Statistical Issues Associated With Multi-way Contingency Tables & Links to Algebraic Geometry

Stephen E. Fienberg

Cylab, Department of Statistics, &

Machine Learning Department

Carnegie Mellon University

&

IMA

(Joint work with A. Dobra, A. Rinaldo, & Y. Zhou)

Preliminaries

- I am an "A" at IMA for ***Applications of Algebraic Geometry*.**
- **This talk:**
  - **Continuation from last week's seminar by Serkan Hosten.**
    - **I won't provide a notational translation table but I will overlap and give links.**
  - **Introduction to a number of statistical problems for the analysis of categorical data.**

Overview

Three data examples and two statistical problems:

1. **Bounds for cell counts in contingency tables given marginals.**
2. **Maximum likelihood estimation for log-linear models and large sparse contingency tables.**

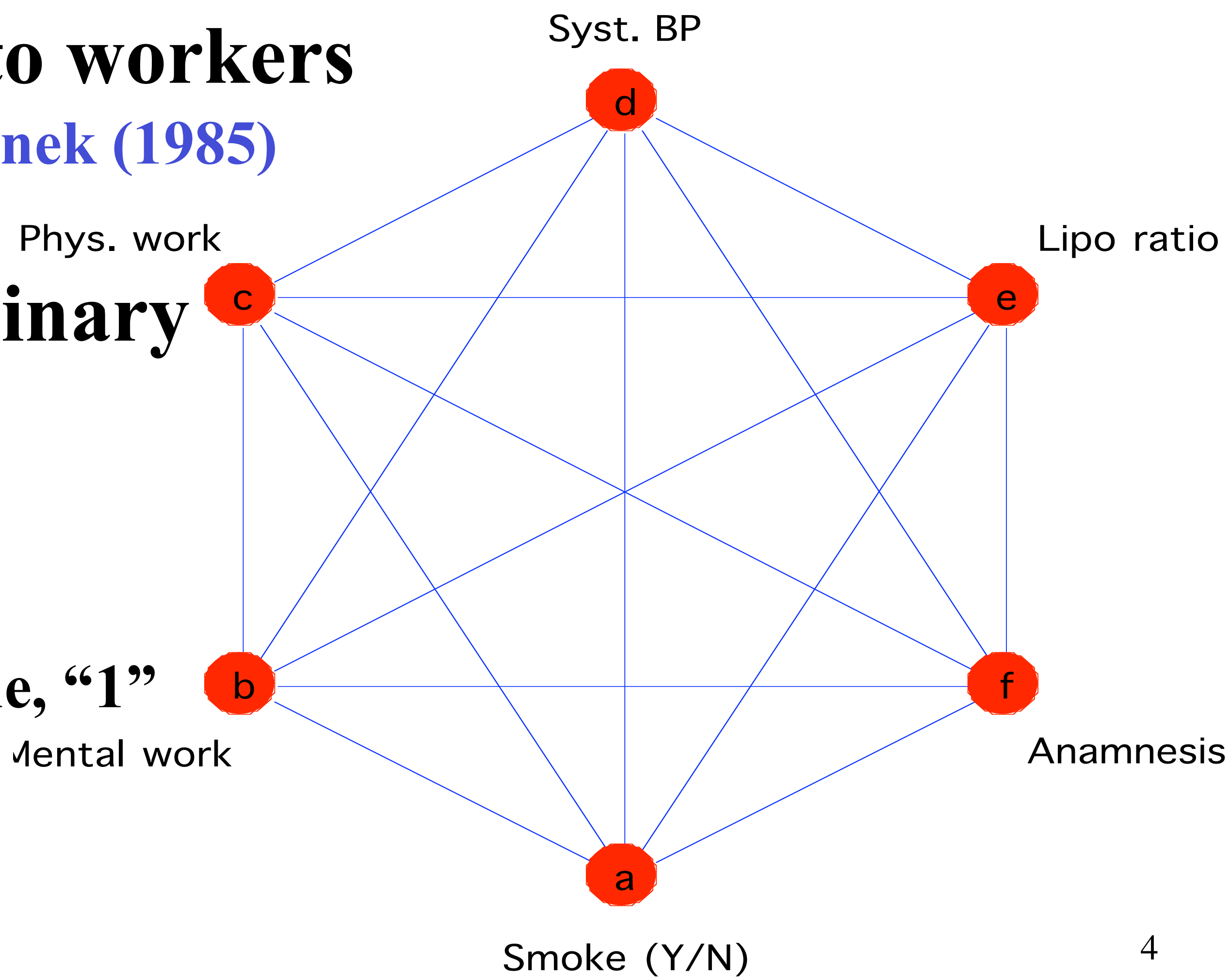
How are they interrelated?

Where do algebraic and other geometry tools fit in?

Scaling up computations to deal with large sparse tables.

Ex. 1: Risk Factors for Coronary Heart Disease

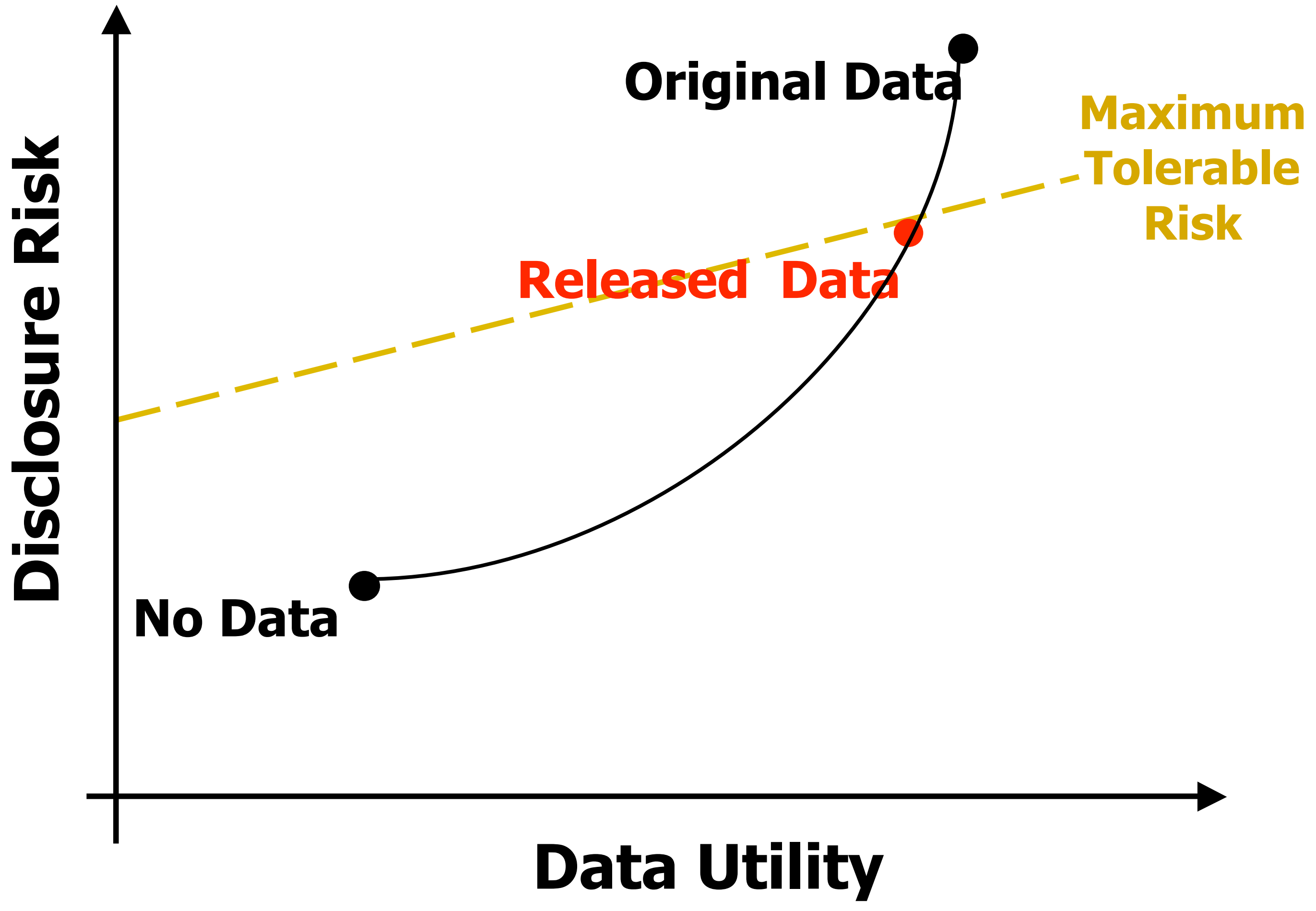
- **1841 Czech auto workers**
  Edwards and Havanek (1985)
  Biometrika
- **Selection of 6 binary variables**
- **2^6 table**
  - **"0" cell**
  - **population unique, "1"**
  - **2 cells with "2"**



Ex. 1: The Data

				B	no		yes	
F	E	D	C	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

R-U Confidentiality Map



(Duncan, et al. 2004)

Disclosure Limitation for Sparse Count Data

- **Uniqueness in population table ⇔ cell count of "1":**
  - **Uniqueness allows intruder to match characteristics in table with other data bases that include same variables to learn confidential information.**
- **Utility typically tied to usefulness of marginal totals for statistical inference.**
- **Risk concerned with small cell counts.**
  - **Assess using bounds for cell counts given marginal totals.**

Marginals as Data Releases

- **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.**
- ***National Institute of Statistical Sciences* Project and some of my former students have dealt with other models and other types of releases.**

Ex. 2: Genetics Linkage

- **Data come from a barley milkdew experiment.**
  - **Edwards (1992).** ***Comp. Stat. Data Anal.***
  - **37 binary variables (genes) and 81 cases (5% missing data).**
- **Subset of 6 genes that appear closely linked on basis of marginal distributions?**
- **On same chromosome?**

Ex. 2: The Data

			1				2				D
			1		2		1		2		E
			1	2	1	2	1	2	1	2	F
1	1	1	0	0	0	0	3	0	1	0	
		2	0	1	0	0	0	1	0	0	
	2	1	1	0	1	0	7	1	4	0	
		2	0	0	0	2	1	3	0	11	
2	1	1	16	1	4	0	1	0	0	0	
		2	1	4	1	4	0	0	0	1	
	2	1	0	0	0	0	0	0	0	0	
		2	0	0	0	0	0	0	0	0	
A	B	C									

Ex. 3: Australian Census Data

- **10-dimensional highly sparse contingency table extracted from 1981 Australian population census (based on10 million people):**

Variable	BPL	SEX	AGE	REL	MST	DUR	QAL	INC	FIN	TIS
# Categ.	102	2	11	27	5	62	11	15	16	18

- **892,533,945,600 cells!**

Collapsed Tables

- **Collapsed 5-way table with 105,600 cells of which 65% are zero**

Variable	BPL	MST	QAL	INC	FIN
# Categ.	8	5	11	15	16

- **Collapsed 6-way table with 48,000 cells of which 41% are zero**

Variable	BPL	SEX	AGE	REL	MST	QAL
# Categ.	8	2	11	5	5	11

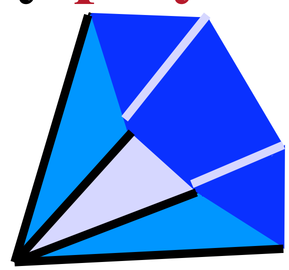
Two Faces of Algebraic Statistics & Contingency Tables

1. **Representation of statistical models for cell probabilities: Description of parameter space.**
   - A. **Characterizing joint distributions.**
   - B. **Log-linear models including those with "graphical representation" via conditional independencies.**
2. **Statistical inference: Studying and characterizing portions of sample space:**
   - A. **Minimal sufficient statistics (sufficient data summaries) for models—marginal totals.**
   - B. **Maximum likelihood estimation.**
   - C. **Distribution over all possible having given marginals ("exact distribution")—related bounds.**

Its All About Geometry

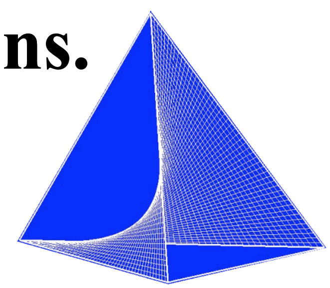
- ***Polyhedral Geometry***: **virtually all data-related quantities can be described by polyhedra.**

Polytope



Polyhedral Cone

- ***Algebraic Geometry***: **a statistical model is specified by a polynomial map. The set of probability distributions is a hyper-surface of points satisfying polynomial equations.**



Algebraic (Toric) Variety

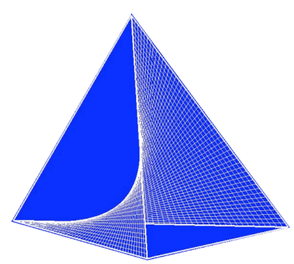
2×2 Table: The Model

p_{11}	p_{12}	p_{1+}
p_{21}	p_{22}	p_{2+}
p_{+1}	p_{+2}	1

- **We are interested in the distribution of the 4 cells in the table specified by the vector of log probabilities:**
- **Model of independence: $p_{ij} = p_{i+}p_{+j}$**

$$log(p_{11}, p_{12}, p_{21}, p_{22}) = A'\theta = (p_{1+}, p_{2+}, p_{+1}, p_{+2})$$

- **The set of all probability distributions for model of independence need to satisfy one polynomial equation: $p_{11}p_{22} - p_{12}p_{21} = 0$, and belong to surface of independence:**



Segre Variety

2×2 Table: The Data

Model of independence: $p_{ij} = p_{i+}p_{+j}$

Observed Counts

n_{11}	n_{12}
n_{21}	n_{22}

→

MSS Margins

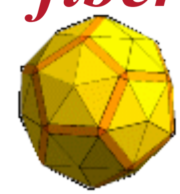
$t = An$

Design Matrix

	n_{11}	n_{12}	n_{21}	n_{22}
$t_1 = n_{1+}$	1	1	0	0
$t_2 = n_{2+}$	0	0	1	1
$t_3 = n_{+1}$	1	0	1	0
$t_4 = n_{+2}$	0	1	0	1

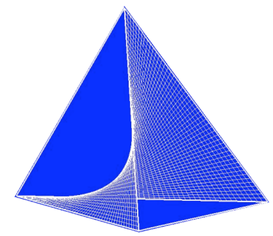
- **Set of all tables having margins *t* are integer points inside a polytope and form the *fiber*:**

$$\{n \in \mathbb{R}^4_{\geq 0}, An = t\}$$



Design Matrix A

MLE



Sample Space

A identifies the fiber: the set of all tables having the same margins:

$$\{x \geq 0, Ax = t\}$$

Leads to the generalized hypergeometric probability distribution.

[Set of all tables are lattice points in the simplex.]

Parameter Space

A specifies the set of polynomial equations that encode the dependence among the variables.

All probability vectors satisfy binomial equations:

$$p^{u+} - p^{u-} = 0$$

all integer $u \in$ kernel (A).

Maximum Likelihood Estimation

- **Distribution for n given p:**

$$f(n \mid p) \propto p_{11}^{n_{11}} p_{12}^{n_{12}} p_{21}^{n_{21}} p_{22}^{n_{22}}$$

- **For model of independence: $p_{ij} = p_{i+} p_{+j}$ minimal sufficient statistics for parameters are:**
  - $t = An = (n_{1+}, n_{2+}, n_{+1}, n_{+2})$
- **Maximum likelihood equations:**
  - $p_{i+} = n_{i+}/n \quad i = 1, 2; \; p_{+j} = n_{+j}/n \quad j = 1, 2.$
- **Solution (MLEs):** $\hat{p}_{ij} = n_{i+} n_{+j} / n^2.$
- **Rescale by total n to count scale $np_{ij} = m_{ij}$:**

$$\hat{m}_{ij} = n_{i+} n_{+j} / n.$$

Two-Way Fréchet Bounds

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

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

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

- **Link to independence:** $\hat{m}_{ij} = n_{i+} n_{+j} / n.$
- **Interested in multi-way generalizations involving higher-order, overlapping margins.**

Log-linear Models for 2^3 Tables

- **In 3-way table of counts, $\{n_{ijk}\}$, we model logarithms of expectations, $E(n_{ijk})=m_{ijk} > 0$:**

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

- **MSSs are margins corresponding to highest order u-terms: $\{n_{ij+}\}$, $\{n_{i+k}\}$, $\{n_{+jk}\}$.**
  - **MSSs describe *simplicial complex*: [12][13][23].**
- **Alternative ways to write model:**

$$m_{ijk} = \alpha_{ij}\beta_{ik}\gamma_{jk}$$

$$\frac{m_{111}m_{221}}{m_{121}m_{211}} = \frac{m_{112}m_{222}}{m_{122}m_{212}}$$

$$m_{111}m_{221}m_{122}m_{212} - m_{121}m_{211}m_{112}m_{222} = 0$$

Log-linear Models (cont.)

- Maximum likelihood estimates (MLEs) found by setting MSSs equal to their expectations:

$$\hat{m}_{ij+} = n_{ij+} \quad \text{for } i = 1,2,, j = 1,2,$$

$$\hat{m}_{+jk} = n_{+jk} \quad \text{for } j = 1,2,,k = 1,2,$$

$$\hat{m}_{i+k} = n_{i+k} \quad \text{for } i = 1,2,k = 1,2.$$

- Set: $m_{ijk} = n_{ijk} \pm \delta$
- Solve cubic equation for δ:

$$m_{111}m_{221}m_{122}m_{212} - m_{121}m_{211}m_{112}m_{222} = 0$$

- When do we get +ve solutions for $\{m_{ijk}\}$?

Existence of MLEs for 2×2×2 Table

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

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

n_{11+}	n_{12+}
n_{21+}	n_{22+}

Delta must be zero and MLE doesn't exist.

Two Other 3-Way Examples With [12][13][23]

- 3^3 table where MLE exists

3	0	0
0	4	0
0	0	4

0	0	1
5	0	0
0	2	0

0	1	0
0	0	5
3	0	0

- 4^3 table where MLE does not exist

0	0	0	4
0	0	1	2
0	1	2	3
5	1	2	3

4	0	0	2
5	0	5	2
5	6	5	2
1	0	0	0

1	5	0	2
5	3	4	2
0	2	0	0
1	2	0	0

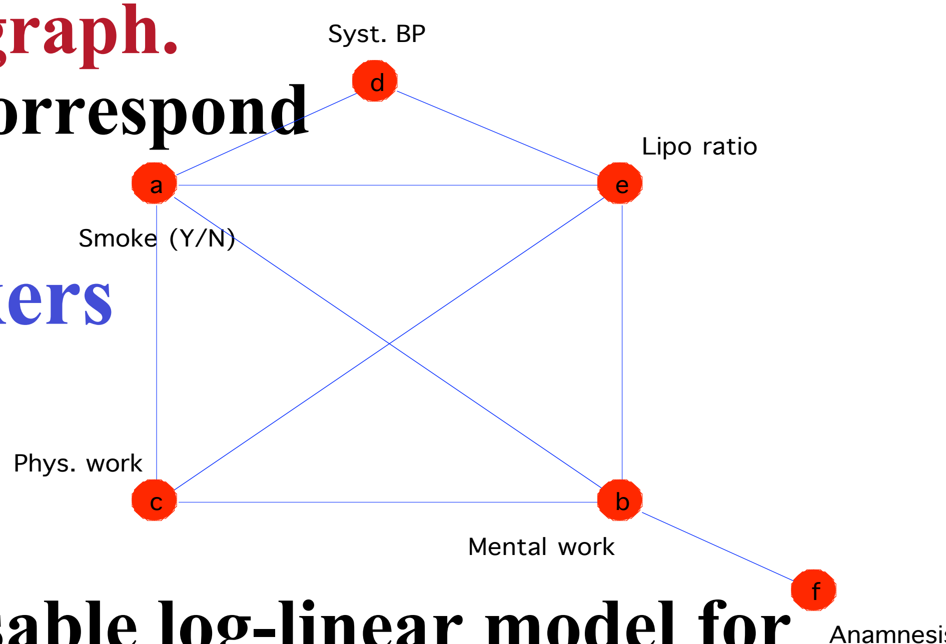
1	5	3	2
0	0	2	0
0	2	4	0
1	2	3	0

MLEs for Log-Linear Models for *k*-Way Tables

- Log-linear models and algebraic geometry representations generalize.
- Sampling distributions for $f(n \mid p)$ are key!
  - ML equations then have similar form.
- Existence of MLEs linked to pattern of zeros:
  - Discoverable by defining basis for models and using algebraic and polyhedral geometry.
  - Examples discovered using *Polymake*.
- General theorem in Haberman (1974) and "constructive" version in Rinaldo (2005).

Graphical & Decomposable Log-linear Models

- *Graphical log-linear models:* defined by simultaneous conditional independence relationships:
  - Absence of edges in graph.
- *Decomposable models* correspond to triangulated graphs.



Ex. 1: Czech autoworkers

- Graph has 3 cliques:

[ADE][ABCE][BF]

- "Interesting" decomposable log-linear model for data!

MLEs for Decomposable Log-linear Models

- **For decomposable models, expected cell values are explicit function of margins, corresponding to highest order terms in model (*cliques* in graph):**

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

  - **e.g., cond. indep. in 3-way table:** $m_{ijk} = \frac{m_{ij+} m_{i+k}}{m_{i++}}$

- **Substitute observed margins for expected in explicit formula to get MLEs.**
  - **Hosten: ML degree 1.**

Bounds for *k*-way Table Entries Given Set of Marginals

- **Methods for several special cases:**
  - When margins corresponding to decomposable models, bounds have explicit formulae.
  - When margins corresponding to reducible graphs, calculation can be broken up into smaller problems.
  - Simple bounds result for 2^k tables with release of all (k-1)-dimensional margins fixed.
- **General, less efficient methods for searching over lattice points of convex polytope:**
  - Integer programming; MCMC with Groebner bases; general shuttle algorithm (Dobra, 2001).

2^3 Table Given 2×2 Margins

n_{111}	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}	n_{222}	n_{2+2}
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: $n_{111} + n_{222}$**

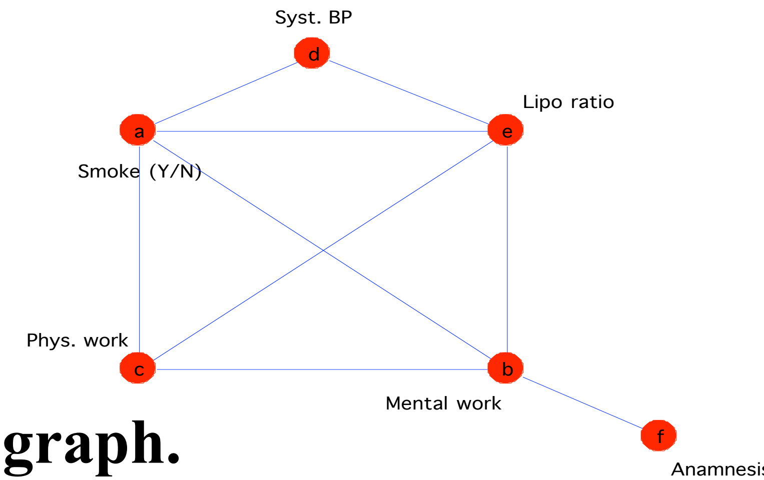
Multi-way Bounds

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

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

- ***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. Fienberg and Dobra (2000)**
  - **Link to Markov bases in Diaconis and Sturmfels (1998).**

Ex. 1: Czech Autoworkers



- Released margins:

  [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 / A	no / no	no / yes	yes / no	yes / 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]

Ex. 1: Counting Tables

- **How many tables are there with various sets of given marginals?**
  - **For release of only the 15 two-way margins there are 705,884 possible tables, although many of these may correspond to same bounds! [via 4ti2]**
  - **For [ACDEF][ABDEF][ABCDE][BCDF][ABCF][BCEF]**

    There are only 810 tables.
  - **For release of all 5-way margins, there are only 2 tables!**
    - **Almost identical upper and lower values; they all differ by 1.**

Ex. 1: 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].**
- **If treat everything else as safe, i.e., we release [ACDE][ABCDF][ABCEF][BCDEF][ABDEF]**
  - **Can fit all reasonable models including our "favorite"one: [ADE][ABCE][BF].**

Ex. 2: Genetic Linkage Data

			1				2				D
			1		2		1		2		E
			1	2	1	2	1	2	1	2	F
1	1	1	0	0	0	0	3	0	1	0	
		2	0	1	0	0	0	1	0	0	
	2	1	1	0	1	0	7	1	4	0	
		2	0	0	0	2	1	3	0	11	
2	1	1	16	1	4	0	1	0	0	0	
		2	1	4	1	4	0	0	0	1	
	2	1	0	0	0	0	0	0	0	0	
		2	0	0	0	0	0	0	0	0	
A	B	C									

Ex. 2: Existence of MLEs?

- When we fit model corresponding to [ACD][ADE][ADF][CE][CF][EF][BCD][BDE][BDF]

			1				2				D
			1		2		1		2		E
			1	2	1	2	1	2	1	2	F
1	1	1	0	0	0	0	+	0	+	0	
		2	0	+	0	0	0	+	0	0	
	2	1	+	0	+	0	+	+	+	0	
		2	0	0	0	+	+	+	0	+	
2	1	1	+	+	+	0	+	0	0	0	
		2	+	+	+	+	0	0	0	+	
	2	1	0	0	0	0	0	0	0	0	
		2	0	0	0	0	0	0	0	0	
A	B	C									

Ex. 2: Cont.

- **For [ACD][ADE][ADF][CE][CF][EF][BCD][BDE][BDF] there are 42 problematic zero cells:**
  - **Detected by generalized shuttle algorithm for bounds and verified by MLE software.**
  - **Correspond to zeros in all 255,880 tables.**
  - **Extended MLE exists here.**
- **For no-2nd-order interaction model there are 15 MSS marginals and no problematic zeros.**
  - **Based on shuttle algorithm and verified by MLE software.**
  - **8,628,046 tables.**

Discovering Non-Existence Using Bounds

- Replace positive counts by counts of 1.
- Run bounds algorithm and/or LP on 0-1 table.
  - Look for: upper bound = lower bound = 0.
  - Fractional LP bounds may not detect non-existence.
- Compare with methods for detecting non-existence of MLEs.
  - Is bounds software simpler than MLE software?

Degenerate MLE

- **Fixing all 15 positive 3-way margins produces following bounds using integer programming procedure in "*lp solve*":**

A	B	C									
			1				2				D
			1		2		1		2		E
			1	2	1	2	1	2	1	2	F
1	1	1	[0, 1]	[0, 0]	[0, 2]	[0, 0]	[1, 4]	[0, 1]	[0, 2]	[0, 1]	
		2	[0, 0]	[0, 2]	[0, 0]	[0, 2]	[0, 1]	[0, 2]	[0, 1]	[0, 1]	
	2	1	[0, 1]	[0, 0]	[0, 2]	[0, 0]	[6, 9]	[0, 1]	[1, 4]	[0, 1]	
		2	[0, 0]	[0, 1]	[0, 0]	[0, 2]	[0, 1]	[1, 4]	[0, 1]	[9, 12]	
2	1	1	[15, 18]	[0, 1]	[0, 4]	[0, 1]	[0, 1]	[0, 0]	[0, 1]	[0, 0]	
		2	[0, 1]	[2, 5]	[1, 2]	[1, 5]	[0, 0]	[0, 1]	[0, 0]	[0, 1]	
	2	1	[0, 1]	[0, 0]	[0, 2]	[0, 1]	[0, 1]	[0, 0]	[0, 1]	[0, 0]	
		2	[0, 0]	[0, 1]	[0, 1]	[0, 2]	[0, 0]	[0, 1]	[0, 0]	[0, 1]	

Ex. 3: Collapsed Tables

- **Collapsed 5-way table with 105,600 cells of which 65% are zero**

Variable	BPL	MST	QAL	INC	FIN
# Categ.	8	5	11	15	16

- **Collapsed 6-way table with 48,000 cells of which 41% are zero**

Variable	BPL	SEX	AGE	REL	MST	QAL
# Categ.	8	2	11	5	5	11

Ex. 3: 5-way Table

- **Table has 105,600 cells; 65% are 0.**
  - **We set counts in all positive cells = 1 to simplify the problem.**
- **Then we use LP to find upper bounds of cells when all the 2-way margins are fixed.**
  - **We can run the LP solver for the table cells in parallel.**
  - **In our experiment, we used cluster of 64 processors and it took about 4 hours.**
  - **Upper bounds of the cells are all positive, so there are no structural zeros found for this 5-way table.**

Ex. 3: 6-way Table

- **Table has 48,400 cells and 41% have zero cells.**
  - **Use 0-1 representation again.**
  - **Fixed all 2-way margins.**
  - **All upper bounds found are positive–MLEs exist.**
  - **Took about 1 hour on the cluster of 64 processors.**
- **Issue: Can we scale to larger models and bigger tables.**

Summary

- What do we mean by sparseness:
  - 3 examples of contingency tables; 2 sparse.
- Statistical problems involving log-linear models:
  - Confidentiality & bounds for cell entries in tables.
  - Existence of MLEs for contingency tables.
- Role of computational algebraic and polyhedral geometry
- Exploring linkages between bounds and MLEs.
- Undone: Scaling up computations.

The End

- **Many related papers available for downloading at**
  - www.niss.org
  - www.stat.cmu.edu/~fienberg/DLindex.html
  - www.stat.washington.edu/adobra/html/research.html
  - www.stat.cmu.edu/~arinaldo

Algebraic Stat. References

- **Diaconis, P. & Sturmfels, B. (1998). *Ann. Statist.*, 26, *363-397.***
- **Dobra, A. and Fienberg, S. E. (2000). *Proc. Nat. Acad. Sci.*, 97, 11885–11892.**
- **Dobra, A. (2002). CMU PhD thesis.**
- **Eriksson, N., Fienberg, S. E., Rinaldo, A., and Sullivant, S. (2005). *J. Symbolic Comp.*, 41, 222–233.**
- **Fienberg, S. E. & Rinaldo, A. (2007). *J Stat. Plan. Infer.***
- **Fienberg, S. E. and Slavkovic, A. B. (2005). *Data Mining and Knowledge Discovery*, 11, 155-180.**
- **Geiger, D., Meek, C. & Sturmfels, B. (2006). *Ann. Statist.*, 34, 1463-1492.**
- **Slavkovic, A. B. (2004). CMU PhD thesis.**
- **Rinaldo, A. (2005). CMU PhD thesis.**
- **Sullivant, S. (2005). UC Berkeley PhD thesis.**

Contingency Table References

- **Bishop, Y.M. M., Fienberg, S.E. and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* MIT Press, Cambridge, MA.**
- **Fienberg, S.E. (1980). *The Analysis of Cross-Classified Categorical Data*. *2nd Ed*. MIT Press, Cambridge, MA.**
- **Haberman, S.J. (1974). *The Analysis of Frequency Data*. University of Chicago Press.**
- **Lauritzen, S.L. 1996. *Graphical Models*. Clarendon Press, Oxford.**

Warning: Bounds and Gaps

- **Bounds may not not be sufficient to understand degree of protection for confidentiality.**
  - **Gaps in range of values for specific cells are possible!**
- **Consider possible 6×4×3 tables:**
  - **Specify values for (1,1,1) cell: 0 and 2 (with gap at 1).**
  - **Can construct margins for which gaps are realized:**

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

deLoera & Ohn (2006)
J. Symb. Comp.