Duke University OPM Synthetic Data Project
Privacy Assessment, Gender
T Balmat
Version 1.0, August 23, 2016
Version 1.1, September 28, 2016

Using version 0.6 of the OPM synthetic data, the 1988-2011 non-DOD status CPDF data received from OPM, and simple assumptions on potential attack strategies, the proportion of synthetic observations with true gender revealed is assessed.

In what follows, it is assumed that an attacker has access to:

- Publicly available federal employee data elements Year, Agency, Occupation, and Salary (these are readily available at various on line at sites, such as www.federalpay.org. Note that employee name and duty station state (U.S.) are also typically available, but neither are contained in the synthetic data.
- The entire database of synthesized Year, Agency, Occupation, BasicPay (Salary), and Sex observations.
- Joint frequencies of authentic Year, Agency, Occupation, and Salary, although they must be accumulated from individual on line queries.

It is assumed that an attacker does not have access to:

- Data synthesis algorithms.
- Any knowledge of potential mapping of synthetic to authentic records.
- Authentic Sex values or other data within the Duke Human Capital system.

To reduce the number of joint categories, BasicPay is truncated to thousands of dollars.

1. Joint Category Uniqueness

An attacker might compose a table of synthetic observation frequencies by Year, Agency, Occupation, and BasicPay then limit queries to joint categories with frequency 1, hoping that many carry Sex from their corresponding authentic observation. First, not all authentic Year, Agency, Occupation, and BasicPay joint categories exist in the synthetic data (22.6% are not represented, the average authentic frequency count of these is 1.3). Further, lacking knowledge of synthetic data algorithms, the attacker has no measure of expected proportion of accurate Sex values in the synthetic data. Joining frequency 1 authentic categories to corresponding synthetic categories reveals that 54.7% have identical Sex values. Recall that 22.6% of the authentic categories have no corresponding synthetic observation.

Figure 1 shows the distribution, within the authentic data, of joint Year, Agency, Occupation, and BasicPay categories by observation count. The x-axis corresponds to the number of observations within category and the y-axis corresponds to the number of categories. Note that, due to the range of values (1 to 100,000 and 1 to 1,000,000) both axes are on log-10 scale. It is readily observed that over 1,000,000 categories have frequency 1 (the actual quantity is approximately 1,600,000).

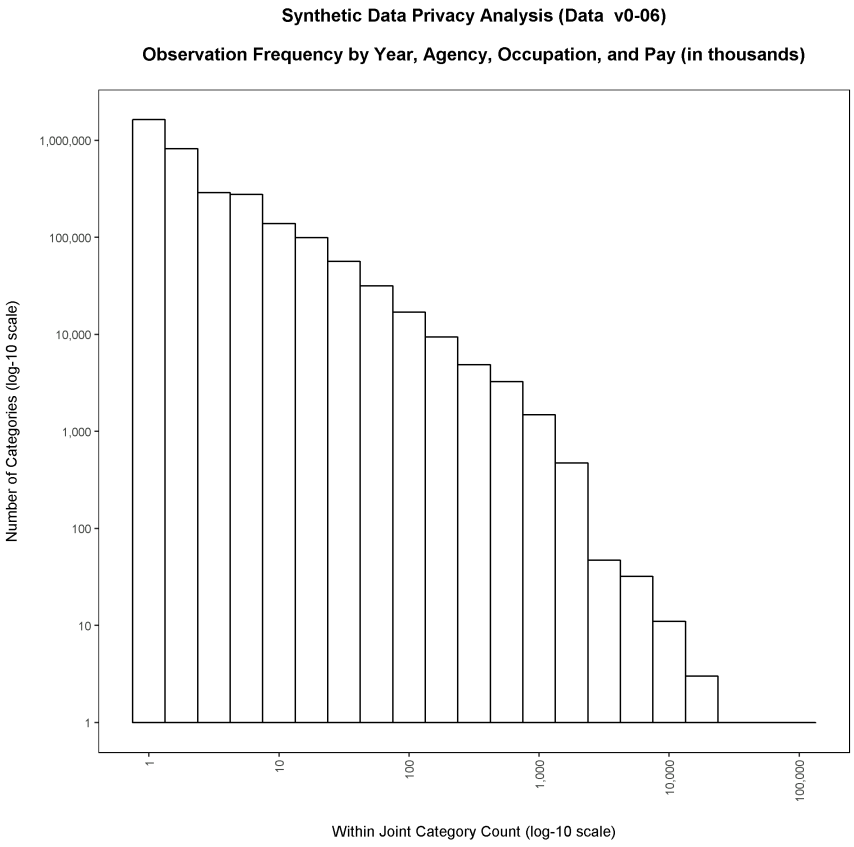

Figure 1, Year, Agency, Occupation, and BasicPay category frequency counts

2. Proportion Sex revealed using synthetic proportions and a classification rule

Assuming an attacker has no data on Sex (otherwise, why is there an attack?) other than the synthetic values, what is the proportion of true Sex values revealed? First, there are no record identifiers linking synthetic to authentic records, making a simple comparison of synthetic to authentic values impossible. One attack approach is:

- Assemble categories of Year, Agency, Occupation, and BasicPay from authentic public data.
- Calculate proportion female for corresponding joint categories in the synthetic data (each category forms a subgroup of observations).
- Classify all observations in subgroups with proportion female ≥ 0.5 as Female, those in subgroups with proportion female < 0.5 as Male. Given no knowledge of proportion gender reassignment between the authentic and synthetic data sets, this maximizes the apparent number of correctly encoded observations (synthetic gender equals authentic gender).

The proportion of true Sex revealed is estimated as follows:

- Using the classified gender of each synthetic subgroup, the number of correctly encoded observations is taken as the minimum of the synthetic subgroup observation count and the number of authentic observations for the corresponding joint category with gender corresponding to that in the synthetic group. Note that, although there may be more observations in an authentic data subgroup with gender equal to that in the corresponding synthetic subgroup, it is assumed that an attacker's observation count (number of observations for which gender is revealed) is taken from the synthetic data, since only that is visible to him.

- Proportion correctly encoded gender is calculated as the ratio of correctly encoded observations to the total synthetic observation count. Note that certain synthetic joint categories may not appear in the authentic data. Although they reveal nothing about authentic federal employee gender, they are included in the total observation count for proportion calculation. The proportion measures, of all synthetic observations, the number that reveal true gender.

Using this classification rule, the estimated proportion of true Sex revealed is 0.68, which is sufficiently low to be considered unreliable.

3. Distribution of differences in synthetic and authentic proportion female by Year, Agency, Occupation, and BasicPay category subgroup size

Close agreement of proportion female between synthetic and authentic subgroups improves synthetic data utility, but also increases the probability of true Sex classification. Along with Sex, the public identifiers (Year, Agency, Occupation, and BasicPay) were synthesized, resulting in masking of Sex through direct modeling and through the virtual transfer of federal employees throughout years, agencies, and occupations. Greater masking (lower probability of correct classification) is expected with greater frequency of large differences in subgroup proportion female and large differences in subgroup observation count. Figure 2 shows the distribution of synthetic to authentic subgroup frequency difference (as a proportion) for six levels of difference in proportion female. Synthetic and authentic proportion female are those observed in the respective data sets. It is apparent that higher variability in category frequency is associated with lower frequencies for all proportion difference cases. This suggests improved masking for low frequency categories.

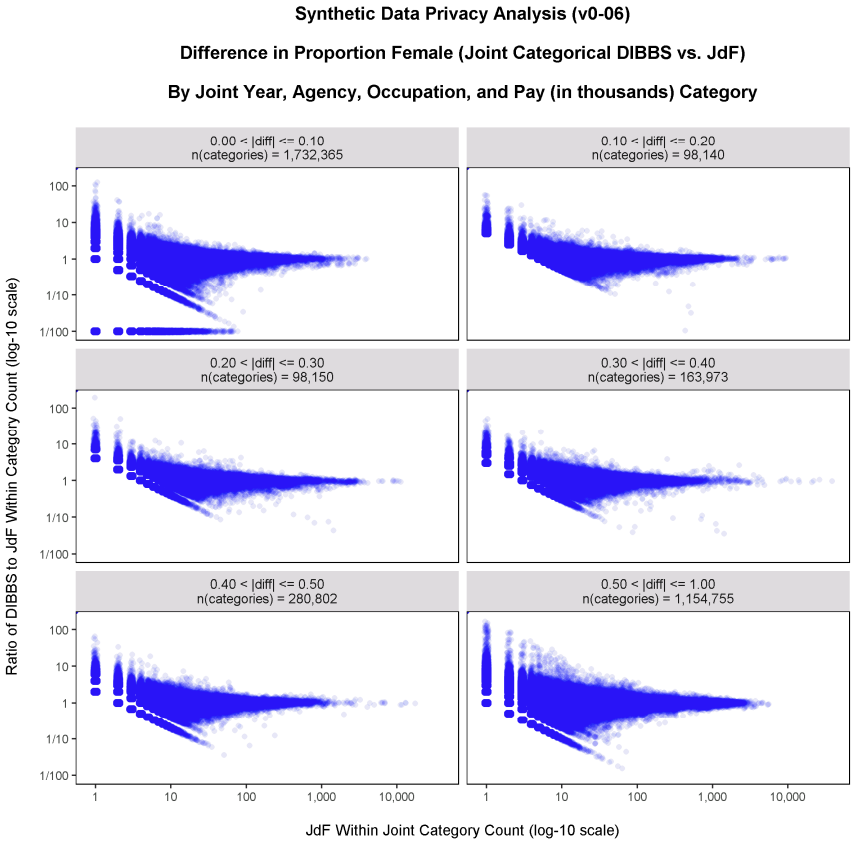

Figure 2, differences in synthetic and authentic data category frequency count using observed proportion female

Figure 3 shows differences in synthetic data category frequencies when synthetic proportion female is contrasted with marginal proportion female from the authentic data (0.48). Patterns are similar to those observed when difference in category proportion female are used (figure 2).

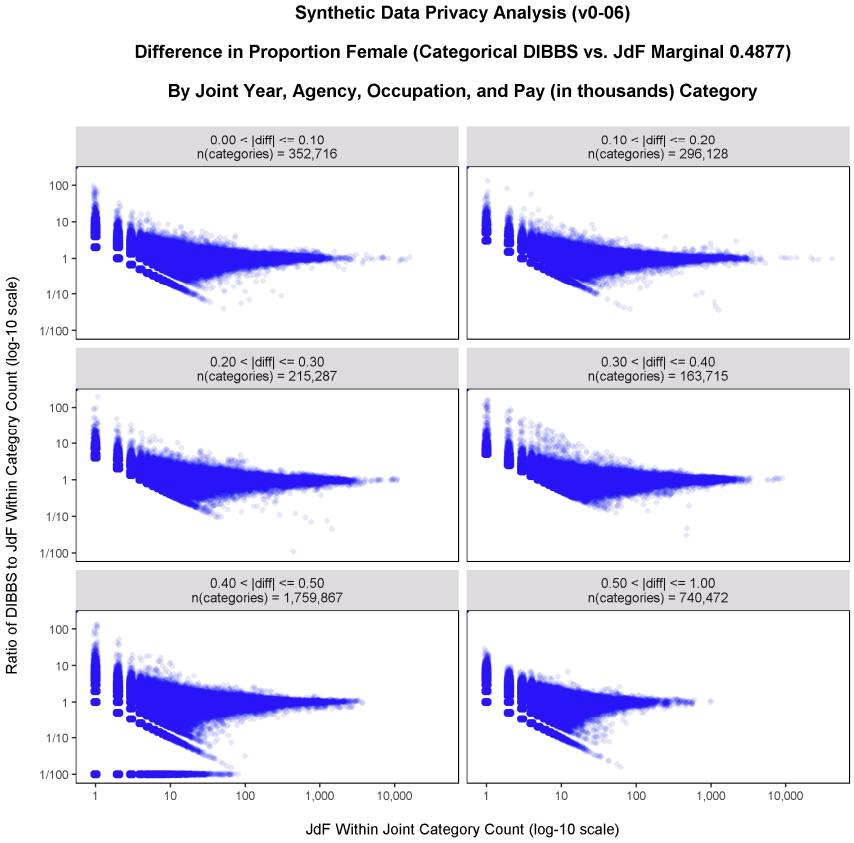

Figure 3, differences in synthetic and authentic data category frequency count using marginal proportion female

4. ROC curve, probability of correctly classified Sex

Using proportion female observed ($\hat{p}$) in synthetic data subgroups (one for each combination of Year, Agency, Occupation, and BasicPay) as estimates of authentic data proportions, it is of interest to measure true positive (Female) and true negative (Male) classification proportions or rates (TPR and TNR).

Method:

- Generate list of unique $\hat{p}$ values using synthetic data subgroups (on the assumption that an attacker has access to the entire synthetic data set and nothing in the authentic data).
- For each $\hat{p}$, classify all authentic observations in subgroups with proportion female > $\hat{p}$ as Female. Classify all observations in subgroups with proportion female $\leq \hat{p}$ as Male.
- Identify true positives (authentic Female observations classified as Female) and true negatives are (authentic Male observations classified as Male).
- Accumulate true positive and true negative classification quantities and convert to TPR and TNR by dividing true positives by number of authentic positives and true negatives by number of authentic negatives.
- Plot TPR by TNR. This is the ROC curve (figure 4).

- Measure area under the ROC curve (AUC) using trapezoidal areas bounded by neighboring TPR, TNR coordinates. High AUC indicates agreement between synthetic and authentic proportion by Sex within subgroups along with increased probability of revealing true Sex when observed synthetic $\hat{p}$ values are used as classifiers.
- Identify $\hat{p}$ associated with maximum simultaneous true positive and true negative frequencies; call it $\hat{p}_m$. Using $\hat{p}_m$ as a classifier maximizes the proportion of true Sex (Female and Male) revealed in the synthetic observations. An attacker might view 1- $\hat{p}_m$ as an error rate were he to know its value and employ it in a classification strategy.

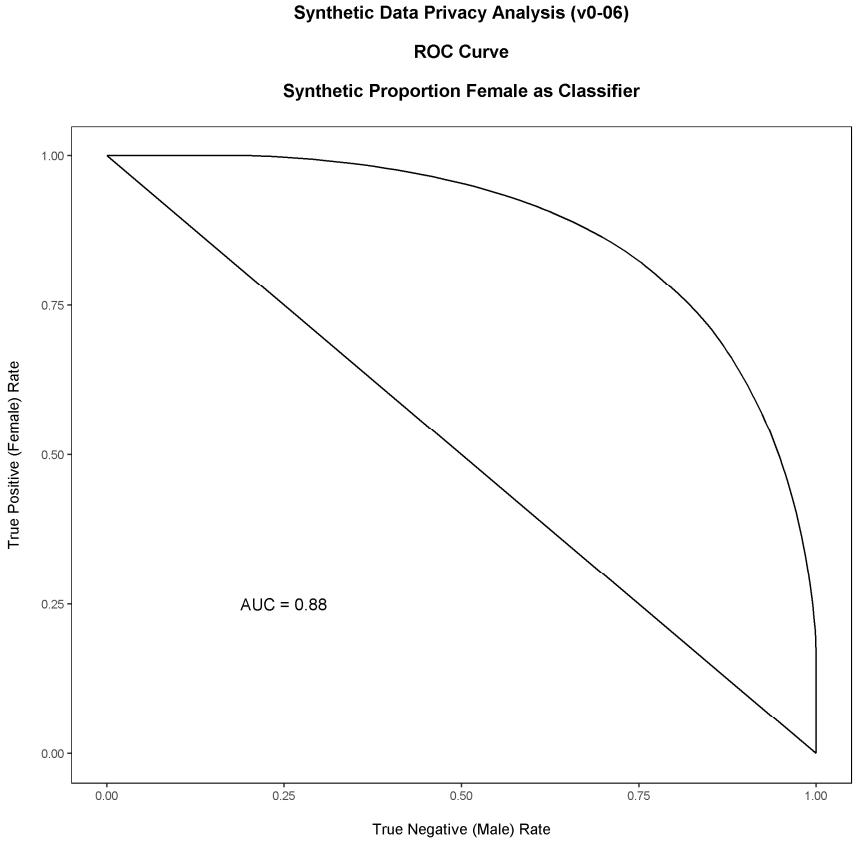

Figure 4, ROC curve using synthetic proportion female as classifier

Using the classification strategy described above (observed proportion > $\hat{p}$ assigned Female, $\leq \hat{p}$ assigned Male), each $\hat{p}$ generates simultaneous TPR and TNR values. These are plotted as coordinates which are then connected to form the ROC curve. Of interest are two classes of curves:

1. those with large AUC (AUC resides on the interval [0.5, 1.0]) indicating a high frequency of $\hat{p}$ values associated with high true Sex classification rates, and
2. those with low AUC (near 0.5), corresponding to TPR, TNR coordinates near the straight reference line and indicating a situation where choice of classification $\hat{p}$ (evaluating different values of $\hat{p}$) results in corresponding loss of TNR for any gain of TPR (the total true classification rate remains low).

Visually analyzing the above ROC curve reveals the existence of a $\hat{p}_m$with TPR, TNR coordinate near 0.8, 0.8. Using the marginal proportion female from the authentic data (0.48) this indicates that, knowing $\hat{p}_m$, an attacker can classify Sex accurately at approximately 0.8 probability (0.8 of both sex are accurately classified). An interesting

extension of this would be to estimate the prior probability of an attacker choosing $\hat{p}_m$, or of choosing from a specified vicinity about $\hat{p}_m$, and applying this to the true classification rate as a measure of overall probability of disclosure. For the record, the ROC curve generated here has $\hat{p}_m$ of exactly 0.50, a likely choice with no prior knowledge of Sex distribution. Low AUC is associated with flatter curves with weak TPR, TNR coordinates and poorly predicting $\hat{p}_m$ values. This is preferable to high AUC, from a privacy protection standpoint, but is also associated with poor utility (authentic data proportions are distant from corresponding synthetic $\hat{p}$ values).