

# **Bounds and Exact Distributions for Categorical Statistical Databases**

---

**Stephen E. Fienberg**

**Department of Statistics &**

**Center for Automated Learning and  
Discovery**

**Carnegie Mellon University**

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

# CMU Activities

---

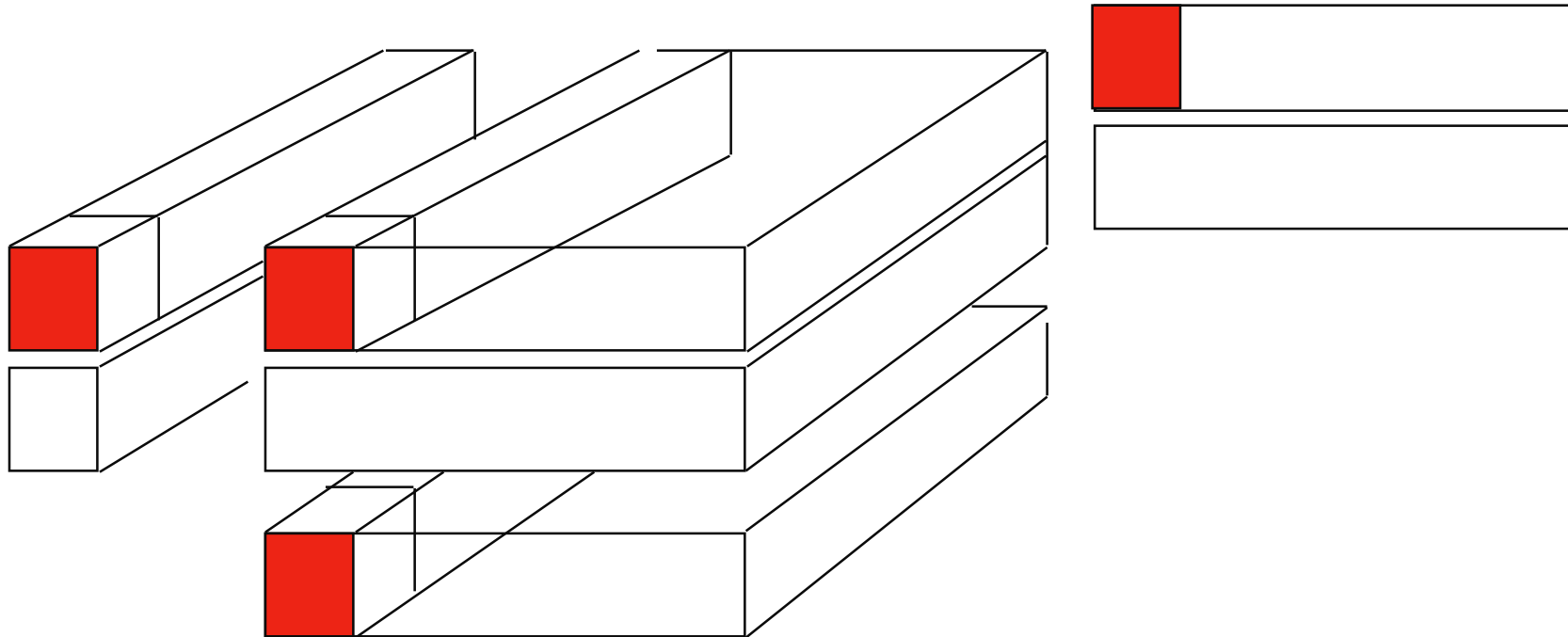
- **\*Work on bounds for tables and perturbation methods**
  - Fienberg, Roehrig, and Dobra.
- **Assessing Risk of Disclosure**
  - Duncan and Keller-McNulty.
  - Fienberg and Trottini.
    - » Trottini thesis proposal.\*\*
- **CS Data Base Issues**
  - Andrew Moore.

# Query System For $k$ -way Table of Counts

---

- ***Queries:*** Can only come in the form of 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 and thus simulated releases also become highly constrained.

# Query System Illustration ( $k=3$ )

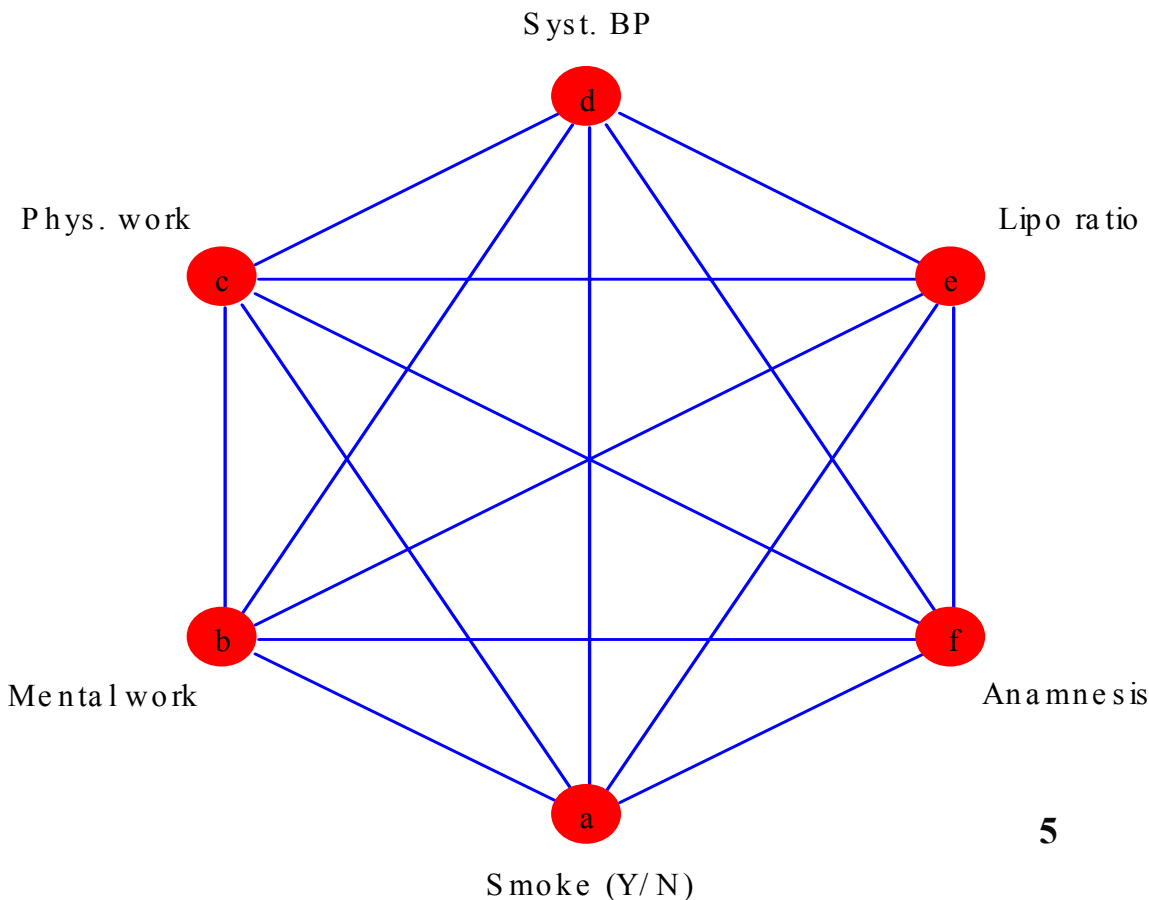


***Challenge:*** Scaling up approach for large  $k$ .

# Example: Risk Factors for Coronary Heart Disease

c	d	e	f	Counts
[a=1 b=1]	[a=1 b=2]	[a=1 b=2]	[a=2 b=2]	
1	1	1	1	44
1	1	1	2	5
1	1	2	1	23
1	1	2	2	7
1	2	1	1	35
1	2	1	2	4
1	2	2	1	24
1	2	2	2	4
2	1	1	1	129
2	1	1	2	9
2	1	2	1	50
2	1	2	2	9
2	2	1	1	109
2	2	1	2	14
2	2	2	1	51
2	2	2	2	5

- 1841 Czech auto workers  
– 6 binary factors:  $2^6$  table



# Fréchet Bounds

- For 2x2 tables of counts  $\{n_{ij}\}$  given the marginal totals  $\{n_{1+}, n_{2+}\}$  and  $\{n_{+1}, n_{+2}\}$ :

$n_{11}$	$n_{12}$	$n_{1+}$
$n_{21}$	$n_{22}$	$n_{2+}$
$n_{+1}$	$n_{+2}$	$n$

$$\min\{n_{+1}, n_{1+}\} \geq n_{11} \geq \max\{n_{+1} + n_{1+} - n, 0\}$$

- New multi-way generalizations involving higher-order, overlapping margins.

# Bounds for Multi-Way Tables

---

- ***k*-way table of counts,  $k \geq 3$ .**
  - Direct generalizations to tables with non-negative entries.
- **Release sequence of marginal totals, possibly overlapping.**
- ***Goal*: Compute bounds for cell entries.**

# Some Approaches

---

- **Network models (Cox):**
  - Needs formal structure to work even for  $k=3$ !
  - No formulation for  $k \geq 4$ .
- **Simplex method (Roehrig, et al.):**
  - Works well for small problems and dimensions.
    - » Fractional bounds.
  - “Linear programming” problem is NP-hard.
- **Shuttle Algorithm (Buzzigoli/Giusti)**
  - Doesn’t work!

# Our Strategy

---

- **Develop efficient methods for variety of special cases.**
- **Exploit linkage to statistical theory**
- **Log-linear models for multi-way tables:**
  - **Marginal totals are MSSs**
  - **Existence of MLEs.**
  - **Subclasses of decomposable and graphical models have special properties.**

# Progress to Date

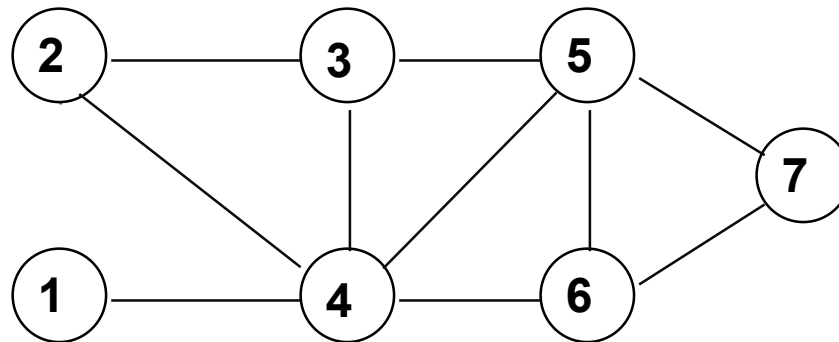
---

- **Explicit formulas for sharp bounds in decomposable case for  $k$ -way tables.**
- **Extension for reducible graphical case.**
  - To appear in Dobra & Fienberg: *Proceedings of the National Academy of Sciences*, (2000).
- **$k$ -way tables with  $(k-1)$  dimensional margins fixed.**
  - Dobra, forthcoming technical report.

# Graphical & Decomposable Log-linear Models

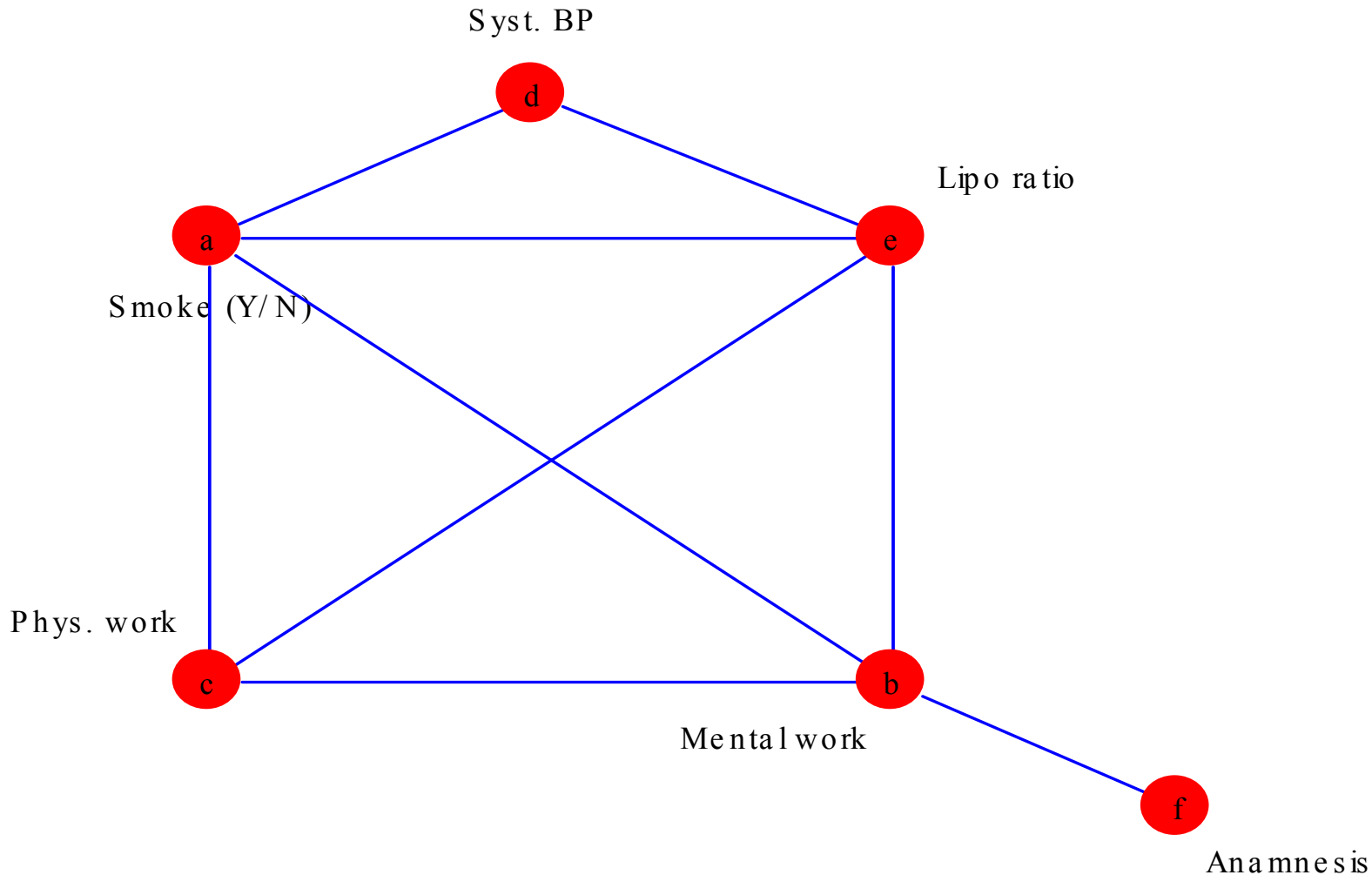
- “Graphical” models defined in terms of simultaneous conditional independence relationships or absence of edges in graph.

Example:



- Decomposable models have closed form structure and special properties:
  - Correspond to triangulated graphs.

# Example: Coronary Heart Disease Risk Factors



# More on Bounds (cont.)

- **For decomposable log-linear models:**
  - Expected cell values are explicit function of margins.
    - » These are cliques in graph.
  - *Upper bound*: minimum of relevant margins.
  - *Lower bound*: maximum of zero, or sum of relevant margins minus separators.
  - Bounds are sharp.
- **Example: Complete Independence**

$$\min\{n_{1+\dots+}, n_{+1+\dots+}, \dots, n_{+1+\dots+}, \dots\} \\ \geq n_{11\dots 1} \geq \max\{n_{1+\dots+} + n_{+1+\dots+} + \dots + n_{+\dots+1} - n(k-1), 0\}$$

# Example: Coronary Heart Disease Risk Factors

---

- **Table 2 in handout contains bounds for example when released margins are  $\{ABCE\}\{CDE\}\{BF\}$ .**
  - **Corresponding to decomposable graph.**
- **Cell containing population unique has bounds  $[0, 22]$ .**
  - **Cells with entry of “2” have bounds:  $[0,20]$  and  $[0,38]$ .**
  - **Lower bound are all “0”.**
- **“Safe” to release these margins, they do not pose substantial risk of disclosure.**

# Example (cont.)

---

- **For 10% from original table (Table 4), and the same released margins, Table 5 contains bounds:**
  - Many sample uniques and many values of “2”.
  - Some lower bounds are non-zero.
  - Bounds for cells with entries of “1” and “2” are relatively tight.
- **But given sampling fraction, it is still be acceptable to release table from disclosure risk perspective.**

# More on Bounds (cont.)

---

- **Computationally useful extensions of bound results for models and margins corresponding to reducible graphs.**
- **For some special cases of non-graphical models can construct bounds formulas:**
  - $2^k$  tables with  $(k-1)$  dimensional margins fixed (need one extra bound here and it comes from log-linear model theory).
  - Extend to general  $k$ -way case by looking at all possible collapsed  $2^k$  tables.

# Example Again

---

- **For example, if we release all 5-way margins, Table 3 gives bounds:**
  - **Highly constrained.**
  - **Almost identical upper and lower values; they all differ by 1.**
- **For 10% sample the 5-way margins determine the cell values exactly.**
- **No longer safe to release margins given the population unique.**

# **Perturbation/Simulation and Disclosure Limitation**

---

- **Idea of simulation subject to constraints to match original data base:**
  - Preserve marginal distributions.
- **Simulation process intended to produce data that can be used “as if” they were the real data for statistical analyses.**
  - Need to know probabilistic characterization of process that generates the simulated data.
- **Simulated cases should not correspond to real people.**

# Exact Distribution of Table Given Margins

---

- **Intertwined with bounds results.**
- **Role of Gröbner bases for generation:**
  - Using ideas and approaches outlined in Fienberg, Makov, and Steele (1998) *JOS*; and Fienberg, Makov, Meyer, and Steele (2000), Fraser volume.
- ***Problem*: Complex calculations to generate Gröbner bases.**
  - Scalability of approaches.

# New Results

---

- **Specification of Gröber basis elements for decomposable case.**
- **Efficient generalization for reducible case.**
  - **Dobra: new technical report.**
- **Improved algorithms for Gröbner basis generation in general cases.**
  - **Roerig: forthcoming technical report.**

# Example Yet Again

---

- **Implications of tight bounds for perturbation and exact distributions.**

# Next Steps

---

- **Code for algorithms for bounds and exact distributions for use in query system.**
- **Technical reports with code and examples.**
- **Extending to more interesting classes of special cases.**
- **Scaling up.**

# Scaling Up

---

- **How effective are such devices for limiting disclosure, i.e., protecting against intruder?**
- **What is information loss when we compare actual data with those released?**
- **How to implement simulation strategies effectively to generate disclosure-limited samples when  $k$  is large?**
  - Large sparse contingency tables.
  - Multiplicity of models of interest.
- **Data base issues (Andrew Moore).**

# Presentations

---

- **Recent Presentations**

- **IAOS Human Rights and Economic Development Conference, Montreux, Switzerland, September, 2000.**

- **Forthcoming Presentations**

- **Israel Conference on Foundations of Statistics and Its Application, December, 2000.**
- **Skopje Confidentiality Conference, March, 2001.**
- **Tulane Conference on Gröbner Bases, September, 2001.**

# Publications

---

- **Fienberg et al. Fraser volume (in press).**
- **Dobra and Fienberg, PNAS (in press).**
- **Fienberg, IAOS Conference Proceedings (in press).**
- **Fienberg and Makov, ISBA Conference Proceedings (under review).**

# Technical Reports

---

- **Dobra, on Gröbner bases for decomposable and reducible cases.**
- **Dobra and Fienberg, on  $k$ -way tables given  $(k-1)$ way margins.\***
- **Trottini, thesis proposal on criteria for release.**