NistDp3 – Differential Privacy 3
Write-up and Privacy Proof

John Gardner (gardn999)
jmg@john-gardner.net

Introduction:

For the third round of the NIST Differential Privacy series, synthetic data is created for the 1940 census data set. The provided data set is for Colorado, while the system test will use another location. This data set has a total of 98 fields. Not only are there more fields than in previous rounds, but many of these fields have a very large number of possible values. However, the epsilon values to be tested have been relaxed to 8.0, 1.0 and 0.3.

Pre-processing:

First, my solution defines the values allowed for all the fields in the input data set. Valid field values are defined based on the information from colorado-specs.json and codebook.cbk. By default, if a field is not in codebook.cbk, ”maxval” is used from colorado-specs.json to allocate maxval+1 bins from 0 to maxval. However, some of these fields not in codebook.cbk have had special bins assigned to them. This is important, because some of these fields would have a very large number of bins assigned otherwise. For example, VALUEH would have a million bins. Finally, there are the 5 state-dependent fields which we have been allowed to set distinct valid values for by sampling the data directly.

Highly correlated fields are then grouped together. The public Colorado data set was studied to determine which fields are highly correlated. This information is used to group highly correlated fields into single counting histograms. The resulting number of bins is the product of the number of bins in each field in the group. To preserve differential privacy, these grouping decisions are made without input from the data set to be privatized. Any fields which are not grouped with any others are placed into their own separate group. The result is the formation of 39 groups from the 98 fields. Finally, for each row of the data set, a single bin in each group's counting histogram is incremented by one.

Privatization and Privacy Proof:

Differential privacy is achieved by adding Laplacian noise to every bin in every group's counting histogram. The privacy budget is split equally among each group and each group's histogram has a sensitivity of one. So, for the purpose of adding noise, the epsilon used is the total epsilon divided by the number of groups: (epsilon per group) = epsilon / (number of groups). Therefore the scale of the Laplacian noise = 1/(epsilon per group) = (number of groups) / epsilon.

Post-processing:

A threshold cut is made after adding Laplacian noise in the privatization step. The threshold cut is constant for all bins in a histogram and proportional to scale * log10(number of bins in the histogram). This is necessary to prevent noise addition from producing a massive number of non-zero bins and greatly inflating the count total for some very large histograms. The bin count is set to 0 if less than the threshold cut or otherwise rounded to the nearest integer.

The synthetic data is finally written out using the noisy counts. For each row, a random bin is chosen from each group's histogram where the probability for a bin to be chosen is proportional to the bin's noisy counts. The field values corresponding to each histogram bin are written out in the corresponding output data column. All of these randomly selected field values are then used to write out a single row of synthetic data. A total of 800,000 rows are written this way.