Overview of Statistical Disclosure Limitation

Statistical Methods for Data Privacy, Confidentiality and Disclosure Limitation

Aleksandra B. Slavković
Department of Statistics, Penn State University
sesa@stat.psu.edu

Data Privacy Course -- Aug 30, 2007
Privacy & Breaches of Privacy

- Expansion of government and company databases & growing use of web and mobile devices had led to increase of sensitive information in the public domain.

- Consumer & Workplace Privacy
- Internet/ Computer Security & Privacy
- Medical Privacy
- Genetics
- Surveillance & Wiretapping, and “Homeland” security
- Cryptography
- Data Theft

- Various definitions of privacy


- Focus on privacy & confidentiality of data collected into statistical databases, then utilized for analyses and released
Agencies seek to release confidential data to the public, and to improve analyses by sharing their confidential data, but via strategies that

- do not reveal identities or sensitive attributes,
- are useful for a wide range of analyses,
- are easy for analysts and agencies to use

Data Utility: Are released data (often tables) useful for statistical inference?

Risk of Disclosure: Are the results of their analysis safe to be released?
Why do we care about privacy & confidentiality?

- **Ethical:** Keeping promises

- **Legal:** Required under law
  - Privacy is the right to keep one’s personal information out of the public view
  - Confidentiality is the dissemination without public identification

- **Pragmatic:**
  - Respondents may not provide data
  - Respondents may provide inaccurate data

- Context dependent definitions of privacy: Consumer & Workplace Privacy, Internet/Computer Security & Privacy, Medical Privacy, Genetics, Surveillance & Wiretapping, and “Homeland” security

Ref: http://www.amstat.org/comm/cmtepc/index.cfm?fuseaction=main
Privacy -- Context dependent

- Privacy is an individual’s right to keep one’s personal information (data) out of public view.

- "Informational privacy encompasses an individual's freedom from excessive intrusion in the quest for information and an individual's ability to choose the extent and circumstances under which his or her beliefs, behaviors, opinions, and attitudes will be shared with or withheld from others.” Report of the Committee on National Statistics' Panel on Confidentiality and Data Access, Duncan et al., 1993. Private Lives and Public Policies, Washington, DC: National Academy Press, p. 22.

Ref: http://www.amstat.org/comm/cmtepc/index.cfm?fuseaction=main
Confidentiality -- Context dependent

- "[Confidential should mean that dissemination] of data in a manner that would allow public identification of the respondent or would in any way be harmful to him is prohibited and that the data are immune from legal process . . ." Report of the Committee on National Statistics' Panel on Confidentiality and Data Access, Duncan et al., 1993. *Private Lives and Public Policies*, Washington, DC: National Academy Press, p. 23.

- Individuals and organizations

Ref: http://www.amstat.org/comm/cmtepc/index.cfm?fuseaction=main
Confidentiality History in the U.S.A.

- Statement by President Taft 1910 for protecting census data
- 1929 Census Act
- Privacy Act of 1974
- 2001 U.S. Patriot Act
- 2003 CIPSEA law
Definition of Disclosure

- Tore Dalenius from Working Paper No. 2, pp. 7 and 9:

  “If the release of the statistics S makes it possible to determine the value [of confidential statistical data] more accurately than is possible without access to S, a disclosure has taken place.”
What constitutes disclosure?

- **Identity disclosure** = when a specific person’s record can be found in a released file
  - The richest person in State College participates in a survey. We can identify the person’s record.

- **Attribute disclosure** = when sensitive information about a specific person is revealed through the released file, sometimes with additional knowledge
  - Can someone figure out how much income the richest person in SC has even without identifying the personal record, may be via record linkage or some other additional information

- **Inferential disclosure** = if from the released data one can determine the value of some characteristic of an individual more accurately than otherwise would have been possible
  - The aggregate information on a survey has been released. You are one of the 3 richest people in SC. Now you have 50-50 chance of identifying the others and their income.

Public Access & Unique Identification

- Uniqueness or small counts in the sample or the population can lead to unique identification

- As the amount of publicly available data increases so does the threat to our privacy and confidentiality

- Record linkage methodologies
  - More sensitive information on income, voting, sexual preferences, etc…

- Sweeney (2000): Date of birth, gender, 5-digit ZIP
  - Likely unique identification of 87% U.S. population

Data protection approaches

- Not sharing/releasing data
- Sharing data via restricted access or altering data in some way

- Cryptography

- Privacy-Preserving Data Mining (PPDM)
  - Data mining algorithms and their security
  - Aggregating multiple sources into a single warehouse
  - Mining multiple distributed databases
  - Record linkage and privacy protection

- Statistical Disclosure Limitation/Control (SDL)
  - Data masking techniques
  - Introduce bias and variance
Goal of Statistical Disclosure Limitation

- Preserving confidentiality

- Providing access to useful statistical data, not just few numbers
  - Inferences should be the same as if we had original complete data
    - Requires ability to reverse disclosure protection mechanism, not for individual identification, but for inferences about parameters in statistical models (e.g., likelihood function for disclosure procedure)
  - Sufficient variables and statistics to allow for proper multivariate analyses
  - Ability to assess goodness of fit of models
    - Need most summary information, residuals, etc

- Strike a balance between *data utility* and *disclosure risk*
The R-U Confidentiality Map

- Duncan et al. (2001)
- Bayesian framework (Trottini (2003))

Two possible releases of the same data

Disclosure Risk

Data Utility
The R-U Confidentiality Map

For many disclosure limitation methods, we can choose one or more parameters that we will vary.
Protecting confidentiality

- Restricted access
  - Conditions are imposed on who may access the data, for what purpose, at what location, which variables may be accessed, what can be published, etc.
  - Sworn Employees at Census Bureau
  - Licensed Researchers at National Center for Education Statistics or at External Sites such as Research Data Centers

- Restricted data
  - Data will come in two formats
    - Microdata files
    - Tabular data
  - Restricting the amount of information in released files via various statistical techniques

- Combination of the two, e.g. online access
  - Online queries & Remote access servers
  - Secure computation & Distributed databases
**Microdata**

- **Microdata**: a set of records containing information on individual respondents
  - $X$ is a data matrix with $n$ records (individuals) and $p$ attributes

- Suppose *you* are supposed to release microdata, for the public good, about individuals. The data include:
  - Name, Address, City, Occupation, Age, Sex, Income, Voting record

- Typically remove “identifiers”
  - HIPPA says it’s OK as long as the researcher promises not to attempt re-identification.
  - Is this reasonable?
  - City=State College, Occupation=CEO, Age=45, Sex=Male, Income=$300,000 Voting=Bush
Tabular Data

- Contingency tables: cross-cross-classifies individuals by attributes
- Publicly available data as **marginal** and **conditional** tables

- Strike a balance between *data utility* and *disclosure risk*
  - *Utility* tied to usefulness of *marginal totals* & *log-linear models*
  - *Risk measure* is ability to identify small cell counts (e.g. via bounds, Dobra (2002))

<table>
<thead>
<tr>
<th>County</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
<th>Very High</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha</td>
<td>15</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>Beta</td>
<td>20</td>
<td>10</td>
<td>10</td>
<td>15</td>
<td>55</td>
</tr>
<tr>
<td>Gamma</td>
<td>3</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>Delta</td>
<td>12</td>
<td>14</td>
<td>7</td>
<td>2</td>
<td>35</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>35</td>
<td>30</td>
<td>20</td>
<td>135</td>
</tr>
</tbody>
</table>

- What can we release about this data to achieve the balance?

Delinquent Children by County & Education Level
Statistical Disclosure Limitation (Control) methods

- Apply to microdata and/or tabulated data before release

- **Data masking**: Transform the original data (matrix $X$) to the disseminated data ($Y$)
  - $Y=AXB + C$
  - $A$=record transformation, $B$=attribute transformation, $C$=noise addition

- Traditional approaches
  - Aggregation: Rounding, Topcoding & Tresholding
  - Suppression, e.g., cell suppression
  - Data Perturbations
  - Data Swapping

- Modern approaches: Sampling and Simulation techniques
  - Synthetic data
  - Remote access servers
  - Secure computation
  - Partial information releases
Aggregation: Rounding & Topcoding

- A way of aggregating data values

- Rounding the reported income to the closest $1000

- Topcoding or thresholding
  - Grouping similar income values together such as all individuals over $500,000

- Works ok for large samples and middle of the distribution
- Problems in the upper tails due to high difference in income
- If intervals are not uniform reporting the middle value will be biased
- If that are cross-classified we may get different tables based on the rounding scheme (Srinivasan, N. http://www.fhwa.dot.gov/ctpp/sr0404.html)
- Possible get different answers to the same questions
Suppress Sensitive Cells & Others

- Not releasing sensitive fields of the microdata file or a table
- Mostly applied to tabular data, e.g., cell suppression -- Cox (1980, 1995, 1999)

<table>
<thead>
<tr>
<th>County</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
<th>Very High</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha</td>
<td>15</td>
<td>p</td>
<td>s</td>
<td>p</td>
<td>20</td>
</tr>
<tr>
<td>Beta</td>
<td>20</td>
<td>10</td>
<td>10</td>
<td>15</td>
<td>55</td>
</tr>
<tr>
<td>Gamma</td>
<td>3</td>
<td>10</td>
<td>s</td>
<td>p</td>
<td>25</td>
</tr>
<tr>
<td>Delta</td>
<td>12</td>
<td>s</td>
<td>7</td>
<td>p</td>
<td>35</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>35</td>
<td>30</td>
<td>20</td>
<td>135</td>
</tr>
</tbody>
</table>

“Swiss cheese”

Issue with correlation structure

NP complete

Can get cell intervals via linear programming
Data Perturbation

- Perturbation generally means altering values before releasing

- Adding random noise to certain fields in microdata file
  - E.g., adding some noise from a normal or lognormal distribution to the income data

- Better works for numeric data than categorical data

- Perturbation of the cells in the table need to provide a consistency between the original table and a perturbed table
  - Random and controlled rounding
  - Controlled tabular adjustment, e.g., “safe but sufficiently close value”
  - Cyclic perturbation (Roehrig & Duncan (2003))
## Data Perturbation: Controlled Rounding

Every cell is a multiple of some rounding bases, e.g. base 3
- Easy for 2-D tables, sometimes impossible for 3-D
- Changes the marginal counts (distributions), as well as the observed proportions, thus may produce faulty inference and data mining pattern matching

<table>
<thead>
<tr>
<th>County</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
<th>Very High</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha</td>
<td>15</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td>Beta</td>
<td>21</td>
<td>9</td>
<td>12</td>
<td>15</td>
<td>57</td>
</tr>
<tr>
<td>Gamma</td>
<td>3</td>
<td>9</td>
<td>9</td>
<td>3</td>
<td>24</td>
</tr>
<tr>
<td>Delta</td>
<td>12</td>
<td>15</td>
<td>6</td>
<td>3</td>
<td>36</td>
</tr>
<tr>
<td>Total</td>
<td>51</td>
<td>33</td>
<td>30</td>
<td>21</td>
<td>135</td>
</tr>
</tbody>
</table>
Data perturbation: Data swapping

- Proposed by Dalenius & Reiss (1982), and is used by several agencies
  - The Census Bureau for decennial census releases

- Exchange a fraction of data at random between two respondents
  - Small fraction swapped
  - Reduces or destroys correlations

- Doubly random swapping more recently Karr, A. et al. (2006)
  - Pick randomly two records to swap and then swap the random pairs of records
  - Better preserves correlations

- No generalizations for more complex structures and k-way tables

- Data swapping toolkit for categorical data at NISS:
  http://www.niss.org/software/dstk.html
**Example: Data Swapping**

<table>
<thead>
<tr>
<th>Location</th>
<th>Age</th>
<th>Sex</th>
<th>Candidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pittsburgh</td>
<td>Young</td>
<td>M</td>
<td>Bush</td>
</tr>
<tr>
<td>Pittsburgh</td>
<td>Young</td>
<td>M</td>
<td>Gore</td>
</tr>
<tr>
<td>Pittsburgh</td>
<td>Young</td>
<td>F</td>
<td>Gore</td>
</tr>
<tr>
<td>Pittsburgh</td>
<td>Old</td>
<td>M</td>
<td>Gore</td>
</tr>
<tr>
<td>Pittsburgh</td>
<td>Old</td>
<td>M</td>
<td>Bush</td>
</tr>
<tr>
<td>State College</td>
<td>Young</td>
<td>F</td>
<td>Bush</td>
</tr>
<tr>
<td>State College</td>
<td>Young</td>
<td>F</td>
<td>Gore</td>
</tr>
<tr>
<td>State College</td>
<td>Old</td>
<td>M</td>
<td>Gore</td>
</tr>
<tr>
<td>State College</td>
<td>Old</td>
<td>F</td>
<td>Bush</td>
</tr>
<tr>
<td>State College</td>
<td>Old</td>
<td>F</td>
<td>Gore</td>
</tr>
</tbody>
</table>

Unique on location, age and sex

Find a match in another location

Flip a coin to see if swapping is done
Recent methods: Simulation & Sampling

- Digital Government Project I & II at NISS
- World “without original microdata”

- Synthetic & Partially synthetic data uses Bayesian methodology
  - Raghunathan, Reiter, and Rubin (2003, JOS)
  - Reiter (2003, Surv. Meth., 2005, JRSS A)

- Partial data releases for tabular data & logistic regression with algebraic tools
  - Dobra et al. (2003)

- Remote access servers
  - Rowland (2003, NAS Panel on Data Access).

- Secure computation
  - Benaloh (1987, CRYPTO86)
  - Karr, Lin, Sanil, and Reiter (2005, NISS tech. rep.)
Synthetic Data

- Fully synthetic data
  - Rubin (1993, *JOS*) -- create multiple, fully synthetic datasets for public release such that no unit in released data has sensitive data from actual unit in population.

- Partially synthetic data
  - Little (1993, *JOS*) -- create multiple, partially synthetic datasets for public release such that released data comprise mix of observed and synthetic values

And for both it should hold that
- Released data look like actual data.
- Statistical procedures valid for original data are valid for released data.

- Procedure
  - Randomly sample new units from the sampling frame.
  - Impute the variables for new units using model fit from observed data, i.e., posterior predictive distribution
  - Repeat multiple times and release multiple datasets.
Synthetic Data

- Does it guarantee confidentiality?
  - Not fully but can significantly lower the risk
  - Because a synthetic data record is not any respondent’s actual data record, identity disclosure is directly impossible
  - Attribute disclosure is still possible
  - With extreme values, it may still be possible to re-identify a source record
  - Some simulated individuals may have data values virtually identical to original sample individuals, so the possibility of both identity and attribute disclosure remain (Fienberg 1997, 2003)

- Is it valid for inferences?
  - It depends on the model used to generate the data
    - Not unless we are careful in how it is synthesized
    - Choosing the models and posterior distributions can be tricky

- Raghunathan, Reiter, and Rubin (2003, *JOS*)
Remote access servers

- Server is a system that allows users to submit queries for output from statistical analyses of microdata, but does not give direct access to microdata.

- Queries: Users request results from fitting a statistical model to the data, e.g., linear regression, logistic regression, a table.

- Response from the server:
  - Answerable query: model output & ideally model diagnostics.
  - Unanswerable query: no results.

- Table Servers / Model Servers
  - Prototype at http://www.niss.org/dg/table-server.html

- Challenges
  - Non-statistical: Operation costs, server security, etc.
  - Statistical:
    - Disclosure risks from smart queries (e.g., subsets, transformations).
    - Inferential disclosure risks
    - Enabling complex model fitting.
  - Limit to the number of releases
“Secure” computation

- Setting:
  - A “global” database partitioned among multiple agencies
  - Agencies want to perform statistical analysis on the “global” database, but are not willing or unable to combine the databases
  - Proprietary and/or confidential data

- Goals:
  - Share “valid” statistical analysis as if had the “global” database but without any actual data integration in order to improve each own analysis
  - Each agency protects its own data from the other agencies
  - Do not reveal identities or sensitive attributes

- Type of partitioning:
  - Horizontally Partitioned = Agencies have different records but same variables.
  - Purely Vertically Partitioned = Agencies have same records but different variables.
  - Partially Overlapping, Vertically Partitioned = Agencies have different records and different variables, with some common records and variables.
“Secure” computation

- Tools from the Computer Science literature based on secure multiparty computation

- Use secure summation or matrix manipulation procedure to share
  - shares sums without sharing data
  - allows linear & logistic regressions, clustering, classifications
  - assumes semi-honesty

- Issues:
  - How to choose a model a priori?
  - How to account for data quality, e.g., reconcile measurement errors?
  - Partially overlapping databases with measurement error

- Benaloh (1987, CRYPTO86) 
- Karr, Lin, Sanil, and Reiter (2005, NISS tech. rep.)
“Secure” Summation

Goal: compute \( f(v_1,\ldots,v_k) = \sum v_k \)

Algorithm:
- Agency 1 (A₁): Creates a large random number R and adds to \( v_1 \), then passes \( R + v_1 \) to A₂
- Agency 2 (A₂): Adds \( v_2 \), and passes \( R + v_1 + v_2 \) to A₃
- ...
- Agency k (Aₖ): Adds \( v_k \), and passes \( R + v_1 + v_2 + \ldots + v_k \) back to A₁

- A₁ subtracts R from \( R + \sum v_k \), and shares the result with ALL other agencies.

Note:
- Need a good random number
- Semi-honesty crucial:
  - No collusions
  - Companies use correct data only for the agreed computation
  - Retain intermediate results
Partial data releases for tabular data

- Goal: Determine safe releases in terms of arbitrary set of marginals and/or conditionals
  - Assume data reported without error, compatible margins and conditionals, and unweighted counts
  - Currently we are exploring extensions to odds ratios, and to these assumptions

- Given the information
  - if we can uniquely identify the joint distribution that is a full disclosure
  - If we have a partial specification of the joint distribution, we may use bounds and/or distributions over the space of possible solutions to assess the risk of disclosure and the data utility

- Using tools from linear/integer programming, specification of joint distributions and algebraic statistics

- Close links to perturbation, data swapping, synthetic data and remote access servers
- Dobra et al. (2003), Slavkovic (2004, 2005)
What is Algebraic Statistics?

- Algebraic statistics exploits the use of polynomial algebra and algebraic geometry for statistical inference.

- Polynomial rings, ideals, and algebraic varieties give a way of representing tables of counts and discrete probability distributions.

- The philosophy of Algebraic Statistics is: “Statistical models are algebraic varieties”.
  A statistical model is the zero sets of system of polynomial equations in several unknowns.
Two Sides of Algebraic Statistics

1. **Representation of a Statistical Model**: alternative description of the parameter space.
   - Log-Linear Models and in general Exponential Families of Discrete Distributions are **toric varieties** (defined by binomial equations).
   - Affect the quality of inference
   - The margins are sufficient statistics and inference is **conditional** on the observed margins and **MLE**-based.

2. **Conditional Inference**: study and characterization of portions of the sample space and, in particular, of all datasets (i.e., tables) having the observed margins and/or conditionals
   - **counting** and/or **sampling** over the possible space of tables
   - **compute** sharp upper and lower **bounds** on cell entries
   - Important for assessment of risk of disclosure.
   - “exact distributions” of the sample statistics
### Delinquent Children by County & Education Level

#### Education Level of Head of Household

<table>
<thead>
<tr>
<th>County</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
<th>Very High</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha</td>
<td>15</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>Beta</td>
<td>20</td>
<td>10</td>
<td>15</td>
<td></td>
<td>55</td>
</tr>
<tr>
<td>Gamma</td>
<td>3</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>Delta</td>
<td>12</td>
<td>14</td>
<td>7</td>
<td>2</td>
<td>35</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>35</td>
<td>30</td>
<td>20</td>
<td>135</td>
</tr>
</tbody>
</table>

18,272,363,056 tables have our margins (De Loera & Sturmfels).

**Data Source:** OMB Statistical Policy Working Paper 22 & S. Roehrig
Delinquent Children by County & Education Level

- Release observed conditional frequencies

\[ P(Education \mid County) = \begin{pmatrix} 0.750 & 0.050 & 0.150 & 0.050 \\ 0.364 & 0.182 & 0.182 & 0.273 \\ 0.120 & 0.400 & 0.400 & 0.080 \\ 0.343 & 0.400 & 0.200 & 0.057 \end{pmatrix} \]

- IP: no feasible solution
- LP relaxation bounds:

<table>
<thead>
<tr>
<th>County</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
<th>Very High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha</td>
<td>15  [15, 74.6]</td>
<td>1 [1, 4.97]</td>
<td>3 [3, 14.9]</td>
<td>1 [1, 4.97]</td>
</tr>
<tr>
<td>Beta</td>
<td>20  [1.99, 30.8]</td>
<td>10 [1, 15.5]</td>
<td>10 [1, 15.5]</td>
<td>15 [1.5, 23.2]</td>
</tr>
<tr>
<td>Gamma</td>
<td>3   [1.5, 11.0]</td>
<td>10 [5, 36.8]</td>
<td>10 [5, 36.8]</td>
<td>2 [1, 7.36]</td>
</tr>
</tbody>
</table>

- Is it safe to release this conditional? NO, only 1 table!

IP=Integer Programming
LP=Linear Programming
Example: Clinical trial data (Koch (1983))

- Effectiveness of an analgesic drug measured at two different centers, and two different health conditions, with two treatments (1=Active, 2=Placebo), and responses (1=Poor, 2=Not Poor).

<table>
<thead>
<tr>
<th>Center</th>
<th>Status</th>
<th>Treatment</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Poor</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Active</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Placebo</td>
<td>11</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Active</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Placebo</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Active</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Placebo</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>Active</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Placebo</td>
<td>6</td>
</tr>
</tbody>
</table>

- Possible margins release for well-fitted models:
  
  \[[CST][CRT][CTR]\]  \[[CST][CRT][TR]\]  \[[CST][CRT]\]
Conditional Inference: sampling with Markov bases

- It is possible to perform a random walk on the space of all the tables with a given set of margins (or conditionals).
  - It requires the identification of moves: integer valued vectors in the kernel of $A$ that, added to the current table, will produce a table with same margins.

- **Markov Bases:** minimal set of moves that preserve connectedness in the fiber.
  - Computed with algebraic software, but this amounts to finding the minimal generators of a set of polynomials defined by $A$:

  $$I = \left\langle x^u^+ - x^u^-, \forall u \in \text{kernel}(A) \cap \mathbb{N}^I \right\rangle$$

- Using Markov Bases, it is possible to build a Gibbs sampler that can be used to explore the fiber and estimate the posterior distribution of the tables given the margins and the distribution of statistics over the fiber (usually Likelihood Ratio, Pearson’s $\chi^2$).
  - Known for the margins generalized hypergeometric distribution
  - For the conditionals, no generalization yet.
Conditional Inference: Optimizing

- **Table entry data security problem**: compute sharp lower and upper bounds for the entries in a table with given margins, conditionals and/or odds-ratios.
  - This gives a way of assessing potential risk of disclosure.
  - It is a Linear Integer Program and Non-linear Integer program.

- Algebraic techniques are being developed to solve Integer Programs.
  - Available symbolic software packages can be used to solve (small) problems: Latte, 4ti2.

- Another active area of research is the study of the integer gap: difference from the solution obtained using LP-relaxations. Serkan and Sturmfels (2003). Results indicate that the gap can be considerable.
Conditional inference given the margins: counting & optimizing

- Need to include margin for explanatory variables [CST].
- Two interesting well-fitting models with $\Delta G^2=5.4$ on 2 d.f.:
  - 1. [CST][CRS] 65,419,200 tables and 2. [CST][CRS][RT] 108,490 tables

<table>
<thead>
<tr>
<th>Center</th>
<th>Status</th>
<th>Treatment</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Poor</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Active</td>
<td>3 [0,14]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Placebo</td>
<td>11 [0,14]</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Active</td>
<td>3 [0,9]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Placebo</td>
<td>6 [0,9]</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Active</td>
<td>12 [2,21]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Placebo</td>
<td>11 [2,21]</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>Active</td>
<td>3 [0,9]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Placebo</td>
<td>6 [0,9]</td>
</tr>
</tbody>
</table>

- Safe to release

Software: Latte, 4ti2
Conditional inference given the conditionals: counting & optimizing

- Release full conditional $[R|CST]$ and sample size

<table>
<thead>
<tr>
<th>Center</th>
<th>Status</th>
<th>Response Treatment</th>
<th>Poor</th>
<th>Moderate</th>
<th>Excellent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Active</td>
<td>3</td>
<td>[1, 17.03]</td>
<td>20 [6.67, 113.55]</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Placebo</td>
<td>11</td>
<td>[1.38, 51.26]</td>
<td>14 [1.75, 65.23]</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Active</td>
<td>3</td>
<td>[1, 16.48]</td>
<td>14 [4.67, 76.91]</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Placebo</td>
<td>6</td>
<td>[1.2, 38.61]</td>
<td>13 [2.60, 83.66]</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Active</td>
<td>12</td>
<td>[1.10, 79.44]</td>
<td>12 [1, 72.26]</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Placebo</td>
<td>11</td>
<td>[1.10, 79.48]</td>
<td>10 [1, 72.26]</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>Active</td>
<td>3</td>
<td>[1, 29.06]</td>
<td>9 [3, 87.17]</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>Placebo</td>
<td>6</td>
<td>[2, 51.89]</td>
<td>9 [3, 77.83]</td>
</tr>
</tbody>
</table>

- There are 7,703,002 tables
- These are LP relaxation bounds, but IP are much sharper
- For data privacy, it is safe to release this conditional

Results of clinical trial for effectiveness of analgesic drug
Data source: Koch et al. (1982)
Bounds from the posterior distribution of $R | CST$

**Histogram of $x_{[1]}$**

<table>
<thead>
<tr>
<th>Cell</th>
<th>True Value</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
<th>Mean</th>
<th>Median</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,1,1)</td>
<td>3</td>
<td>0</td>
<td>6</td>
<td>2.709</td>
<td>3</td>
<td>0.9176728</td>
</tr>
</tbody>
</table>

Compared to LP: $[1, 17.03]$

Presence of integer gaps which can strongly influence the disclosure risk and utility.
Practical Implications

- Synthetic data methods currently look promising and are being implemented
- Currently more focus on alterations in microdata, then additional privacy layers in tabular data
- Agencies already release conditionals in 2-way and 3-way tables, and conditionals reveal zero counts
- K-way table
  - Releasing full conditionals too risky
  - Small conditionals may release less information (less disclosure) than corresponding marginals
- Algebraic geometry useful for exploring the space of tables for smaller problems
  - Size of the move may determine uniqueness
  - Number of tables as a measure for disclosure evaluation
    - Space of tables too small
    - May reveal margins
- Related area: Privacy-Preserving Datamining
- Need to combine methods: Crypto, PPDM, SDL
The issues that remain

- Most methods that agencies use are still ad-hoc
- Implementations of new (and old) methods are nontrivial
- Increasing the sample size does NOT necessarily decrease the risk
- Definitions of utility and risk, and disclosure

- The usual hard problems remain hard
  - Preserving the structure of sampling design
  - Multi-wave surveys
  - Geographical data
  - Longitudinal data
  - Capturing the multivariate statistical characteristics
  - Modeling the joint distribution of multivariate categorical data (especially in presence of sparse data)
The ASA Committee on Privacy and Confidentiality

“We expect that privacy, confidentiality, and data security issues will become increasingly important for those professionals who depend upon the cooperation of individuals and businesses in research and statistical studies. We hope our efforts here will also lay the groundwork for future development of additional Internet-based education initiatives in this field.”

http://www.amstat.org/comm/cmtepc/index.cfm