers available for downloading at**

    **http://www.niss.org**

    **http://www.stat.cmu.edu/~fienberg/disclosure.html**

**Workshop on Computational Algebraic Statistics**

**December 14 to 18, 2003**

**American Institute of Mathematics**

**Palo Alto, California**

**http://aimath.org/ARCC/workshops/compalgstat.html**

# Three-way Illustration (*k*=3)
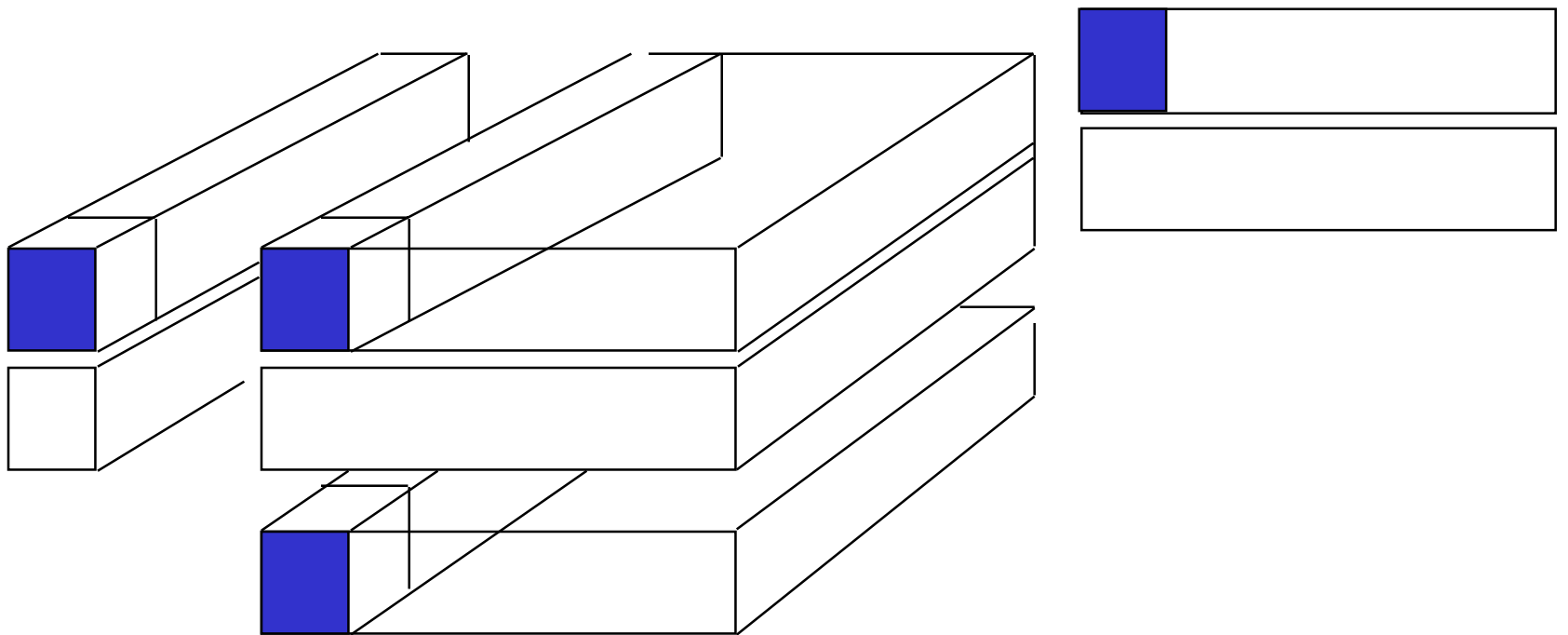


*Challenge:* Scaling up approach for large *k.*

# Existence of MLEs for 2×2×2 Table

| | | |
|---|---|---|
| $0$ | $n_{121}$ | $n_{1+1}$ |
| $n_{211}$ | $n_{221}$ | $n_{2+1}$ |
| $n_{+11}$ | $n_{+21}$ | $n_{++1}$ |

| | | |
|---|---|---|
| $n_{112}$ | $n_{122}$ | $n_{1+2}$ |
| $n_{212}$ | $0$ | $n_{2+2}$ |
| $n_{+12}$ | $n_{+22}$ | $n_{++2}$ |

| | |
|---|---|
| $n_{11+}$ | $n_{12+}$ |
| $n_{21+}$ | $n_{22+}$ |

- Require all estimated expected cell values to be positive.

46

# Existence of MLEs for 2×2×2 Table

| $0+\delta$ | $n_{121}-\delta$ | $n_{1+1}$ | | $n_{112}-\delta$ | $n_{122}+\delta$ | $n_{1+2}$ |
|---|---|---|---|---|---|---|
| $n_{211}-\delta$ | $n_{221}+\delta$ | $n_{2+1}$ | | $n_{212}+\delta$ | $0-\delta$ | $n_{2+2}$ |
| $n_{+11}$ | $n_{+21}$ | $n_{++1}$ | | $n_{+12}$ | $n_{+22}$ | $n_{++2}$ |

| | |
|---|---|
| $n_{11+}$ | $n_{12+}$ |
| $n_{21+}$ | $n_{22+}$ |

$\delta$ must be zero and MLE doesn't exist.

# $2^3$ Table Given $2\times2$ Margins

| $n_{111}$ | $n_{121}$ | $n_{1+1}$ |     | $n_{112}$ | $n_{122}$ | $n_{1+2}$ |
|-----------|-----------|-----------|-----|-----------|-----------|-----------|
| $n_{211}$ | $n_{221}$ | $n_{2+1}$ |     | $n_{212}$ | $n_{222}$ | $n_{2+2}$ |
| $n_{+11}$ | $n_{+21}$ | $n_{++1}$ |     | $n_{+12}$ | $n_{+22}$ | $n_{++2}$ |

| | $n_{11+}$ | $n_{12+}$ | |
|---|-----------|-----------|---|
| | $n_{21+}$ | $n_{22+}$ | |

- Obvious upper and lower bounds for $n_{111}$
- Extra upper bound: $n_{111} + n_{222}$

48

# NISS Table Server: 6-Way Table