Preserving Positivity (and other Constraints?) in Released Microdata

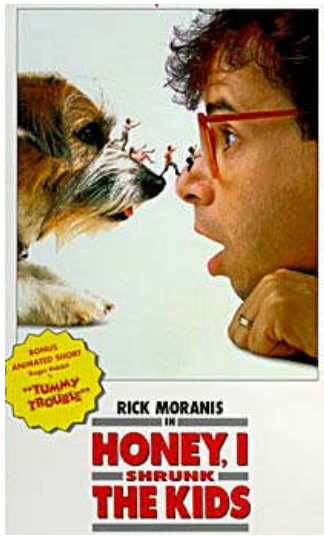
Alan Karr
11/18/05

Problem

- *O* = original microdata
  - Constrained to satisfy $o_{ij} >= 0$ for all data records *i* and (some or all) attributes *j*
- Want masked data release *M* that
  - Satisfies same positivity constraints
  - Has high utility
  - Has low disclosure risk

Motivating Example

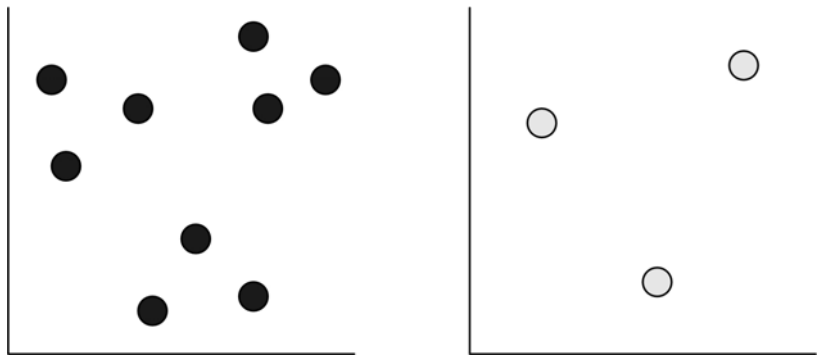
- O = original data
- M_1 = MicZ applied to O
  - Conceptually, data have shrunk
- $D = O - M_1$
- N = normally distributed noise with
  - Mean 0
  - Cov(N) = Cov(D)
- $M_2 = M_1 + N$
  - Works well for utility and risk
  - Does not preserve positivity



Which Methods Preserve Positivity?

Method	**Yes/No**	**Comments**
Mic*	Yes	Edge effects may be significant
Swapping	Yes	
Additive noise	No, in general	Noise distributions that preserve positivity induce edge effects
Transformation	No, in general	
Synthetic data	Possible	If model preserves positivity

The Problem with Mic*



Some (1/2, ¼, 1/1048576)-Wit Ideas

- Transform data to remove constraints (e.g., take logs), do SDL on transformed data, and untransform
  - Microaggregation becomes “weighted”
  - How to choose noise distribution?
- Use rejection sampling for additive noise
  - M+N|M+N>= 0
  - Does it restore full covariance?

More Ideas

- Microaggregation with weights: move points near edges less
  - May be bad for risk
  - No longer preserves first moments
- ?????

Other Questions

- More complex constraints
  - $x_1 <= x_2$ (time precedence; net income <= gross)
  - $x_1 + x_2 <= x_3$
  - Are all “convexity-like” constraints alike?
- Multiple constraints
  - Methods need to be scalable
- “Soft” constraints