Optimal Tabular Releases from Confidential Data

Alan F. Karr, Adrian Dobra, Ashish P. Sanil
National Institute of Statistical Sciences
{karr,adobra,ashish}@niss.org

Problem Formulation

- **Underlying database:** Large contingency table containing counts
- **Releases:** Marginal sub-tables (low-dimensional; defined by subset of attributes)
- **Disclosure risk:** values of small count cells
- **Problem:** disclosure risk is cumulative as more sub-tables are released, i.e., queries interact

Responses

- Table servers
  - Dynamic assessment of disclosure risk
  - Problems: scale, user equity, release rules
- Optimal tabular releases (OTRs): static, one-time releases that *maximize data utility* subject to a *constraint on disclosure risk*

OTR Components -1

- K-way table
  - 2^K sub-tables, partially ordered by set inclusion of attributes
  - Many more possible releases R
- Data utility measure DU
  - Example: number of sub-tables in R
  - DU(R) = #{R}
- Disclosure risk measure DR
  - Example: Tightness of bounds on small cells
  - DR(R) = min{UB(*C*,R) - LB(*C*,R): 0 < #{*C*} <= 3}

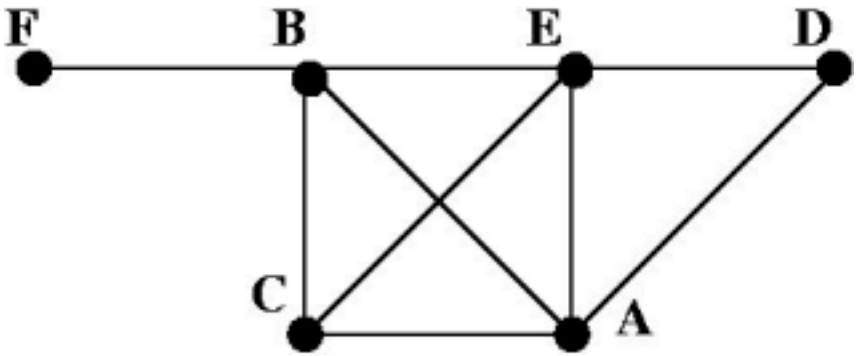
Optimization Formulation

$$\max DU(R)$$
$$\text{s.t. } DR(R) >= \alpha$$

- Not tractable computationally

Alternative 1: Decomposable Models

- Restrict R to correspond to a *decomposable statistical model*



- +: Bounds computable explicitly
- -: Optimization still has to be done using simulated annealing

Example: CPS Table

- 299,285 records
- 13 attributes
  - 8192 sub-tables
- 2,592,000 cells, of which
  - 41,672 are non-zero
  - 22,996 are 1
  - 6,435 are 2

Results

α	Released Frontier	#{R*}	% of Tables Released
3	2 7-way 5 6-way	351	4.25
4	2 7-way 3 6-way 2 5-way	319	3.89
5	4 6-way 3 5-way	239	2.92

Alternative 2: Ordered Marginals The Idea

- Order marginals in a particular way that represents *individual riskiness*
- Add to the overall release R in order of decreasing individual riskiness until *cumulative risk* of R is unacceptable

Ordered Marginals: The Details

- Cell *C* is *at risk* if $0 < \#\{C\} <= 2$
- Release R is *bad* at bound width *w* if

$$\min\{\text{UB}(C,R) - \text{LB}(C,R): C \text{ is at risk}\} <= w$$

- *Most parsimonious release* for sub-table *T* is MPR(*T*) = {*T*, all 1-d sub-tables for attributes not in *T*} (which is decomposable)
- *T* is bad at width *w* if MPR(*T*) is bad at width *w*
- *Critical width* of *T* – CW*(T)* = max{*w*: *T* is bad at *w*} – measures individual riskiness

Example

- Table on disease in Czech auto workers
  - 1841 records
  - 6 variables A, B, C, D, E, F
  - 3 at risk cells: 1 with count 1, 2 with count 2
- Optimal release
  - *All* sub-tables other than [ABCDEF], [ACDEF], [ABCDE], [ABDEF]
  - Largest decomposable release contains 48 sub-tables

Critical Width Visualization