INFO 7470/ILRLE 7400
Methods of Confidentiality Protection

John M. Abowd and Lars Vilhuber
April 19, 2011
Outline

• What is Statistical Disclosure Limitation?
• What is Privacy-preserving Data Mining?
• Basic Methods for Disclosure Avoidance (SDL)
• Rules and Methods for Model-based SDL
• Noise Infusion Methods
Statistical Disclosure Limitation

• Protection of the confidentiality of the underlying micro-data
  – Avoiding identity disclosure
  – Avoiding attribute disclosure
  – Avoiding inferential disclosure

• Identity disclosure: who (or what entity) is in the confidential micro-data

• Attribute disclosure: value of a characteristic for that entity or individual

• Inferential disclosure: improvement of the posterior odds of a particular event
Privacy-preserving Datamining

• Formally define the properties of “privacy”
• Introduce algorithmic uncertainty as part of the statistical process
• Prove that the algorithmic uncertainty meets the formal definition of privacy
General Methods for Statistical Disclosure Limitation

• At the Census Bureau SDL is called Disclosure Avoidance Review
• Traditional methods
  – Suppression
  – Coarsening
  – Adding noise via swapping
  – Adding noise via sampling
• Newer methods
  – Explicit noise infusion
  – Synthetic data
  – Formal privacy-preserving sanitizers
Suppression

• This is by far the most common technique
• Model the sensitivity of a particular data item or observation ("disclosure risk")
• Do not allow the release of data items that have excessive disclosure risk (primary suppression)
• Do not allow the release of other data from which the sensitive item can be calculated (complementary suppression)
Suppression in Model-base Releases

• Most data analysis done in the RDCs is model-based
• The released data consist of summary statistics, model coefficients, standard errors, some diagnostic statistics
• The SDL technique used for these releases is usually suppression: the suppression rules are contained (up to confidential parameters) in the RDC Researcher’s Handbook
Coarsening

- Coarsening is the creation of a smaller number of categories from the variable in order to increase the number of cases in each cell
- Computer scientists call this “generalizing”
- Geographic coarsening: block-block group-tract-place-county-state-region
- Top coding of income is a form of coarsening
- All continuous variables in a micro-data file can be considered coarsened to the level of precision released
- This method is often applied to model-based data releases by restricting the number of significant digits that can be released
Swapping

• Estimate the disclosure risk of certain attributes or individuals
• If the risk is too great, attributes of one data record are (randomly) swapped with the same attributes of another record
• If geographic attributes are swapped this has the effect of placing the risky attributes in a different location from the truth
• Commonly used in household censuses and surveys
• Rarely used with establishment data
Sampling

• Sampling is the original SDL technique

• By only selecting certain entities from the population on which to collect additional data (data not on the frame), uncertainty about which entity was sampled provides some protection

• In modern, detailed surveys, sampling is of limited use for SDL
Rules and Methods for Model-based SDL

• Refer to Chapter 3 of the RDC Researcher’s Handbook

• Suppress: coefficients on detailed indicator variables, on cells with too few entities

• Smooth: density estimation and quantiles, use a kernel density estimator to produce quantiles

• Coarsen: variables with heavy tails (earnings, payroll), residuals (truncate range, suppress labels of range)
Noise Infusion

- Introduction
- Application to the Quarterly Workforce Indicators
- Measures of Protection
- Measures of Analytical Validity
Explicit Noise Infusion

- Adding noise to the published item or to the underlying micro data to disguise the true value
- Example: QWIs and work place data in OTM
The Quarterly Workforce Indicator System

- Multiplicative noise infusion system
- Establishment level micro data are distorted according to a permanent distortion factor
- Distortion factor always moves the fuzzed item away from the actual item by a minimum and maximum percentage
- All release data are computed from the fuzzed items
References for QWI


Noise Factor Distribution

\[ p(\delta_j) = \begin{cases} 
  \frac{(b-\delta)}{(b-a)^2}, & \delta \in [a,b] \\
  \frac{(b+\delta-2)}{(b-a)^2}, & \delta \in [2-b, 2-a] 
\end{cases} \]

\[ F(\delta_j) = \begin{cases} 
  0.5 + \frac{\left[ (b-a)^2 - (b-\delta)^2 \right]}{2(b-a)^2}, & \delta \in [a,b] \\
  \frac{\left[ (\delta+b-2)^2 \right]}{2(b-a)^2}, & \delta \in [2-b, 2-a] 
\end{cases} \]
Graph of Noise Distribution
Implementation of Noise Infusion

- Counts: \( B, E, M, F, A, S, H, R, FA, FS, W1, W2, W3, NA, NH, NR, \) and \( NS \)
- Ratios: \( Z_{W2}, Z_{W3}, Z_{WA}, Z_{WS}, Z_{NA}, Z_{NH}, Z_{NR}, \) and \( Z_{NS} \)
- Differences: \( JF, JC, JD, FJF, FJC, FJD, DWA, DWFA, DWS, DWFS \)
- In \( OTM, B \) is distorted as a count
Multiplicative Noise Infusion

- $B$ is beginning of quarter employment; $E$ is end of period; $E$-bar is the average.
- $Z \_ W2$ is end of quarter employee earnings, $W2$ is total payroll for end of quarter employees.
- $JF$ is net job flows
- Asterisk indicates distorted values.

\[ B^*_{jt} = \delta_j \times B_{jt} \]

\[ Z \_ W^*_{2,jt} = \frac{W^*_{2,jt}}{E_{jt}} = \frac{\delta_j \times W_{2,jt}}{E_{jt}} \]

\[ JF^*_{kt} = G_{kt} \times \overline{E}^*_{kt} = JF_{kt} \times \frac{\overline{E}^*_{kt}}{E_{kt}} \]

\[ Z \_ DWA^*_{kt} = \frac{\Delta WA_{kt}}{A_{kt}} \times \frac{A^*_{kt}}{A_{kt}} \]
Weighting

• Each fuzzed micro-data item is weighted by the QWI final weight before aggregation
• This means that all input data are real numbers (not integers)
• Final disclosure control formulas must reflect rounding of the counts
Interpreting the Algorithm

• Based on the public use version of the spreadsheet
<table>
<thead>
<tr>
<th>Variable</th>
<th>Fuzz_type</th>
<th>Fuzz_value_numerator</th>
<th>Fuzz_value_denominator</th>
<th>Fuzz_change_numerator</th>
<th>Fuzz_change_denominator</th>
</tr>
</thead>
<tbody>
<tr>
<td>BeginEmp</td>
<td>count</td>
<td>count</td>
<td>delta_B</td>
<td>unity</td>
<td>unity</td>
</tr>
<tr>
<td>EndEmp</td>
<td>count</td>
<td>count</td>
<td>delta_E</td>
<td>unity</td>
<td>unity</td>
</tr>
<tr>
<td>JobFlowNet</td>
<td>count</td>
<td>count</td>
<td>delta_EBAR</td>
<td>unity</td>
<td>unity</td>
</tr>
<tr>
<td>JobCreate</td>
<td>count</td>
<td>count</td>
<td>delta_EBAR</td>
<td>unity</td>
<td>unity</td>
</tr>
<tr>
<td>JobDest</td>
<td>count</td>
<td>count</td>
<td>delta_EBAR</td>
<td>unity</td>
<td>unity</td>
</tr>
<tr>
<td>Acc</td>
<td>count</td>
<td>count</td>
<td>delta_A</td>
<td>unity</td>
<td>unity</td>
</tr>
<tr>
<td>Hire</td>
<td>count</td>
<td>count</td>
<td>delta_H</td>
<td>unity</td>
<td>unity</td>
</tr>
<tr>
<td>Recl</td>
<td>count</td>
<td>count</td>
<td>delta_R</td>
<td>unity</td>
<td>unity</td>
</tr>
<tr>
<td>Sep</td>
<td>count</td>
<td>count</td>
<td>delta_S</td>
<td>unity</td>
<td>unity</td>
</tr>
<tr>
<td>FullEmp</td>
<td>count</td>
<td>count</td>
<td>delta_F</td>
<td>unity</td>
<td>unity</td>
</tr>
<tr>
<td>JobCreate</td>
<td>count</td>
<td>count</td>
<td>delta_FBAR</td>
<td>unity</td>
<td>unity</td>
</tr>
<tr>
<td>JobDest</td>
<td>count</td>
<td>count</td>
<td>delta_FBAR</td>
<td>unity</td>
<td>unity</td>
</tr>
<tr>
<td>JobDest</td>
<td>count</td>
<td>count</td>
<td>delta_FBAR</td>
<td>unity</td>
<td>unity</td>
</tr>
<tr>
<td>JobDest</td>
<td>count</td>
<td>count</td>
<td>delta_FBAR</td>
<td>unity</td>
<td>unity</td>
</tr>
<tr>
<td>FulQrtTurn</td>
<td>ratio</td>
<td>aggregate</td>
<td>delta_FA</td>
<td>unity</td>
<td>unity</td>
</tr>
<tr>
<td>FulEmpFlw</td>
<td>count</td>
<td>count</td>
<td>delta_FA</td>
<td>unity</td>
<td>unity</td>
</tr>
<tr>
<td>FulHire</td>
<td>count</td>
<td>count</td>
<td>delta_FA</td>
<td>unity</td>
<td>unity</td>
</tr>
<tr>
<td>FulSep</td>
<td>count</td>
<td>count</td>
<td>delta_FS</td>
<td>unity</td>
<td>unity</td>
</tr>
<tr>
<td>EarnEnd</td>
<td>ratio</td>
<td>ratio</td>
<td>delta_W2</td>
<td>unity</td>
<td>unity</td>
</tr>
<tr>
<td>EarnFul</td>
<td>ratio</td>
<td>ratio</td>
<td>delta_W3</td>
<td>unity</td>
<td>unity</td>
</tr>
<tr>
<td>EarnFulAcc</td>
<td>ratio</td>
<td>ratio</td>
<td>delta_WFA</td>
<td>unity</td>
<td>unity</td>
</tr>
<tr>
<td>ChgEmpAcc</td>
<td>ratio</td>
<td>ratio</td>
<td>change_ratio</td>
<td>dWA</td>
<td>A</td>
</tr>
<tr>
<td>NonEmpAcc</td>
<td>ratio</td>
<td>ratio</td>
<td>change_ratio</td>
<td>dWA</td>
<td>A</td>
</tr>
<tr>
<td>NonEmpHire</td>
<td>ratio</td>
<td>ratio</td>
<td>change_ratio</td>
<td>dWA</td>
<td>A</td>
</tr>
<tr>
<td>NonEmpHire</td>
<td>ratio</td>
<td>ratio</td>
<td>change_ratio</td>
<td>dWA</td>
<td>A</td>
</tr>
<tr>
<td>NonEmpHire</td>
<td>ratio</td>
<td>ratio</td>
<td>change_ratio</td>
<td>dWA</td>
<td>A</td>
</tr>
<tr>
<td>EarnFulHire</td>
<td>ratio</td>
<td>ratio</td>
<td>change_ratio</td>
<td>dWA</td>
<td>A</td>
</tr>
<tr>
<td>EarnFulHire</td>
<td>ratio</td>
<td>ratio</td>
<td>change_ratio</td>
<td>dWA</td>
<td>A</td>
</tr>
<tr>
<td>Payroll</td>
<td>sum</td>
<td>sum</td>
<td>sum</td>
<td>unity</td>
<td>unity</td>
</tr>
<tr>
<td>TotalEmp</td>
<td>count</td>
<td>count</td>
<td>delta_M</td>
<td>unity</td>
<td>unity</td>
</tr>
</tbody>
</table>
Protection Properties of the QWI Algorithm
### Table 1A

**Variable: B**

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5+</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>99.61</td>
<td>0.39</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>1</td>
<td>0.00</td>
<td>98.57</td>
<td>1.43</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>0.00</td>
<td>1.04</td>
<td>96.10</td>
<td>2.85</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td>0.00</td>
<td>0.00</td>
<td>2.19</td>
<td>93.21</td>
<td>4.60</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>7.30</td>
<td>82.52</td>
<td>10.18</td>
</tr>
<tr>
<td>5+</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.60</td>
<td>98.40</td>
</tr>
</tbody>
</table>

Note: The data represent county data for Maryland/NAICS Industry Group. Cells represent row percentages and sum to 100.
### Table 1B

**Variable: B**

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5+</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>99.19</td>
<td>0.81</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>1</td>
<td>0.14</td>
<td>89.29</td>
<td>10.56</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>0.04</td>
<td>1.39</td>
<td>67.45</td>
<td>30.70</td>
<td>0.42</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td>0.03</td>
<td>0.04</td>
<td>2.19</td>
<td>50.99</td>
<td>42.76</td>
<td>3.99</td>
</tr>
<tr>
<td>4</td>
<td>0.03</td>
<td>0.02</td>
<td>0.03</td>
<td>3.07</td>
<td>41.04</td>
<td>55.81</td>
</tr>
<tr>
<td>5+</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.02</td>
<td>0.33</td>
<td>99.64</td>
</tr>
</tbody>
</table>

Note: The data represent county data for Maryland/NAICS Industry Group. Cells represent row percentages and sum to 100.
## Table 1C

**Variable: B**

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5+</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>99.17</td>
<td>0.82</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>1</td>
<td>7.85</td>
<td>84.74</td>
<td>6.62</td>
<td>0.78</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>0.51</td>
<td>11.93</td>
<td>61.06</td>
<td>24.14</td>
<td>2.24</td>
<td>0.12</td>
</tr>
<tr>
<td>3</td>
<td>0.06</td>
<td>0.76</td>
<td>7.53</td>
<td>47.50</td>
<td>39.13</td>
<td>5.02</td>
</tr>
<tr>
<td>4</td>
<td>0.03</td>
<td>0.11</td>
<td>0.93</td>
<td>7.40</td>
<td>38.84</td>
<td>52.69</td>
</tr>
<tr>
<td>5+</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.11</td>
<td>0.71</td>
<td>99.16</td>
</tr>
</tbody>
</table>

Note: The data represent county data for Maryland/NAICS Industry Group. Cells represent row percentages and sum to 100.
Analytical Validity Properties
Error Distribution for B (Micro-data)
Error Distribution for W (Micro-data)
Table 7

Distribution of the Difference Between Autocorrelation Coefficients
Unweighted/Undistorted vs. Unweighted/Distorted

<table>
<thead>
<tr>
<th>Percentile</th>
<th>B</th>
<th>H</th>
<th>R</th>
<th>E</th>
<th>A</th>
<th>S</th>
<th>M</th>
<th>F</th>
<th>FA</th>
<th>FS</th>
<th>H3</th>
</tr>
</thead>
<tbody>
<tr>
<td>99</td>
<td>0.067</td>
<td>0.058</td>
<td>0.042</td>
<td>0.066</td>
<td>0.061</td>
<td>0.061</td>
<td>0.079</td>
<td>0.059</td>
<td>0.056</td>
<td>0.054</td>
<td>0.056</td>
</tr>
<tr>
<td>95</td>
<td>0.036</td>
<td>0.035</td>
<td>0.025</td>
<td>0.036</td>
<td>0.036</td>
<td>0.036</td>
<td>0.042</td>
<td>0.032</td>
<td>0.034</td>
<td>0.033</td>
<td>0.035</td>
</tr>
<tr>
<td>90</td>
<td>0.025</td>
<td>0.025</td>
<td>0.015</td>
<td>0.024</td>
<td>0.025</td>
<td>0.025</td>
<td>0.028</td>
<td>0.021</td>
<td>0.024</td>
<td>0.023</td>
<td>0.024</td>
</tr>
<tr>
<td>75</td>
<td>0.009</td>
<td>0.009</td>
<td>0.005</td>
<td>0.009</td>
<td>0.010</td>
<td>0.010</td>
<td>0.011</td>
<td>0.008</td>
<td>0.009</td>
<td>0.008</td>
<td>0.008</td>
</tr>
<tr>
<td>50</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>25</td>
<td>-0.008</td>
<td>-0.011</td>
<td>-0.005</td>
<td>-0.009</td>
<td>-0.010</td>
<td>-0.010</td>
<td>-0.011</td>
<td>-0.008</td>
<td>-0.009</td>
<td>-0.008</td>
<td>-0.009</td>
</tr>
<tr>
<td>10</td>
<td>-0.023</td>
<td>-0.028</td>
<td>-0.016</td>
<td>-0.023</td>
<td>-0.026</td>
<td>-0.026</td>
<td>-0.027</td>
<td>-0.022</td>
<td>-0.026</td>
<td>-0.024</td>
<td>-0.026</td>
</tr>
<tr>
<td>5</td>
<td>-0.035</td>
<td>-0.038</td>
<td>-0.025</td>
<td>-0.037</td>
<td>-0.036</td>
<td>-0.038</td>
<td>-0.041</td>
<td>-0.033</td>
<td>-0.037</td>
<td>-0.035</td>
<td>-0.038</td>
</tr>
<tr>
<td>1</td>
<td>-0.061</td>
<td>-0.063</td>
<td>-0.045</td>
<td>-0.065</td>
<td>-0.059</td>
<td>-0.065</td>
<td>-0.074</td>
<td>-0.061</td>
<td>-0.061</td>
<td>-0.059</td>
<td>-0.062</td>
</tr>
</tbody>
</table>

Note: The data represent county data for Maryland/NAICS Industry Group. Cells represent the difference between the autocorrelation coefficients the percentile designated by the rows.
### Table 8

Distribution of the Difference Between Autocorrelation Coefficients
Unweighted/Undistorted vs. Weighted/Distorted w/ Suppressions

<table>
<thead>
<tr>
<th>Percentile</th>
<th>B</th>
<th>H</th>
<th>R</th>
<th>E</th>
<th>A</th>
<th>S</th>
<th>M</th>
<th>F</th>
<th>FA</th>
<th>FS</th>
<th>H3</th>
</tr>
</thead>
<tbody>
<tr>
<td>99</td>
<td>0.318</td>
<td>0.592</td>
<td>0.538</td>
<td>0.325</td>
<td>0.602</td>
<td>0.652</td>
<td>0.330</td>
<td>0.366</td>
<td>0.606</td>
<td>0.656</td>
<td>0.648</td>
</tr>
<tr>
<td>95</td>
<td>0.153</td>
<td>0.326</td>
<td>0.221</td>
<td>0.158</td>
<td>0.294</td>
<td>0.304</td>
<td>0.134</td>
<td>0.181</td>
<td>0.317</td>
<td>0.352</td>
<td>0.359</td>
</tr>
<tr>
<td>90</td>
<td>0.086</td>
<td>0.196</td>
<td>0.136</td>
<td>0.084</td>
<td>0.173</td>
<td>0.181</td>
<td>0.063</td>
<td>0.108</td>
<td>0.196</td>
<td>0.218</td>
<td>0.229</td>
</tr>
<tr>
<td>75</td>
<td>0.021</td>
<td>0.060</td>
<td>0.049</td>
<td>0.019</td>
<td>0.053</td>
<td>0.053</td>
<td>0.017</td>
<td>0.025</td>
<td>0.068</td>
<td>0.070</td>
<td>0.087</td>
</tr>
<tr>
<td>50</td>
<td>-0.002</td>
<td>-0.006</td>
<td>0.000</td>
<td>-0.003</td>
<td>-0.006</td>
<td>-0.008</td>
<td>-0.006</td>
<td>-0.001</td>
<td>-0.004</td>
<td>-0.004</td>
<td>-0.001</td>
</tr>
<tr>
<td>25</td>
<td>-0.029</td>
<td>-0.075</td>
<td>-0.063</td>
<td>-0.031</td>
<td>-0.071</td>
<td>-0.078</td>
<td>-0.037</td>
<td>-0.026</td>
<td>-0.070</td>
<td>-0.084</td>
<td>-0.082</td>
</tr>
<tr>
<td>10</td>
<td>-0.085</td>
<td>-0.188</td>
<td>-0.162</td>
<td>-0.091</td>
<td>-0.170</td>
<td>-0.190</td>
<td>-0.099</td>
<td>-0.077</td>
<td>-0.169</td>
<td>-0.200</td>
<td>-0.190</td>
</tr>
<tr>
<td>5</td>
<td>-0.150</td>
<td>-0.289</td>
<td>-0.242</td>
<td>-0.154</td>
<td>-0.258</td>
<td>-0.288</td>
<td>-0.176</td>
<td>-0.141</td>
<td>-0.262</td>
<td>-0.300</td>
<td>-0.286</td>
</tr>
<tr>
<td>1</td>
<td>-0.393</td>
<td>-0.551</td>
<td>-0.474</td>
<td>-0.426</td>
<td>-0.505</td>
<td>-0.531</td>
<td>-0.489</td>
<td>-0.419</td>
<td>-0.499</td>
<td>-0.538</td>
<td>-0.527</td>
</tr>
</tbody>
</table>

Note: The data represent county data for Maryland/NAICS Industry Group. Cells represent the difference between the autocorrelation coefficients the percentile designated by the rows.
Table 9

Distribution of the Difference Between Autocorrelation Coefficients
Unweighted/Undistorted vs. Weighted/Distorted w/ Synthetic Replacements

<table>
<thead>
<tr>
<th>Percentile</th>
<th>B</th>
<th>H</th>
<th>R</th>
<th>E</th>
<th>A</th>
<th>S</th>
<th>M</th>
<th>F</th>
<th>FA</th>
<th>FS</th>
<th>H3</th>
</tr>
</thead>
<tbody>
<tr>
<td>99</td>
<td>0.107</td>
<td>0.246</td>
<td>0.223</td>
<td>0.096</td>
<td>0.238</td>
<td>0.263</td>
<td>0.210</td>
<td>0.227</td>
<td>0.260</td>
<td>0.157</td>
<td>0.294</td>
</tr>
<tr>
<td>95</td>
<td>0.056</td>
<td>0.144</td>
<td>0.126</td>
<td>0.051</td>
<td>0.134</td>
<td>0.166</td>
<td>0.123</td>
<td>0.112</td>
<td>0.145</td>
<td>0.076</td>
<td>0.185</td>
</tr>
<tr>
<td>90</td>
<td>0.036</td>
<td>0.104</td>
<td>0.092</td>
<td>0.033</td>
<td>0.099</td>
<td>0.120</td>
<td>0.084</td>
<td>0.075</td>
<td>0.108</td>
<td>0.049</td>
<td>0.133</td>
</tr>
<tr>
<td>75</td>
<td>0.011</td>
<td>0.050</td>
<td>0.043</td>
<td>0.010</td>
<td>0.047</td>
<td>0.058</td>
<td>0.037</td>
<td>0.031</td>
<td>0.051</td>
<td>0.018</td>
<td>0.066</td>
</tr>
<tr>
<td>50</td>
<td>-0.004</td>
<td>0.003</td>
<td>0.007</td>
<td>-0.005</td>
<td>0.005</td>
<td>0.007</td>
<td>0.006</td>
<td>0.002</td>
<td>0.007</td>
<td>-0.003</td>
<td>0.009</td>
</tr>
<tr>
<td>25</td>
<td>-0.030</td>
<td>-0.043</td>
<td>-0.013</td>
<td>-0.030</td>
<td>-0.032</td>
<td>-0.032</td>
<td>-0.017</td>
<td>-0.018</td>
<td>-0.030</td>
<td>-0.034</td>
<td>-0.037</td>
</tr>
<tr>
<td>10</td>
<td>-0.065</td>
<td>-0.109</td>
<td>-0.067</td>
<td>-0.064</td>
<td>-0.089</td>
<td>-0.095</td>
<td>-0.049</td>
<td>-0.056</td>
<td>-0.095</td>
<td>-0.082</td>
<td>-0.128</td>
</tr>
<tr>
<td>5</td>
<td>-0.092</td>
<td>-0.167</td>
<td>-0.120</td>
<td>-0.093</td>
<td>-0.142</td>
<td>-0.145</td>
<td>-0.081</td>
<td>-0.088</td>
<td>-0.144</td>
<td>-0.116</td>
<td>-0.194</td>
</tr>
<tr>
<td>1</td>
<td>-0.166</td>
<td>-0.295</td>
<td>-0.257</td>
<td>-0.176</td>
<td>-0.257</td>
<td>-0.268</td>
<td>-0.168</td>
<td>-0.187</td>
<td>-0.277</td>
<td>-0.233</td>
<td>-0.343</td>
</tr>
</tbody>
</table>

Note: The data represent county data for Maryland/NAICS Industry Group. Cells represent the difference between the autocorrelation coefficients the percentile designated by the rows.
QWIPU: Aggregating Formulas

• Counts and magnitudes:
  – Add

• Ratios and differences
  – Multiply by released base
  – Aggregate numerator and denominator separately
  – Add

• Job creations and destructions
  – Handle like counts but understand that there is an inherent loss of information
QWIPU: Handling the Suppressions

• Status flag 5 suppressions
  – Must be treated as missing data and estimated

• Status flag 9 data
  – Not suppressed. Should be used as published.
Research Uses of the Micro-data

• QWI Micro-data (UFF_B)
  – Use the undistorted confidential data (establishment level) in models
  – Use conventional model-based disclosure avoidance rules
  – No fuzzed data should be used

• OTM Micro-data (WHAT_B)
  – All micro-data and all imputations of missing data used to build OTM
  – Contain both fuzzed and unfuzzed values with separate weights
  – Unfuzzed data should be used in models, weighted as appropriate
INFO 7470/ILRLE 7400
Cryptographic and Statistical Advances in Confidentiality Protection

John M. Abowd and Lars Vilhuber
April 19, 2011
Outline

• Motivation: formal privacy models and statistical disclosure limitation
• The basic OnTheMap application
• The statistical structure of the OnTheMap data
• Applying probabilistic differential privacy to OnTheMap
• The trade-off between analytical validity and confidentiality protection
Formal Privacy Models and Statistical Disclosure Limitation

• Formal privacy protection methods are based on open algorithms with provable properties

• The standard in privacy-preserving datamining is based on cryptography:
  – Only the private key (password, encryption key) is confidential; all algorithms and parameters are public
  – Attacker (= user) can have massive amounts of prior information
The Cryptographic Critique of SDL

• Standard SDL techniques fail because:
  – They and do not have provably protective properties when the attacker (= user) is allowed full access to the algorithm
  – They depend upon the realized data and not the algorithm

• Many standard SDL techniques are viewed as very risky when the cryptographic critique is applied
Point of Common Ground

• Federal Committee on Statistical Methodology working paper 22 offers the desirable disclosure avoidance property:

  Disclosure relates to inappropriate attribution of information to a data subject, whether an individual or an organization. Disclosure occurs when a data subject is identified from a released file (identity disclosure), sensitive information about a data subject is revealed through the released file (attribute disclosure), or the released data make it possible to determine the value of some characteristic of an individual more accurately than otherwise would have been possible (inferential disclosure). (page 4)

• Evfimievski, Gehrke and Srikant (2003), Dwork (2006) show that disclosure avoidance in this sense is impossible to achieve in general.
Focus on Synthetic Data and Randomized Sanitizers

• The SDL technique known as synthetic data most closely resembles the cryptographic data protection techniques
• The cryptographic techniques are known as privacy-preserving datamining, randomized sanitizers, differential privacy, and e-privacy.
Definition of Synthetic Data

\[ X \equiv \text{confidential data} \]

\[ \Pr[\tilde{X}|X] \equiv \text{PPD of } \tilde{X} \text{ given } X \]

Release data are samples of \( \tilde{X} \)

- Synthetic data are created by estimating the posterior predictive distribution (PPD) of the release data given the confidential data; then sampling release data from the PPD conditioning on the actual confidential values.

- The PPD is a parameter-free forecasting model for new values of the complete data matrix that conditions on all values of the underlying confidential data.
Connection to Randomized Sanitizers

\( X \equiv \text{confidential data} \)

\( U \equiv \text{random noise} \)

\( \text{San}(X,U):(X,U) \rightarrow \tilde{X} \)

\( \Pr[\tilde{X}|X] \equiv \text{probability of } \tilde{X} \text{ given } X \)

• A randomized sanitizer creates a conditional probability distribution for the release data given the confidential data

• The randomness in a sanitizer is induced by the properties of the distribution of \( U \)

• The PPD is just a particular randomized sanitizer
$\varepsilon$-Differential Privacy

Definition ($\varepsilon$ - Differential Privacy): Let $A$ be a randomized algorithm, let $S$ be the set of all possible outputs of the algorithm, and let $\varepsilon > 0$. The algorithm $A$ satisfies $\varepsilon$ - differential privacy if for all pairs of data sets $(D_1, D_2)$ that differ in exactly one row,

$$\forall S \in S, \frac{P(A(D_1)) = S}{P(A(D_2)) = S} \leq e^{\varepsilon} \text{ or } \ln \left| \frac{P(A(D_1)) = S}{P(A(D_2)) = S} \right| < \varepsilon.$$ 

- Differential privacy (Dwork, and many co-authors) is difficult to maintain in sparse applications when geographically near blocks have very different posterior probabilities.
Disclosure Set

Definition (Disclosure Set): Let $D$ be a table and $\mathcal{D}$ be the set of tables that differ from $D$ in at most one row. Let $A$ be a randomized algorithm and $\mathcal{S}$ be the space of outputs of the algorithm $A$. The disclosure set of $D$, denoted $\text{Disc}(D, \varepsilon)$, is

$$\left\{ S \in \mathcal{S} \left| \exists X_1, X_2 \in \mathcal{D}(D), |X_1 \setminus X_2| = 1 \land \ln \frac{P(A(X_1) = S)}{P(A(X_2) = S)} > \varepsilon \right. \right\}.$$

- This set describes the outcomes where differential privacy fails.
Probabilistic Differential Privacy

Definition (Probabilistic Differential Privacy): Let $A$ be a randomized algorithm and $S$ be the space of outputs of $A$. Let $\varepsilon > 0$ and $0 < \delta < 1$ be constants. Then $A$ satisfies $(\varepsilon, \delta)$-probabilistic differential privacy (or $(\varepsilon, \delta)$-pdp) if for all tables $D$, $P(A(D) \in \text{Disc}(D, \varepsilon)) \leq \delta$.

- PDP allows us to control the probability that differential privacy fails
- The analytical validity of sparse applications can be controlled with PDP because the restrictions on the prior used in the synthesizer are reasonable for use with sparse tables
Disclosure Limitation Definitions

\[ X = x^{(1)} \text{ and } X = x^{(2)} \]

\[ \tilde{X} = \tilde{x}, \text{ realization of the synthesizer} \]

- Consider two confidential data matrices that differ in only a single row, \( x^{(1)} \) and \( x^{(2)} \)
- Use the PPD to evaluate the probability of a particular release data set given the two different confidential data sets
Synthetic Data Can Leak Information about a Single Entity

\[ \Pr[\tilde{X} = \tilde{x} | X = x^{(1)}] \neq \Pr[\tilde{X} = \tilde{x} | X = x^{(2)}] \]

- Changing a single row of the confidential data matrix changes the PPD or the random sanitizer
- The PPD or the random sanitizer define the transition probabilities from the confidential data to the release data
- True for all SDL procedures that infuse noise
Connection Between Synthetic Data and Differential Privacy

\[
\frac{\Pr[X = x^{(1)} | \tilde{X} = \tilde{x}]}{\Pr[X = x^{(2)} | \tilde{X} = \tilde{x}]} = \frac{\frac{\Pr[\tilde{X} = \tilde{x} | X = x^{(1)}]}{\Pr[X = x^{(1)}]}}{\frac{\Pr[\tilde{X} = \tilde{x} | X = x^{(2)}]}{\Pr[X = x^{(2)}]}}
\]

The posterior odds ratio for the gain in information about a single row of \( X \) is equal to the differential privacy from the randomized sanitizer that creates release data by sampling from the specified conditional distribution.
Connection Between Differential Privacy and Inferential Disclosure

The posterior odds ratio for the gain in information about a single row of $X$ is the Dalenius (1977) definition of an inferential disclosure. Bounding the differential privacy therefore bounds the inferential disclosure.
Taking Account of Formal Privacy Models

• A variety of papers in the cryptographic data privacy literature (Dwork, Nissim and their many collaborators, Gehrke and his collaborators, and others) show that the confidentiality protection afforded by synthetic data or a randomized sanitizer depends upon properties of the transition probabilities that relate the confidential data to the release data.

• Exact data releases are not safe. Not surprising since

\[
\Pr[\tilde{X} | X] = I
\]

implies that the sanitizer leaves the confidential data unchanged.

• Off-diagonal elements that are zero imply infinite differential privacy: exact disclosure in some cases with probability 1.

• For a full explanation of the relation between the transition matrix and differential privacy measures see Abowd and Vilhuber (2008).
Relationship to Post-randomization

• Post-randomization (Kooiman et al. 1997) focuses on the diagonal elements of $\Pr[\tilde{X}|X]$.

• When off-diagonal elements of this transition matrix are zero, infinite differential privacy usually results.

• Swapping, shuffling, stratified sampling, and most noise-infusion methods result in off-diagonal elements that are zero.
A DETAILED EXAMPLE: SYNTHETIC DATA
The Multinomial-Dirichlet Model

- The data matrix $X$ consists of categorical variables that can be summarized by a contingency table with $k$ categories.
- $n_i$ are counts.
- $\pi_i$ are probabilities.

$n = (n_1, \ldots, n_k), n = \sum n_i$

$\alpha = (\alpha_1, \ldots, \alpha_k), \alpha_0 = \sum \alpha_i$

$\pi = (\pi_1, \ldots, \pi_k)$

$n \sim M(n, \pi)$

$\pi \sim D(\alpha), \text{a priori}$

$\pi \sim D(\alpha + n), \text{a posteriori}$

$m = (m_1, \ldots, m_k), m = \sum m_i$

$m \sim M(m, \pi)$
The Multinomial-Dirichlet Synthesizer

\[ \Pr[m|n] = E_{\pi|n}[M(m, \pi)] \]

• The synthetic data are samples from the synthesizer, and can be summarized by their counts, \( m \)

• Since all the random variables are discrete, the synthesizer can be expressed as a simple transition probability matrix
- $k = 2$
- $\alpha_i = \frac{1}{2}; \alpha_0 = 1$
- $n = m = 5$
- The table displays the transition probabilities that map $n$ into $m$
\( \varepsilon \)-Differential Privacy

\[
\left| \ln \frac{\Pr[m|n^{(1)}]}{\Pr[m|n^{(2)}]} \right| < \varepsilon
\]

- The two confidential data matrices, \( n^{(1)} \) and \( n^{(2)} \) differ by changing exactly one entity’s data.
- Bounding by \( \varepsilon \) the log inferential disclosure odds ratio in the M-D synthesizer amounts to controlling the probabilities in \( \Pr[m|n] \) appropriately.
The table shows all of the differential privacy ratios for the example problem.

The $\epsilon$-differential privacy of this synthesizer is the maximum element in this table, 5.493061.

The differential privacy limit is attained when the synthesizer delivers (0,5) and the underlying data are either (5,0) or (4,1) (or (0,5) with original data (1,4) or (5,0)).

If I release (5,0) and you know 4 people are in category 2, then the odds are 243:1 (= $\exp(5.493061)$) that the unknown person is in category 1.
Probabilistic Differential Privacy

• This definition of differential privacy allows the $\varepsilon$-differential privacy limit to fail with probability $\delta$ (Machanavajjhala et al. 2008)

• To compute the PDP, the joint distribution of $m$ and $n$ must be examined for outcomes with differential privacy that exceed the limit to ensure that they occur with total probability less than $\delta$
- The table is $\Pr[m,n]$, where the marginal $\Pr[n]$ is based on the prior $D(\alpha)$.
- If we want to have $\varepsilon$-differential privacy of 2, then the synthesizer fails in the highlighted cells.
- With prior $D(\alpha)$, probabilistic differential privacy has $\varepsilon = 2$ and $\delta = 0.000623$, which is just the sum of the highlighted cells.
A DETAILED EXAMPLE: RANDOM SANITIZER
Dwork et al. (2006) show that $\varepsilon$-differential privacy can be achieved in the Multinomial model with a sanitizer using independent double exponential noise (Laplace noise) with mean zero and variance $2/\varepsilon$.

Note that in our application the total $n$ is released without noise.
\[ k = 2 \]
\[ n = m = 5 \]
\[ \varepsilon = 2 \]

The table displays the transition probabilities that map \( n \) into \( m \)

Note that the diagonals are larger than the M-D model and the extreme outcomes have greater probability
The table confirms that the transition matrix on the previous page has $\varepsilon = 2$.
Challenges and Applications

• Realistic problems are all very sparse
• Probabilistic differential privacy can solve the sparseness problem
  – But, it requires coarsening and domain shrinking to deliver acceptable analytical validity.
• The Laplace synthesizer can solve the sparseness problem by adaptive histogram coarsening
  – But the user cannot directly control the coarsening hence analytical validity for some hypotheses is low
• OnTheMap uses probabilistic differential privacy
A REAL APPLICATION: US CENSUS BUREAU’S ONTHEMAP
Where residents of Sausalito, CA with high wages are employed
The OnTheMap Data Structure

- Set of linked data tables with a relational database schema
- Main tables (micro-data)
  - Job: [Person_ID, Employer_ID]
  - Residence: [Person_ID, Origin_Block, ...]
  - Workplace: [Employer_ID, Destination_Block, ...]
  - Geo-code: [Block, Tract, Latitude, Longitude, ...]
Detailed Geo-spatial Data in OTM

• Workplace and residence geographies are defined using Census blocks
• Statistical analysis to estimate the PPD is based on Census tract-to-tract relations
• There are 8.2 million blocks and 65,000 tracts in the U.S.
• Every workplace block with positive employment has its own synthesizer
Dirichlet-Multinomial Synthesizer

- \( I \) origins
- Model each destination \( d \) separately for each demographic segment (age, earnings, industry)
- Sample data \( X \) tabulated into \( n \)
- Synthetic data tabulated into \( m \)
- Usually \( m = n \), but not in the OTM application

\[
\begin{align*}
n &= (n_1, \ldots, n_I), \quad n = \sum n_i \\
\alpha &= (\alpha_1, \ldots, \alpha_I), \quad \alpha_0 = \sum \alpha_i = |\alpha| \\
\pi &= (\pi_1, \ldots, \pi_I) \\
n &\sim M(\pi, n) \\
\pi &\sim D(\alpha), \text{ a priori} \\
\pi &\sim D(\alpha + n), \text{ a posteriori} \\
m &= (m_1, \ldots, m_I), \quad m = \sum m_i \\
m &\sim M(\pi, m), \text{ a posteriori}
\end{align*}
\]
Synthetic Data Model

• Likelihood of place of residence (index i) conditional on place of work (index j) and characteristics (index k):

\[
p(n_{ijk} \mid \pi_{i|jk}) \propto \prod_{i=1}^{I} \pi_{i|jk}^{n_{ijk}}
\]

• The resulting posterior for \( \pi \) is Dirichlet with parameter \( n_{jk} + \alpha_{jk} \) for each unique workplace and characteristic combination (age, earnings, industry).

• Synthesize residence counts by sampling from the posterior predictive distributions conditional on already protected (and published) destination employment counts, \( m_{jk} \)
We rely on the concept of $(\varepsilon, \delta)$-probabilistic differential privacy, which guarantees $\varepsilon$-differential privacy with $1-\delta$ confidence (Machanavajjhala et al. (2008)).

Search algorithm finds the minimum prior sample size to guarantee $\varepsilon$-differential privacy with failure probability $\delta$.

This minimum prior sample size is then used as the lower bound of the prior sample sizes ($\alpha$).

The privacy-preserving algorithm implemented in OnTheMap guarantees $\varepsilon$-differential privacy protection of 8.99 with 99.999999% confidence ($\delta = 0.000001$).
Measures to Improve Validity

• Coarsening of the domain
  – Reducing the number of support points in the domain of the prior

• Editing the prior domain
  – Eliminating the most unlikely commute patterns (from prior and likelihood) based on previously published data

• Use of informative priors
  – Impose likely shape based on previously published data subject to minimum prior sample size that ensures $(\varepsilon, \delta)$-PDP

• Pruning the prior
  – Randomly eliminating a fraction support points with no likelihood support.
  – Pruning comes with a penalty in terms of privacy protection
Refinement: Coarsening the Domain

- Blocks are collected into larger geographic areas- SuperPUMAs, PUMAs, Tracts
- Reduces the dimensionality of the domain of each destination’s synthesizer
- Theorem 5.1 in Machanavajjhala et al. shows that $\varepsilon$-differential privacy, and $(\varepsilon, \delta)$-probabilistic differential privacy both survive coarsening with unchanged parameters
Coarsening Steps

• If origin block very far away from destination block (distance > 90th percentile of CTTP commute distribution) coarsened to Super-PUMA (400,000 population in Census 2000)
• Else if origin block far away from destination block (distance > 50th percentile of CTTP commute distribution) coarsened to PUMA (100,000 population in Census 2000)
• Else if origin block close to destination block (distance < 50th percentile of CTTP commute distribution) coarsened to Census Tract (4,000 population on average).
• Idea: “marginal differences in commute distances between candidate locations have less predictive power in allocating workers the farther away the locations are”
Effects of Coarsening

- Coarsening in formal privacy models is effectively the same as coarsening in traditional methods.
- After coarsening, an entity (in this case a block) is chosen randomly to represent the coarsened unit (one block per SuperPUMA, PUMA, or tract, as appropriate).
- This ensures that the transition matrix has no zero elements at the block level.
- Ratios of the elements of this transition matrix determine the differential privacy.
Refinement: Editing the Prior Domain

• For each work tract:
  – if point in domain has zero probability in prior data then do:
    • eliminate point with $p=0.98$ if distance $> 500$ miles
    • eliminate point with $p=0.9$ if distance $> 200$ miles
    • eliminate point with $p=0.5$ if distance $> 100$ miles
    • do not eliminate if distance $< 100$ miles
  – else retain point

• Note: contribution of any likelihood data in eliminated points also eliminated
## Fraction of Points in the Prior Domain with Positive Counts in Census Transportation Planning Package Data

<table>
<thead>
<tr>
<th>Distance (in miles)</th>
<th>State A</th>
<th></th>
<th>State B</th>
<th></th>
<th>State C</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>- low-10</td>
<td>0.47</td>
<td>0.37</td>
<td>0.40</td>
<td>0.32</td>
<td>0.92</td>
<td>0.18</td>
</tr>
<tr>
<td>- 10-25</td>
<td>0.30</td>
<td>0.26</td>
<td>0.19</td>
<td>0.19</td>
<td>0.63</td>
<td>0.29</td>
</tr>
<tr>
<td>- 25-100</td>
<td>0.01</td>
<td>0.13</td>
<td>0.09</td>
<td>0.10</td>
<td>0.15</td>
<td>0.16</td>
</tr>
<tr>
<td>- 100-500</td>
<td>0.01</td>
<td>0.03</td>
<td>0.01</td>
<td>0.02</td>
<td>0.02</td>
<td>0.04</td>
</tr>
<tr>
<td>- 500-high</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>All</td>
<td>0.18</td>
<td>0.28</td>
<td>0.14</td>
<td>0.23</td>
<td>0.34</td>
<td>0.40</td>
</tr>
</tbody>
</table>
Fraction of Points in the Domain with Positive Counts in CTPP after Eliminating Extremely Unlikely Commute Patterns

<table>
<thead>
<tr>
<th>distance (in miles)</th>
<th>Large State</th>
<th>Medium State</th>
<th>Small State</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>- low-10</td>
<td>0.47</td>
<td>0.37</td>
<td>0.40</td>
</tr>
<tr>
<td>- 10-25</td>
<td>0.30</td>
<td>0.26</td>
<td>0.19</td>
</tr>
<tr>
<td>- 25-100</td>
<td>0.13</td>
<td>0.13</td>
<td>0.09</td>
</tr>
<tr>
<td>- 100-500</td>
<td>0.06</td>
<td>0.09</td>
<td>0.03</td>
</tr>
<tr>
<td>- 500-high</td>
<td>0.07</td>
<td>0.13</td>
<td>0.06</td>
</tr>
<tr>
<td>All</td>
<td>0.21</td>
<td>0.27</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Fraction of likelihood data eliminated by eliminating unlikely commute patterns is about 3-7% depending on state and year.
Support Points in Prior Domain (before pruning)

<table>
<thead>
<tr>
<th>Support points:</th>
<th>Large State (A)</th>
<th>Medium State (B)</th>
<th>Small State (C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support points:</td>
<td>Mean</td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>Total</td>
<td>1,005</td>
<td>583</td>
<td>2,067</td>
</tr>
<tr>
<td>By level of coarsening</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Super-PUMA</td>
<td>526</td>
<td>519</td>
<td>538</td>
</tr>
<tr>
<td>- PUMA</td>
<td>39</td>
<td>9</td>
<td>73</td>
</tr>
<tr>
<td>- Census Tract</td>
<td>438</td>
<td>32</td>
<td>1,506</td>
</tr>
<tr>
<td>By distance (in miles) between centroids</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- low-10</td>
<td>265</td>
<td>1</td>
<td>878</td>
</tr>
<tr>
<td>- 10-25</td>
<td>127</td>
<td>8</td>
<td>794</td>
</tr>
<tr>
<td>- 25-100</td>
<td>85</td>
<td>23</td>
<td>289</td>
</tr>
<tr>
<td>- 100-500</td>
<td>139</td>
<td>119</td>
<td>206</td>
</tr>
<tr>
<td>- 500-high</td>
<td>389</td>
<td>361</td>
<td>412</td>
</tr>
</tbody>
</table>
Refinement: Informative Priors

- In year 2002: Public-use CTTP data
- In year 2003-2008: Public-use previous year OnTheMap data (not posterior)
- \( \alpha = \max[\min_{\alpha}, f(\text{prior density})] \) minimum prior sample size is the larger of the PDP value (\( \min_{\alpha} \)) or the informative prior value
- Priors unique to each employment tract
- Not strictly Bayesian because the posterior is not published, and published data are required for prior by PDP
Refinement: Domain Pruning

• Domain may still have too many blocks for good analytical validity

• Algorithm 2 prunes the domain for a given destination \( j \):
  - Keep all origins in the likelihood support (confidential data)
  - For all other origins, add to domain with probability \( f_i \);
    (generates \( \min_p \) below)
  - From Machanavajjhala et al. 2008:

Theorem 5.2 (summary): Applying the domain pruning algorithm 2 changes the \((\varepsilon, \delta)\)-pdp to

\[
\varepsilon' = \varepsilon + \max_{i \in \{i | n_i = 0\}} \left( \ln \left( \frac{1}{f_i} \right) \right) + \max_{i \in \{i | n_i = 0\}} \left\lceil \alpha_i \right\rceil \ln 2
\]
Effects of Domain Pruning

- Domain pruning leaves all of the support points that appear in the likelihood function in the posterior.
- Domain pruning removes some of the prior support points that have no likelihood.
- Domain pruning improves analytical validity, but because it depends upon the confidential data, it increases the effective differential privacy limit.
Final Privacy Settings for OnTheMap V3

- Unadjusted $\varepsilon = 4.6$
- Probability of failure $\delta = 0.000001$
- Minimum retention probability $\min_p = 0.025$
- Adjusted $\varepsilon = 8.9$
- Kullback-Leibler and Integrated Mean Squared Error loss functions used to set parameters of prior
- Multinomial-Dirichlet Posterior sampled for every workplace block in the U.S. (about 1.4 million)
Analytical Validity Measures

• The divergence between posterior and likelihood for a population is measured by the Kullback-Leibler Divergence index (KL) and the Integrated Mean Square Error (IMSE) over a 29 point grid defined by the cross product of:
  – 8 commute distance categories (in miles: 0, (0-1), [1-4), [4-10), [10-25), [25-100), [100,500), [500+)
  – 5 commute direction categories (NW, NE, SW, SE, “N/A”)
• \( D_{KL} = 0 \) if identical; \( D_{KL} = \infty \) if no overlap

\[
D_{KL} (P \parallel L) = \sum_i L(i) \ln \frac{P(i)}{L(i)}
\]
Kullback-Leibler Divergence by Epsilon: All Populations

Epsilon

Kullback-Leibler Divergence (Posterior from Likelihood)

State C
State B
State A
Kullback-Leibler Divergence by Epsilon: Small Populations

- **State C**
- **State B**
- **State A**

Kullback-Leibler Divergence (Posterior from Likelihood) vs. Epsilon

- **Y-axis:** Kullback-Leibler Divergence (Posterior from Likelihood)
- **X-axis:** Epsilon

The graph shows the relationship between Kullback-Leibler Divergence and Epsilon for different states, indicating how the divergence changes with varying epsilon values.
Summary: Varying $\varepsilon$

- Figures show the population-weighted $D_{KL}$ for all and small (1 to 9) workforce populations for $\varepsilon = 2, 4, 4.6, 10$ and $25$
- Overall, $D_{KL}$ close to zero for values of $\varepsilon > 4$
- Significant gains in analytical validity for small populations as we increase $\varepsilon$ further to 4.6
- The marginal improvements in analytical validity from even higher values of $\varepsilon$ hard to justify in terms the costs in privacy protection loss
Kullback-Leibler Divergence and Prune-adjusted Epsilon by Minimum Retention Probability in Prune Function: All Populations

Adjusted Epsilon

Kullback-Leibler Divergence (Posterior from Likelihood)

Minimum Retention Probability in Prune Function

State C

State B

State A

Adjusted Epsilon

--- State C

--- State B

--- State A

--- Adjusted Epsilon
Kullback-Leibler Divergence and Prune-adjusted Epsilon by Minimum Retention Probability in Prune Function: Small Populations

- State C
- State B
- State A
- Adjusted Epsilon
Summary: Varying $min_p$

- Figures show the population-weighted $D_{KL}$ for all and small (1 to 9) workforce populations and $\varepsilon$ for $min_p = 0.1, 0.05, 0.025$ and 0.001
- Large gains in analytical validity as $min_p$ is decreased from 0.1 to 0.05 for all populations and further large gains for small populations as $min_p$ is decreased to 0.025
- The marginal improvements in analytical validity from even lower values of $min_p$; hard to justify in terms the costs in privacy protection loss
Summary: Varying $\delta$

- We evaluate $\delta = 0.001, 0.0001, 0.00001$ and $0.000001$
- Only very marginal improvements in analytical validity as we decrease confidence from 1 in a million to 1 in a 100
- No reason to consider values of $\delta > 0.000001$
### Posterior, Likelihood and Prior Mass across Commute Ranges for All and for Small Populations

<table>
<thead>
<tr>
<th>Distance</th>
<th>Post.</th>
<th>Lik.</th>
<th>Prior</th>
<th>Post.</th>
<th>Lik.</th>
<th>Prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.07</td>
<td>0.07</td>
<td>0.01</td>
<td>0.30</td>
<td>0.32</td>
<td>0.18</td>
</tr>
<tr>
<td>(0-1)</td>
<td>0.15</td>
<td>0.15</td>
<td>0.03</td>
<td>0.13</td>
<td>0.16</td>
<td>0.03</td>
</tr>
<tr>
<td>[1-4)</td>
<td>0.23</td>
<td>0.23</td>
<td>0.07</td>
<td>0.25</td>
<td>0.27</td>
<td>0.17</td>
</tr>
<tr>
<td>[4-10)</td>
<td>0.26</td>
<td>0.26</td>
<td>0.24</td>
<td>0.28</td>
<td>0.27</td>
<td>0.31</td>
</tr>
<tr>
<td>[10-25)</td>
<td>0.28</td>
<td>0.28</td>
<td>0.39</td>
<td>0.21</td>
<td>0.22</td>
<td>0.17</td>
</tr>
<tr>
<td>[25-100)</td>
<td>0.14</td>
<td>0.13</td>
<td>0.19</td>
<td>0.18</td>
<td>0.16</td>
<td>0.31</td>
</tr>
<tr>
<td>[100-500)</td>
<td>0.03</td>
<td>0.03</td>
<td>0.07</td>
<td>0.04</td>
<td>0.03</td>
<td>0.11</td>
</tr>
<tr>
<td>[500-high]</td>
<td>0.02</td>
<td>0.02</td>
<td>0.05</td>
<td>0.01</td>
<td>0.00</td>
<td>0.08</td>
</tr>
</tbody>
</table>
Overall Summary

• Synthetic data as an privacy protection algorithm is a promising alternative to traditional disclosure avoidance methods, especially when data representation is sparse

• Hard to quantify degree of disclosure protection – synthetic data methods may leak more information than intended

• OnTheMap version 3 demonstrates the successful implementation of formal privacy guarantees based on the concept of probabilistic ε-differential privacy

• To achieve acceptable analytical validity results with privacy guarantees requires experimentation


