Confidentiality Protection and Utility for Contingency Table Data

Stephen E. Fienberg

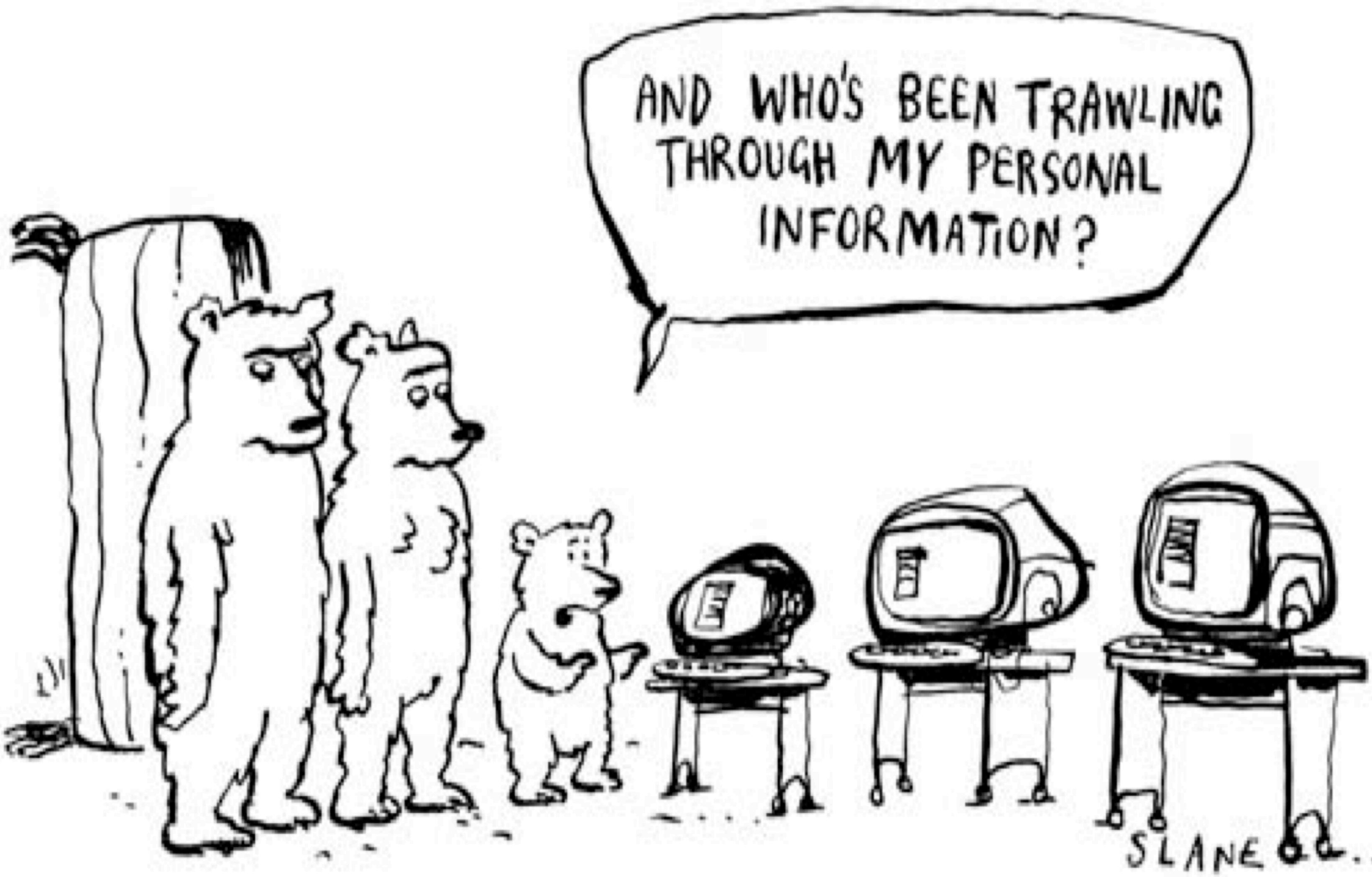
Department of Statistics, Machine Learning Department, and Cylab

Carnegie Mellon University

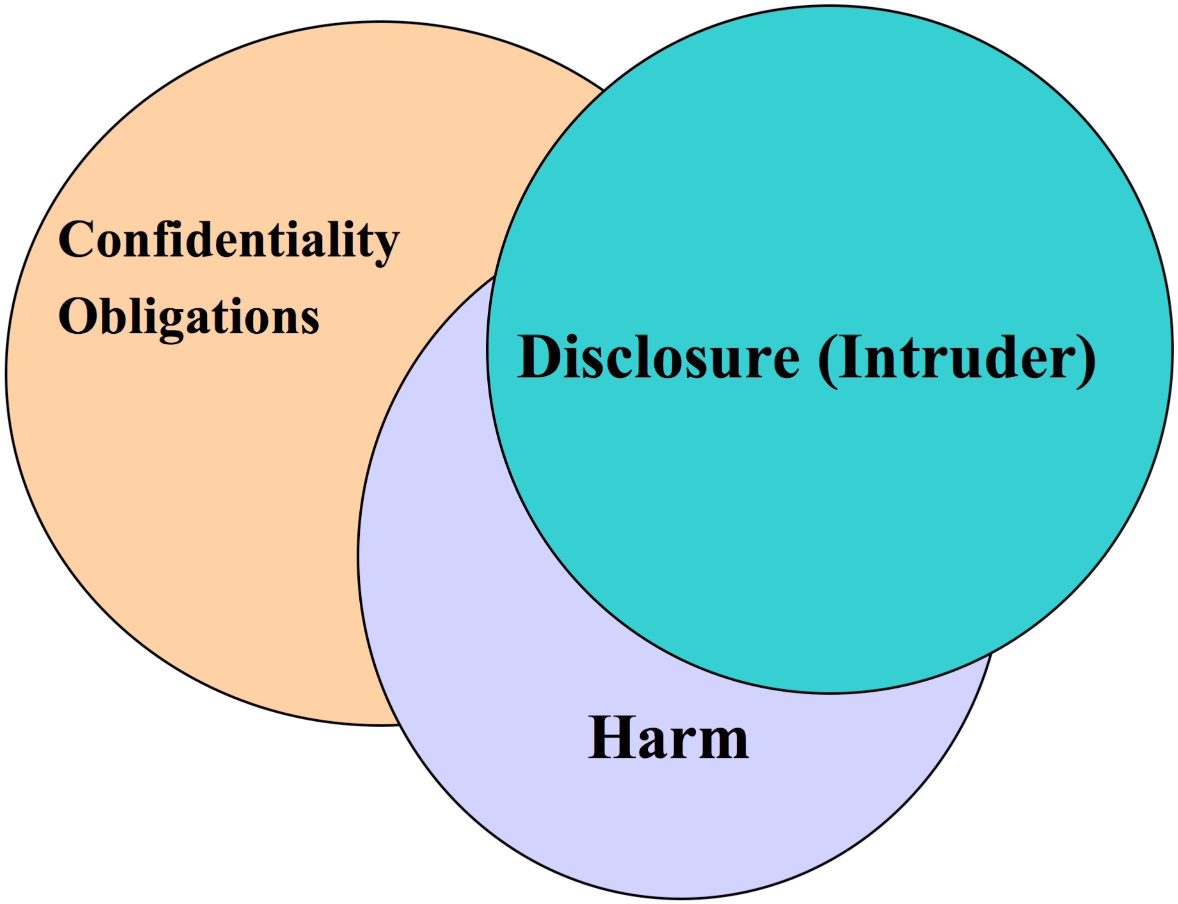
(Joint work with A. Dobra, A. Rinaldo, and Y. Zhou)

Outline

- Privacy and confidentiality
  - Focus individual data (not establishment data)
- Three examples and two problems:
  1. Bounds for cell counts in contingency tables given marginals.
  2. Maximum likelihood estimation for log-linear models.
  - How are they interrelated?
  - What are the mathematical tools? (No details!)
  - Scaling up computations for large sparse tables.

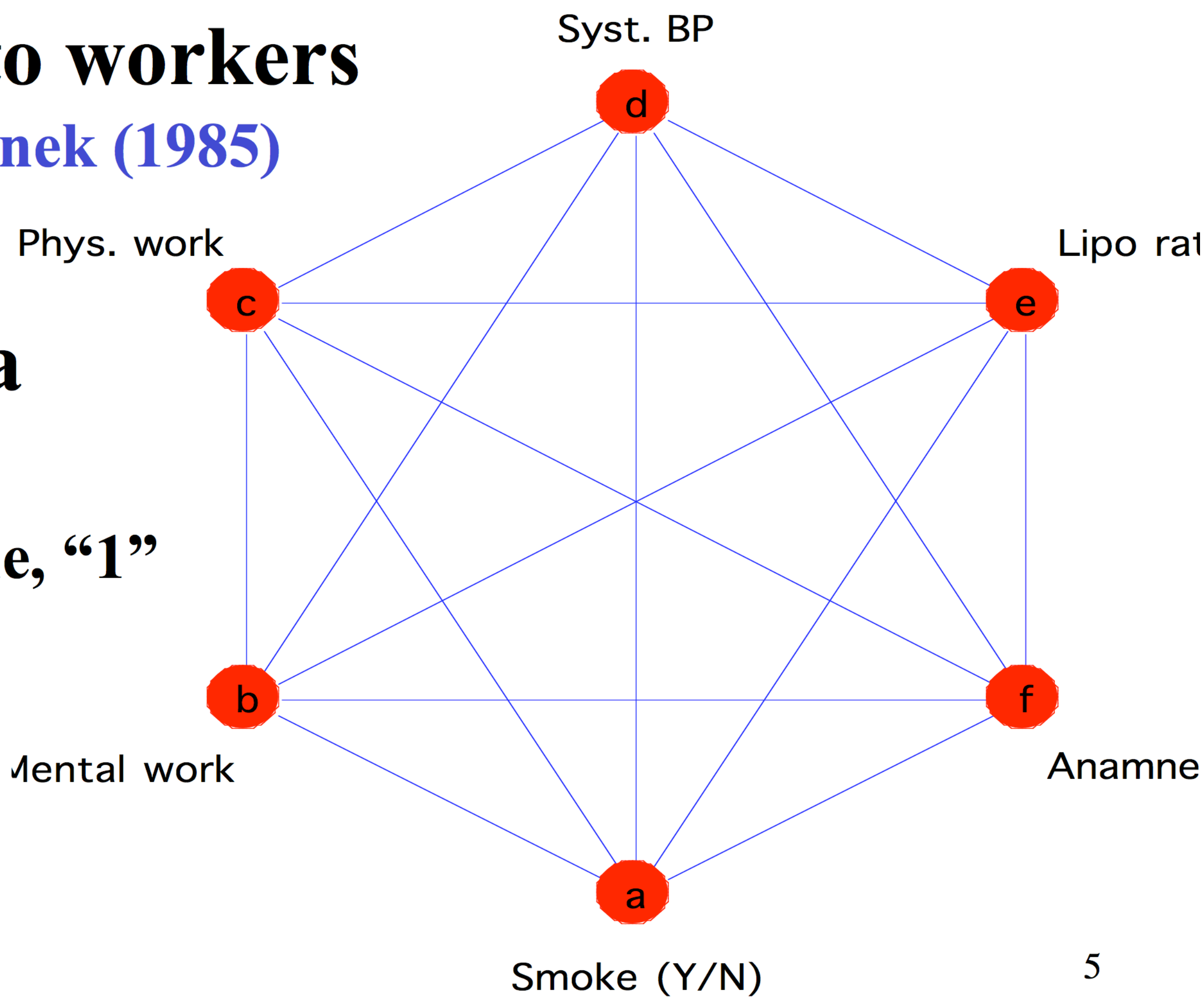

AND WHO'S BEEN TRAWLING THROUGH MY PERSONAL INFORMATION?
SLANE

Issues and Linkages



Ex. 1: Risk Factors for Coronary Heart Disease

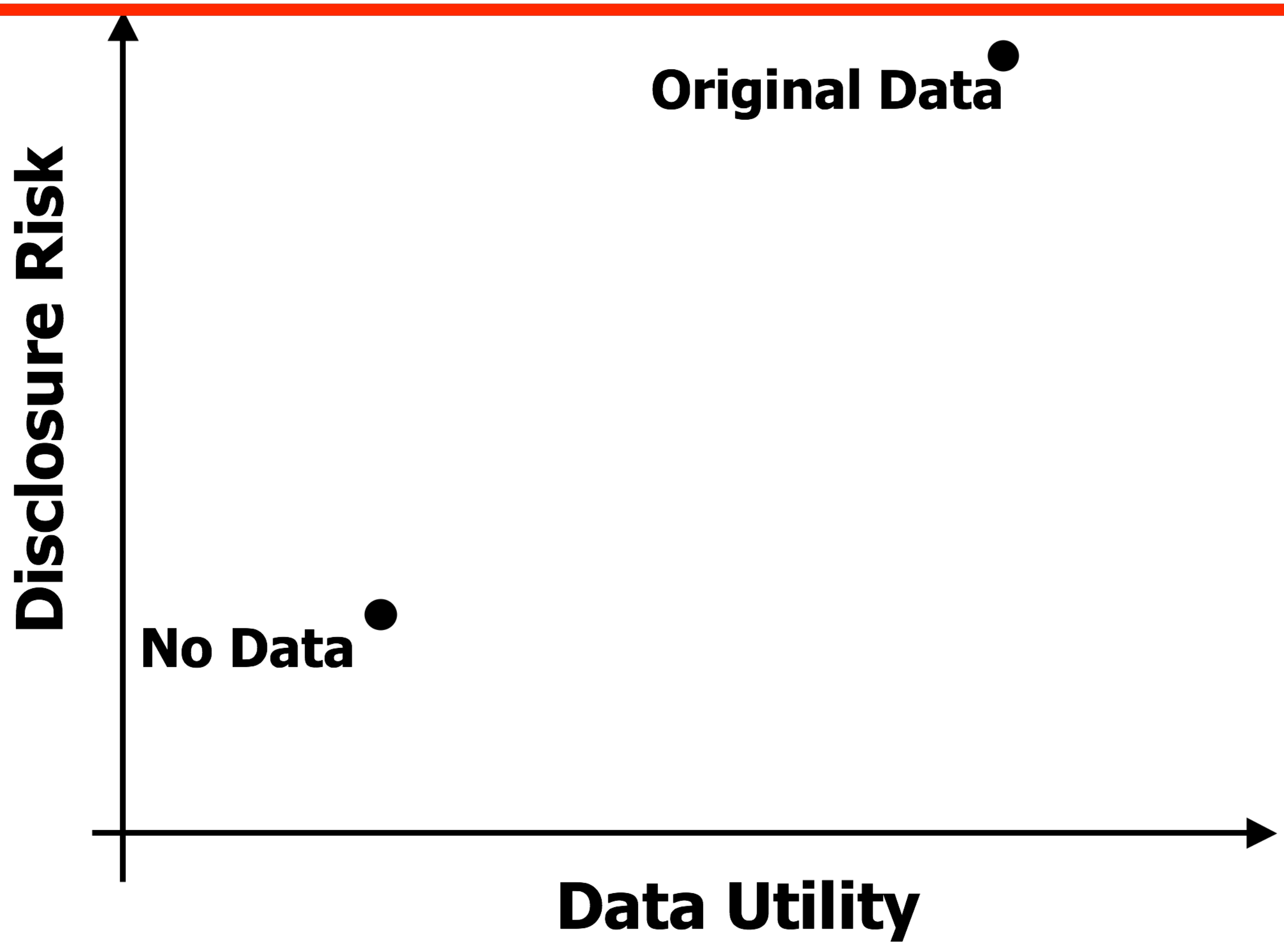
- 1841 Czech auto workers
  Edwards and Havanek (1985)
- 2^6 table
- population data
  - "0" cell
  - population unique, "1"
  - 2 cells with "2"



Ex. 1: The Data

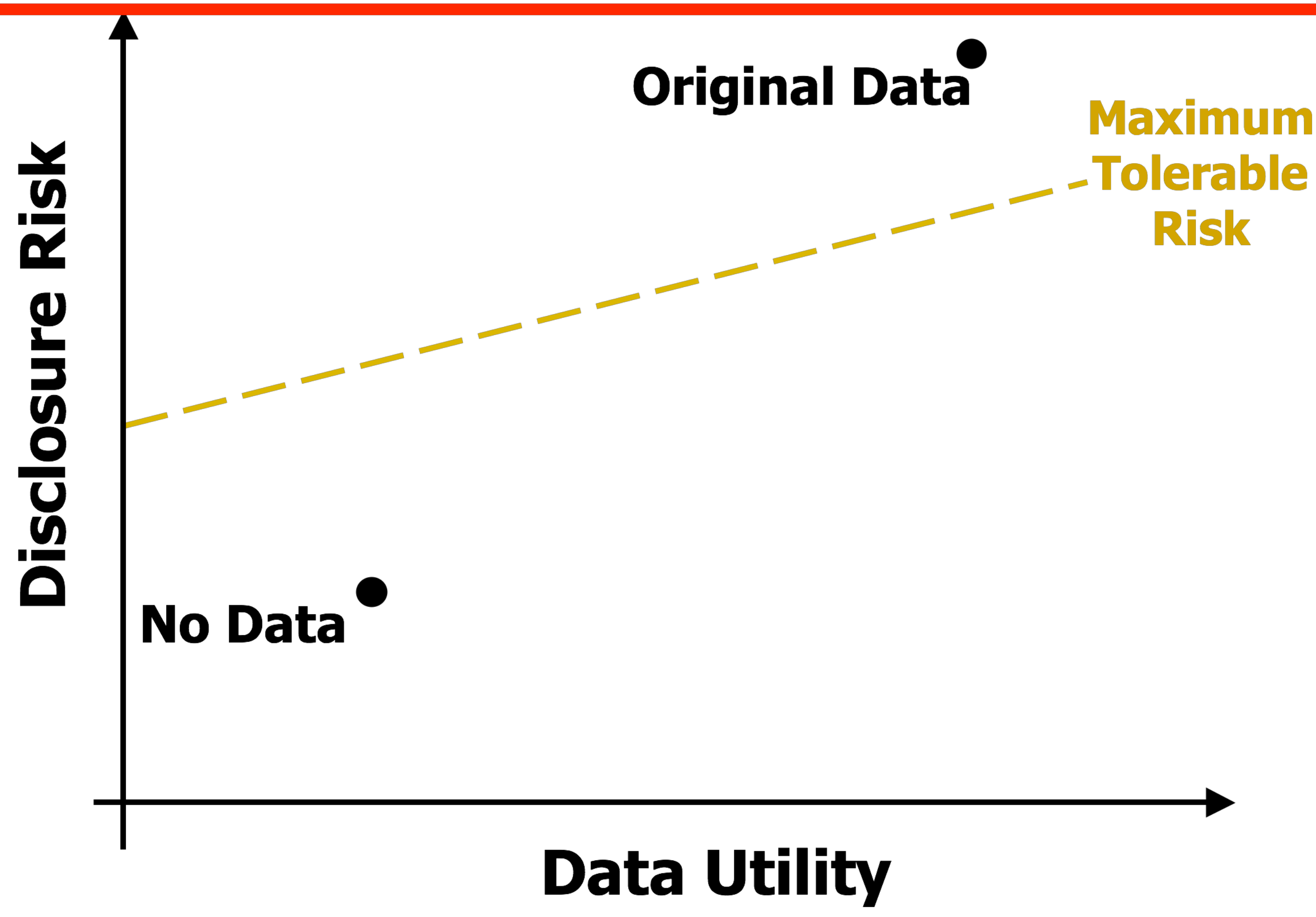
F	E	D	C	B	no		yes	
				A	no	yes	no	yes
neg	< 3	< 140	no		44	40	112	67
			yes		129	145	12	23
		≥ 140	no		35	12	80	33
			yes		109	67	7	9
	≥ 3	< 140	no		23	32	70	66
			yes		50	80	7	13
		≥ 140	no		24	25	73	57
			yes		51	63	7	16
pos	< 3	< 140	no		5	7	21	9
			yes		9	17	1	4
		≥ 140	no		4	3	11	8
			yes		14	17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



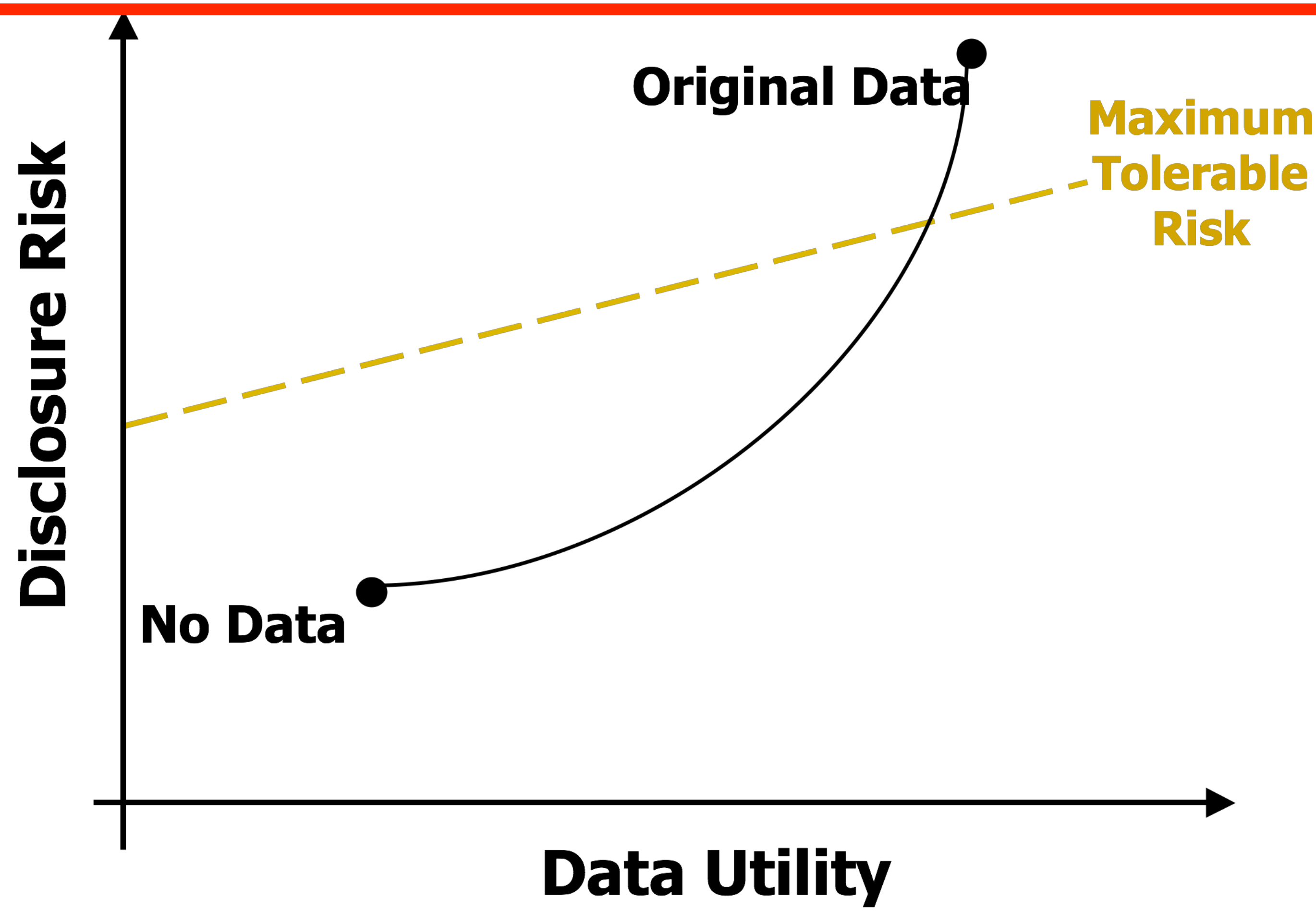
(Duncan, et al. 2004)

R-U Confidentiality Map



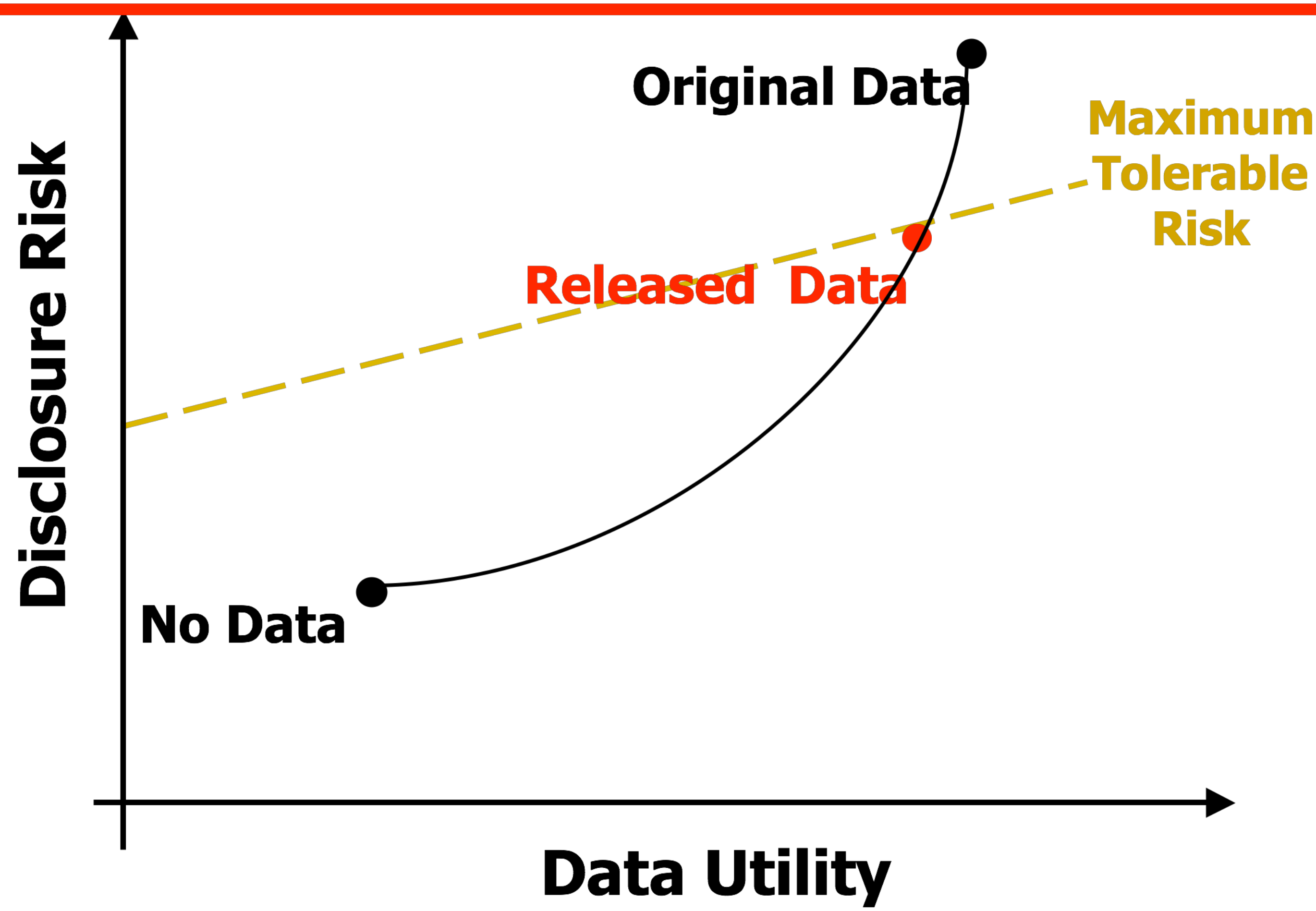
(Duncan, et al. 2004)

R-U Confidentiality Map



(Duncan, et al. 2004)

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:**
  - **Other types of sensible summary statistics?**
- **Risk concerned with small cell counts.**
  - **Assess using bounds given marginal totals.**

Ex. 2: Genetics Linkage

- **Data come from a barley milkdew experiment.**
  - **Edwards (1992). *CDSA*.**
- **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: Australia Census Data

- **10-dimensional highly sparse contingency table extracted from 1981 Australian population census (10 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-Way Fréchet Bounds

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

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

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

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

Multi-way Bounds

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

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

- ***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)

2^3 Table Given 2×2 Margins

x_{111}	x_{121}	x_{1+1}
x_{211}	x_{221}	x_{2+1}
x_{+11}	x_{+21}	x_{++1}

x_{112}	x_{122}	x_{1+2}
x_{212}	x_{222}	x_{2+2}
x_{+12}	x_{+22}	x_{++2}

x_{11+}	x_{12+}
x_{21+}	x_{22+}

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

Role of Log-linear Models?

- **For 2×2 case, lower bound is evocative of MLE for estimated expected value under independence:**

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

  - **Bounds correspond to log-linearized version.**
  - **Margins are *Minimal Sufficient Statistics (MSS)*.**

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

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

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

Log-linear Models (cont.)

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

$$\hat{m}_{ij+} = x_{ij+} \quad \text{for } i = 1,2,\ldots,I,\ j = 1,2,\ldots,J,$$

$$\hat{m}_{+jk} = x_{+jk} \quad \text{for } j = 1,2,\ldots,J,\ k = 1,2,\ldots,K,$$

$$\hat{m}_{i+k} = x_{i+k} \quad \text{for } i = 1,2,\ldots,I,\ k = 1,2,\ldots,K.$$

Existence of MLEs for 2×2×2 Table

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

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

x_{11+}	x_{12+}
x_{21+}	x_{22+}

Delta must be zero and MLE doesn't exist.

Two Other Three-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

Existence of MLEs

- **Linked to pattern of zeros.**
- **Discoverable by defining basis for models and using algebraic and polyhedral geometry.**
- **Examples discovered using algebraic software: *Polymake*.**
- **General theorems in Haberman (1974) and "constructively" in Rinaldo (2005):**
  - **Currently being implemented in C++ and R.**

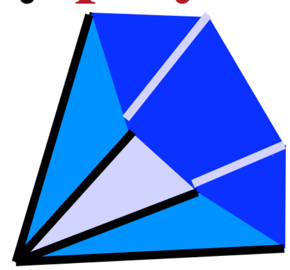
Two Faces of Algebraic Statistics

1. **Conditional Inference**: study and characterization of portions the sample space and, in particular, of all datasets having the observed margins ("exact distribution").
2. **Representation of a Statistical Model**: alternative, more powerful, description of the parameter space.

Its All About Geometry

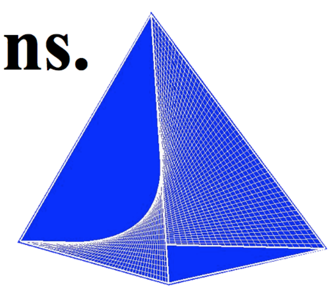
- **Polyhedral Geometry: virtually all data-related quantities can be described by polyhedra.**

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

Graphical & Decomposable Log-linear Models

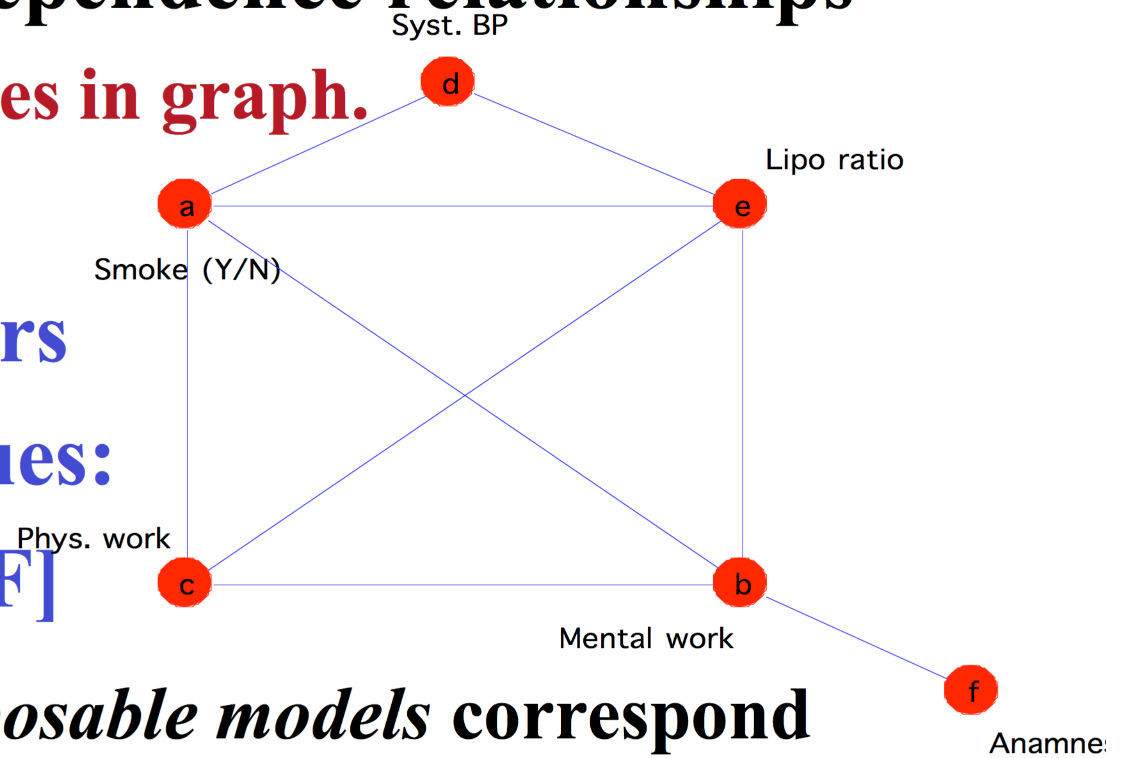
- *Graphical models:* **defined by simultaneous conditional independence relationships**
  - **Absence of edges in graph.**

Example 1:

Czech autoworkers

Graph has 3 cliques:

[ADE][ABCE][BF]



Decomposable models **correspond to triangulated graphs.**

Multi-way Bounds

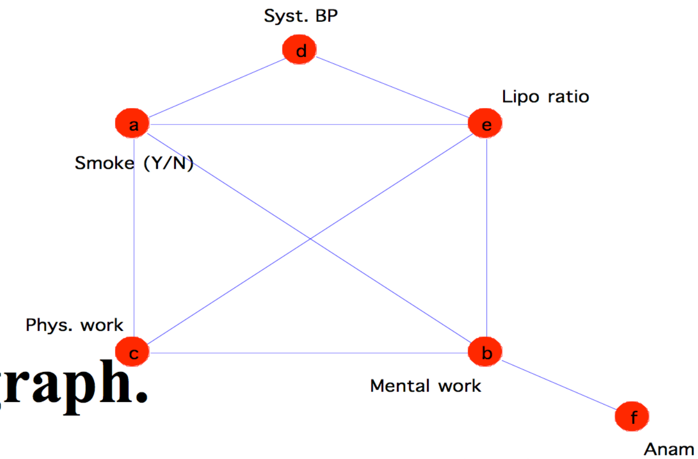
- **For decomposable log-linear models:**

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

- ***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)

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]

				B	no		yes	
F	E	D	C	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]

Example 1: What to Release?

Example 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].**

Example 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].**
- **Can actually show that release of everything else is "safe": i.e., we can release [ACDE][ABCDF][ABCEF][BCDEF][ABDEF]**

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:
  - Three examples of contingency tables
- Confidentiality & bounds for cell entries
- Existence of MLEs for contingency tables
- Role of computational algebraic geometry
- Exploring linkages between bounds and MLEs
- Undone: Scaling up computations

The End

- **Based in part on paper:**

  A. Dobra, S.E. Fienberg, A. Rinaldo, and Y. Zhou: "Confidentiality Protection and Utility for Contingency Table Data: Algorithms and Links to Statistical Theory."

- **Many related papers available for downloading at**

  http://www.niss.org

  www.stat.cmu.edu/~fienberg/DLindex.html

References

- **Dobra, A. and Fienberg, S. E. (2000). Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *PNAS*, 97, 11885–11892.**
- **Dobra, A. & Fienberg, S. E. (2003). In *Foundations of Statistical Inference: Proceedings of Shoresh Conference 2000* (Y. Haitovsky, H.R. Lerche, and Y. Ritov, eds.) 3–16.**
- **Eriksson, N., Fienberg, S. E., Rinaldo, A., & Sullivant, S. (2005). Polyhedral conditions for the nonexistence of the MLE for hierarchical log-linear models. *Journal of Symbolic Computation*, 41, 222–233.**
- **Fienberg, S. E. & Rinaldo, A. (2007). Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. *JSPI*, 137, 3430-3445.**
- **Rinaldo, A. (2006). On maximum likelihood estimation for log-linear models. Submitted for publication.**

Bounds for *k*-way Table Entries

- LP and IP approaches are NP-hard.
- Develop efficient methods for several special cases, exploiting linkage to statistical theory where possible:
  - Released margins corresponding to decomposable models have explicit formulae.
  - Margins corresponding to reducible graphs can be broken up into smaller problems.
  - Simple result for 2^k tables with release of all (k-1)-dimensional margins fixed.
- Generalized Shuttle algorithm (Dobra, 2001) for residual cases.

2×2 Table: The Data

Observed Counts

x_{11}	x_{12}
x_{21}	x_{22}

Released Margins

$$t = Ax$$

Design Matrix

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

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

$$\{x \in R^4_{\geq 0}, Ax = t\}$$

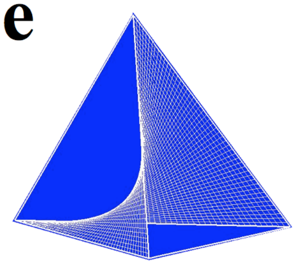
2×2 Table: The Model

p_{11}	p_{12}
p_{21}	p_{22}

- **We are interested in the distribution of the 4 cells in the table specified by the vector of log probabilities:**

$$\log(p_{11}, p_{12}, p_{21}, p_{22}) = A'\theta = (\theta_1 + \theta_3, \theta_1 + \theta_4, \theta_2 + \theta_3, \theta_2 + \theta_4)$$

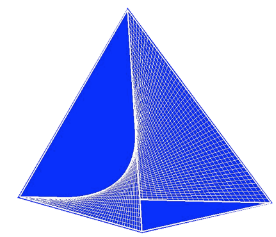
- **The set of all probability distributions for the model of independence need to satisfy one polynomial equation: $p_{11}p_{22}-p_{12}p_{21}=0$, and belong to surface of independence:**



Segre Variet

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 distribution.

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 ∈ kernel (A).

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