

# **The Confidentiality-Data Access Tradeoff: Making Protected Data Statistically Useful**

**Stephen E. Fienberg**  
**CREST, INSEE, Paris, France**

**And**

**Department of Statistics**  
**Center for Automated Learning & Discovery**  
**Center for Computer & Communications Security**  
**Carnegie Mellon University**  
**Pittsburgh, PA, U.S.A.**

# Overview

**What makes data useful for statistical analysis?**

- **Methods for tables of counts:**
  - Results on bounds for table entries.
  - Links to log-linear models, and related statistical theory and methods.
  - Related results on perturbation methods.
- **Some general principles for developing new methods.**

# Statistical Disclosure Limitation

- **What is goal of disclosure limitation?**
- **Statistical disclosure limitation needs to assess **tradeoff** between**
  - **Preserving confidentiality and**
  - **Ensuring usefulness of released data, especially for inferential purposes.**
- **Statistical users want more than retrieval a few numbers—data need to be useful for statistical analysis.**

# **Brief History of Confidentiality in U.S.**

- **Concern arose in early 20th century.**
- **Statement by President Taft in 1910 re protecting census data from other government agencies.**
- **Census Bureau focused initially on protection for businesses.**
- **Title 13 of US Code enacted in 1929 protected individual data as well.**
  - **Resolution of language on confidentiality in 1960s.**
- **U.S. Patriot Act of 2001.**

# What Makes Released Data Statistically Useful?

- Inferences should be the same as if we had original complete data.
  - Requires ability to reverse disclosure protection mechanism, not for individual identification, but for inferences about parameters in statistical models (e.g, likelihood function for disclosure procedure).
- Sufficient variables to allow for proper multivariate analyses.
- Ability to assess goodness of fit of models.
  - Need most summary information, residuals, etc.

# 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. (Rubin, Abowd)
  - Reporting data summaries (sufficient statistics) allowing for inferences AND assessment of fit.

# Example 1: NLTCs

- **National Long Term Care Survey**
  - National Institute on Aging funding to Duke Univ.
  - Main interviews conducted by Census Bureau.
    - 20-40 demographic/background items.
    - More than 30 items on disability status, ADLs and IADLs, most binary but some polytomous.
    - Linked Medicare files.
  - 5 waves: 1982, 1984, 1989, 1994, 1999.  
[Erosheva \(2002\); Dobra, Erosheva, & Fienberg \(2003\)](#)
- **Access via Duke Center for Demographic Studies.**

# Example 2: Avoiding Statistical “Swiss Cheese”

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

# Disclosure Limitation for Tabular Count Data

- Large sparse multi-dimensional tables.
- 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.
- Risk concerned with small cell counts.
- Utility typically tied to usefulness of marginal totals:
  - **Other types of sensible summary statistics!**

# Example 3: 2000 Census

- U.S. decennial census “long form”
  - 1 in 6 sample of households nationwide.
  - 53 questions, many with multiple categories.
  - **Data reported after application of data swapping!**
- 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.

### Search

- keyword
- place name


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...](#)

Your source for population, housing, economic and geographic data

## start with Basic Facts

Tables

	Males	Females
Under 1 year	40,457	37,754
1 and 2 years	115,728	113,145
3 and 4 years	115,134	111,963
5 years	95,278	113,845
5 years	67,528	113,652

Maps



Show me

for

[Where is my state?](#)

[Census 2000 Release Schedule](#)

[American Community Survey](#)

### Items of Interest

#### [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, families, and housing areas are now being released on a state-by-state basis



#### [Housing Unit Counts](#)

Housing unit counts for states, counties, places, and more

#### [Demographic Profiles](#)

Age, sex, race, and counts, and information on households and families

#### [Rankings, Comparisons, and Summaries](#)

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

▶ United States total population: [281,421,906](#) (April 1, 2000)



# 2000 Census Block Data

- **2000 Census Data on Sex, Age, and Race for Pittsburgh Block extracted from AFF**

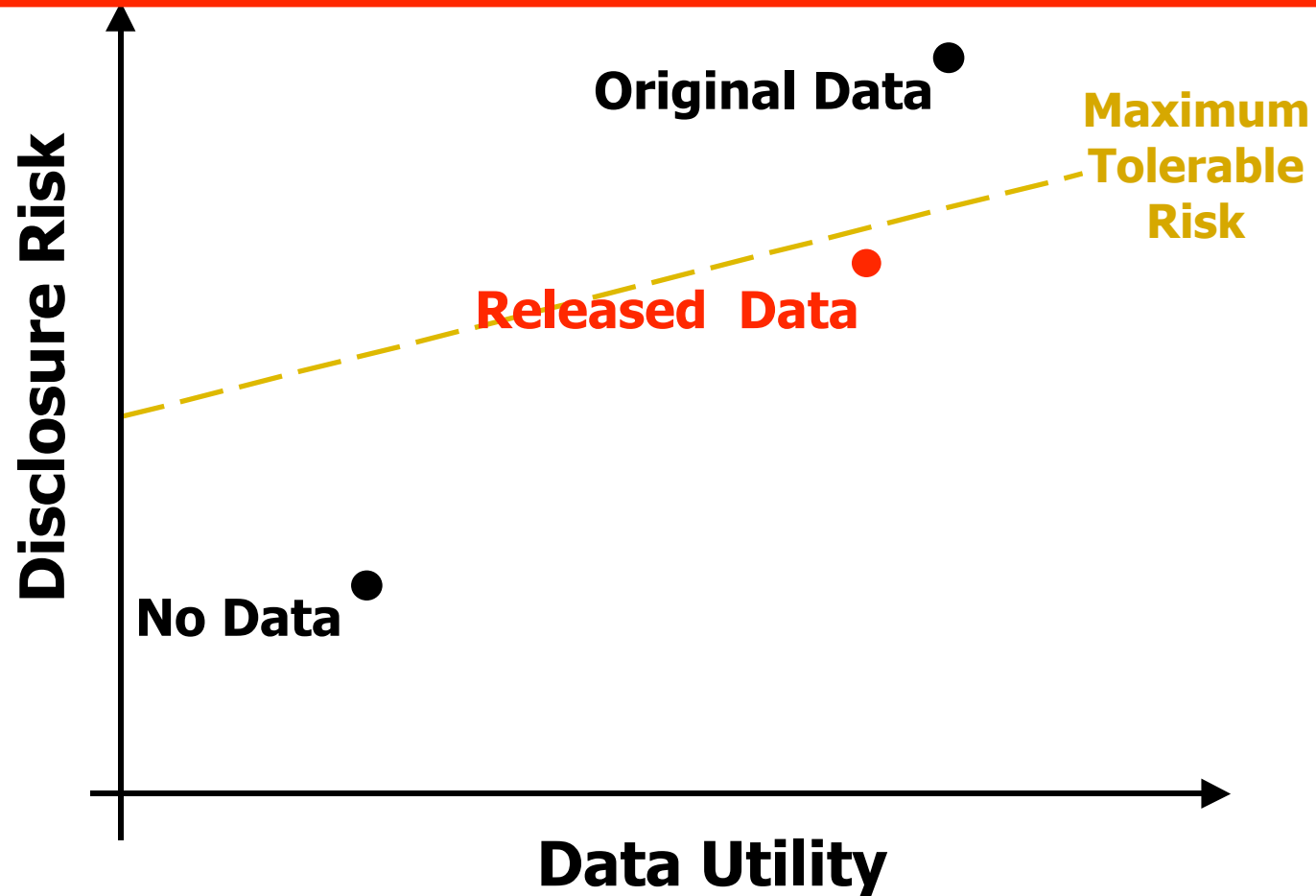
Race	Sex	Male		Female	
	Age	<18	≥18	<18	≥ 18
White		4	31	3	32
Black		0	1	0	0
Asian		1	3	2	3
2+		1	0	2	0

**Note**  
**reporting of**  
**population**  
**uniques!**

# 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.**
- **NISS Project and some of my students are dealing with other models and other types of releases.**

# R-U Confidentiality Map



(Duncan, et al. 2004)

## Ex. 3: American FactFinder

- **Some data reported without protection by law!!**
  - Table in earlier transparency was constructed from such data.
- **Risk** assessed using ad hoc approaches and obsolete criteria.
- **Utility** conceived of in terms of individually released tables and not underlying joint tabular structure, involving all of data.
- **No accounting for census error.**
- **What is alternative?**

# Example 4: Risk Factors for Coronary Heart Disease

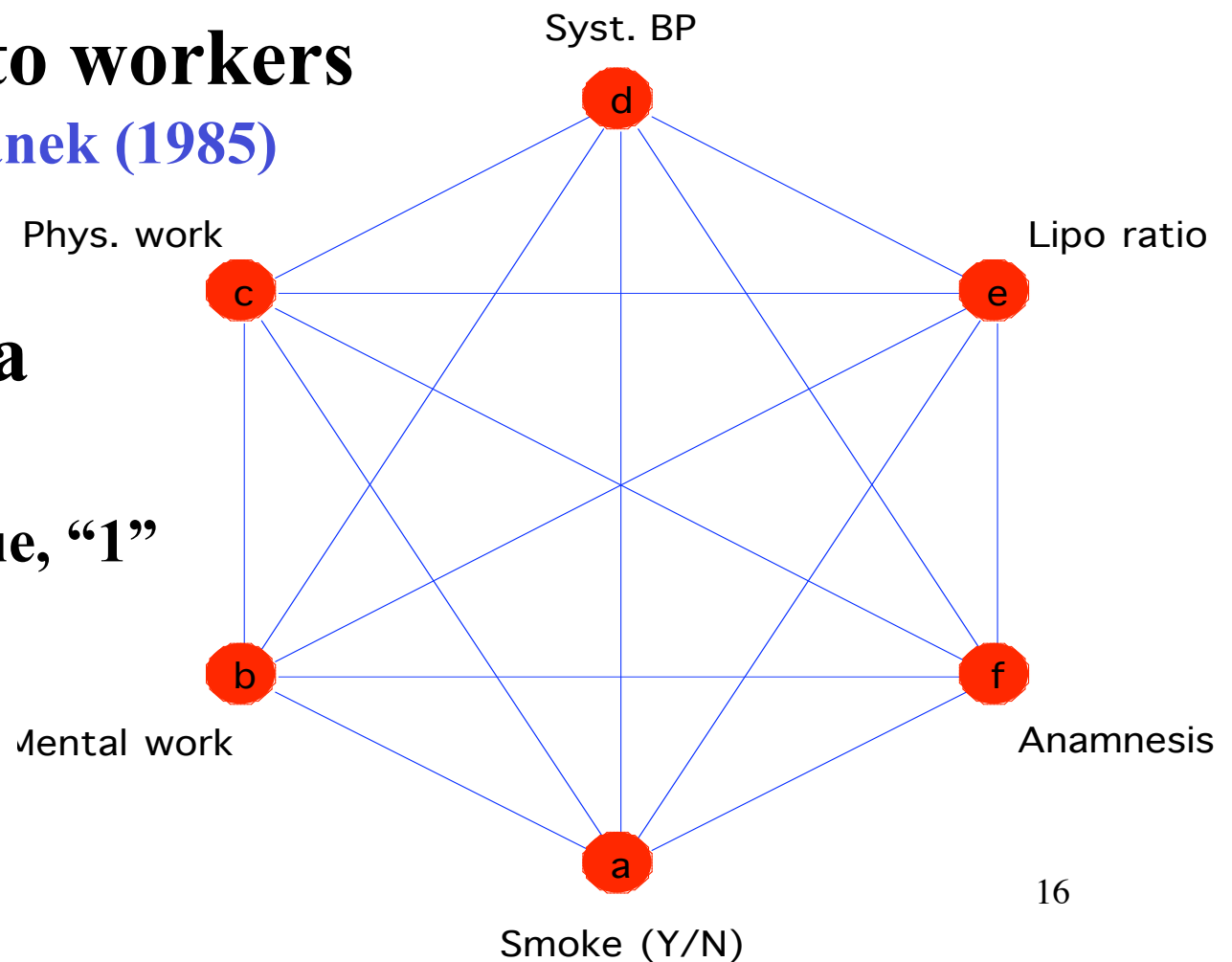
- 1841 Czech auto workers

Edwards and Havanek (1985)

- $2^6$  table

- population data

- “0” cell
- population unique, “1”
- 2 cells with “2”



# Example 4: The Data

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

# Two-Way Fréchet Bounds

- For  $2 \times 2$  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 Tables Entries

- ***k*-way table of non-negative counts,  $k \geq 3$ .**
  - 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.**  
**exploiting 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 logarithms 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

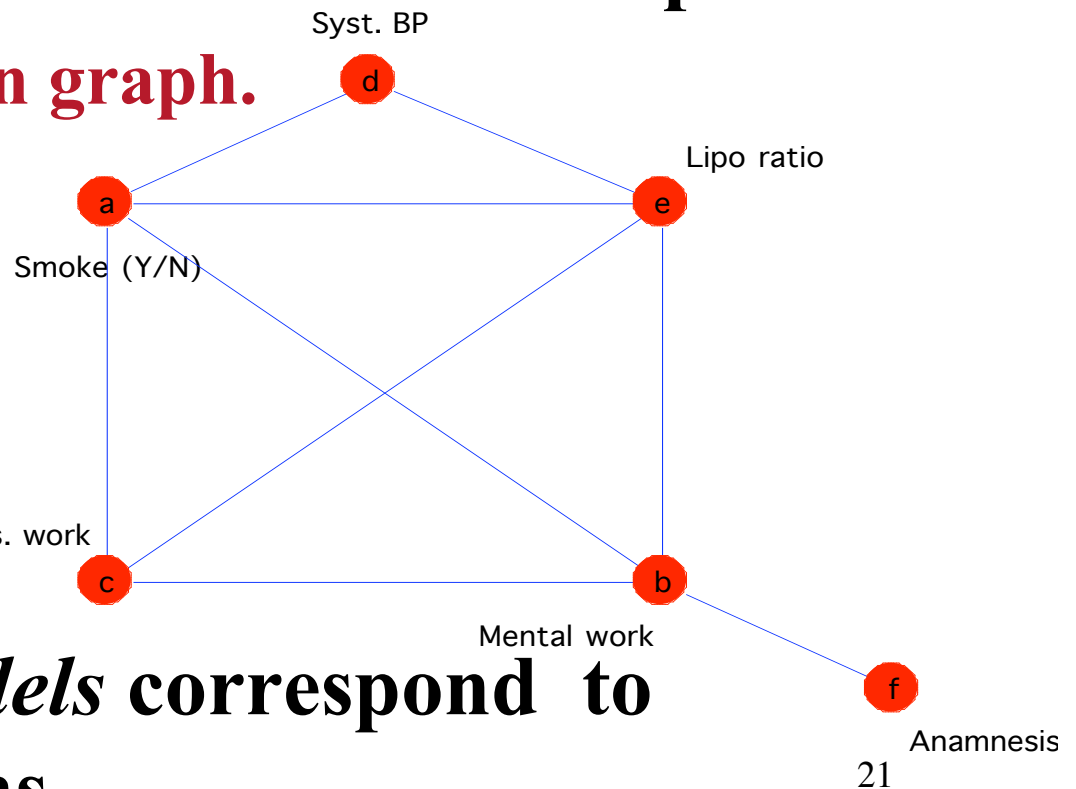
– Absence of edges in graph.

## Example 3:

Czech autoworkers

Graph has 3 cliques:

[ADE][ABCE][BF]



- *Decomposable models* correspond to triangulated graphs.

# 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:

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

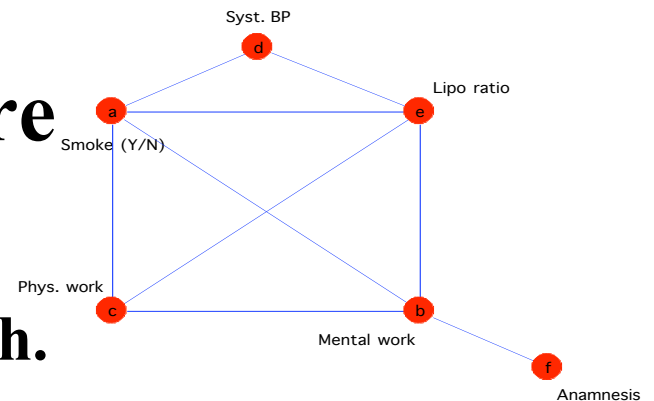
- ***Theorem***: When released margins correspond to those of decomposable model:
  - *Upper bound*: minimum of values from relevant margins.
  - *Lower bound*: maximum of zero, or sum of values from relevant margins minus separators.
  - Bounds are sharp.

# Example 4: 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		yes		
				A	no	no	yes	
neg	< 3	< 140	no		[0,88]	[0,62]	[0,224]	[0,117]
			yes	[0,261]	[0,246]	[0,25]	[0,38]	
			no	[0,88]	[0,62]	[0,224]	[0,117]	
		≥ 3	< 140	no	[0,58]	[0,60]	[0,170]	[0,148]
			yes	[0,115]	[0,173]	[0,20]	[0,36]	
			no	[0,58]	[0,60]	[0,170]	[0,148]	
	pos	< 3	≥ 140	no	[0,88]	[0,62]	[0,126]	[0,117]
				yes	[0,134]	[0,134]	[0,25]	[0,38]
				no	[0,88]	[0,62]	[0,126]	[0,117]
		≥ 3	< 140	no	[0,58]	[0,60]	[0,126]	[0,126]
			yes	[0,115]	[0,134]	[0,20]	[0,36]	
			no	[0,58]	[0,60]	[0,126]	[0,126]	
		≥ 140	yes	[0,115]	[0,134]	[0,20]	[0,36]	

# Statistical Theory & Bounds

- **Computationally efficient extensions for:**
  - Released margins corresponding to log-linear models that have reducible graphs.
  - For  $2^k$  tables with release of all  $(k-1)$ -dimensional margins fixed.
- **General “shuttle” algorithm in Dobra (2002) is computationally intensive but works for all cases:**
  - **Generates special cases with limited extra computation.**
  - Collapsing categories of selected variables.

## **Example 4: Release of All 5-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 4: What to Release?

- 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 and these involve [ABCDE].
- Suppose we deem release of everything else as safe, i.e., we release [ACDE][ABCDF][ABCEF][BCDEF][ABDEF]
- What can user and intruder do?

## Example 4: Making Proper Statistical Inferences

- Can fit all reasonable models including our “favorite” one:  $[ADE][ABCE][BF]$ .
- Can carry out 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 in this example.

# Example 5: Clinical Trial Data

		<b>Response</b>	<b>Poor</b>	<b>Moder.</b>	<b>Excel.</b>
<b>Center</b>	<b>Status</b>				
<b>1</b>	<b>1</b>	<b>Active</b>	<b>3</b>	<b>20</b>	<b>5</b>
<b>1</b>	<b>1</b>	<b>Placebo</b>	<b>11</b>	<b>14</b>	<b>8</b>
<b>1</b>	<b>2</b>	<b>Active</b>	<b>3</b>	<b>14</b>	<b>12</b>
<b>1</b>	<b>2</b>	<b>Placebo</b>	<b>6</b>	<b>13</b>	<b>5</b>
<b>2</b>	<b>1</b>	<b>Active</b>	<b>12</b>	<b>12</b>	<b>0</b>
<b>2</b>	<b>1</b>	<b>Placebo</b>	<b>11</b>	<b>10</b>	<b>0</b>
<b>2</b>	<b>2</b>	<b>Active</b>	<b>3</b>	<b>9</b>	<b>4</b>
<b>2</b>	<b>2</b>	<b>Placebo</b>	<b>6</b>	<b>9</b>	<b>3</b>

# Example 5: Analysis

- Interested in effect of T on R:
  - “Good Model” model: [CST][ CSR]
  - Target model of Inference: [CST][CSR][TR]
  - $\Delta G^2 = 5.4$  with 2 d.f.

– **Bounds:**

Center	Status	Response Treatment	Poor	Moderate	Excellent
1	1	Active	[0,14]	[1,28]	[0,13]
1	1	Placebo	[0,14]	[6,33]	[0,13]
1	2	Active	[0,9]	[3,27]	[1,17]
1	2	Placebo	[0,9]	[0,24]	[0,16]
2	1	Active	[2,21]	[3,22]	[0,0]
2	1	Placebo	[2,21]	[0,19]	[0,0]
2	2	Active	[0,9]	[0,16]	[0,7]
2	2	Placebo	[0,9]	[2,18]	[0,7]

- Releasing more margins produces much tighter bounds.

# Warning re Bounds

- **Bounds may not not be sufficient to understand degree of protection.**
  - **With sufficient marginal constraints, gaps in range of values are possible!**
    - de Loera and Onn (2003, unpublished ms.)**
  - **Consider cell in 3-way table and specify possible values: e.g., 0 and 2 (with gap at 1).**
  - **Then it is possible to find set of 3-way tables whose common 2-way marginals imply that there exists only tables corresponding to these values!**

# Example: 3-way Bounds Gap

- Let the specified values for (1,1,1) cell be 0 and 2.
- Then there exists set of  $6 \times 4 \times 3$  tables with following marginals, all of which possess

gap:

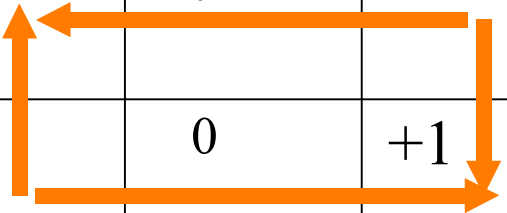
2	1	1	0
1	0	0	1
2	2	0	0
0	0	2	2
2	0	2	0
0	2	0	2

2	2	0
1	1	0
2	0	2
3	0	1
0	2	0
0	1	3

2	3	2
2	1	2
2	1	2
2	1	2

# 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.**

# 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 Sulivant, 2002)**
  - **Simple moves for some but not all  $2^k$  tables (“binomials”) (Aoki and Takamura, 2003)**
  - **Rooted in ideas from likelihood theory for log-linear models and computational algebra of toric ideals.**

# Warning Re Perturbations

- **Perturbation distributions using Markov bases can possess disjoint pieces, if marginals are sufficiently constraining.**
  - [Aoki and Takamura \(2003\)](#)
  - Link to gaps result for bounds.
- **Implications:**
  - Multimodal “exact” distributions.
  - Need to consider carefully how to do perturbation and what to report to users.

# 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 (e.g., MSSs).**
- **Methods should allow for reversibility for inference purposes:**
  - **Missing data should be “ignorable” for inferences.**
  - **Assessing goodness of fit is important.**

# **What If Released Margins Are Not Sufficient?**

- **Approach outlined here works if margins include MSSs of suitable log-linear model for data.**
- **Sometimes no log-linear model fits (Example 1: NLTCS) or no set of releasable margins gives MSSs for acceptable models.**
- **What to do?**

# Summary

- **Discussed purposes of disclosure limitation and the tradeoff with utility.**
- **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.**
- **Articulated some general principles for developing DL methods.**

# The End

- **Most papers available for downloading at**

<http://www.niss.org>

[www.stat.cmu.edu/~fienberg/disclosure.html](http://www.stat.cmu.edu/~fienberg/disclosure.html)