Data Background

The gold standard file (GSF) integrates data elements from seven SIPP panels (1990, 1991, 1992, 1993, 1996, 2001, and 2004) with SSA-provided administrative data from the Summary Earnings Records (SER), the Detailed Earnings Records (DER), the Master Beneficiary Record (MBR), the Supplemental Security Record (SSR), the 831 Disability File (F831), and the Payment History Update System (PHUS). We then use regression-based multiple imputation to fill in the missing data of the GSF to create four completed data sets that are identical in structure and all non-missing values to the GSF but contain independent draws from a probability distribution replacing the missing values. We refer to these four datasets as the completed data. We then use the same modeling techniques to create 16 synthetic datasets (4 synthetic implicates are created from each of the 4 completed implicates). This process is like the imputation to complete the missing data, except now all values are replaced by independent draws from the estimated probability distributions. Every value of every variable (except type of SSA benefit, gender, and the first marital link observed in the SIPP) is synthesized. This document describes our assessment of the degree to which the synthetic data protects the confidentiality of respondents in our data.

The link between administrative earnings, benefits data and SIPP data adds a significant amount of information to an already very detailed survey and could pose potential disclosure risks beyond those originally managed as part of the regular SIPP public use file disclosure avoidance process. The creation of synthetic data is meant to prevent a link between these new public use files and the original SIPP public use files, which are already in the public domain. In addition, the synthesis of the earnings data meets the IRS disclosure officer's criteria for properly protecting the federal tax information. Our disclosure avoidance research uses the principle that a potential intruder would first try to re-identify the source record for a given synthetic data observation in the existing SIPP public use files, which were used to create the SIPP component of our Gold Standard file.

Because the SIPP is public use and our methods of creating the GSF are public, we assume that an intruder has access to all the SIPP data in the GSF but none of the administrative data. As a result, the only unsynthesized variables of use to an intruder as blocking variables in a re-identification exercise are gender and the first marital link (if any) observed in the SIPP. Because of the unsynthesized marital link, we used a wide-version of our GSF and synthetic data where a single record contains all the data for both members of a linked marriage. If there is no linked marriage, the record only contains all the data for that single individual. Table 3 shows the sizes of these unsynthesized cells in the GSF and the synthetic data. The first thing to notice is that the cells are much smaller for the single men and women and slightly smaller for the married couples in the synthetic data. This is because we apply an age cutoff of 15 as of the beginning of the SIPP panel in order to be kept in the synthetic data. Thus original SIPP respondents from the gold standard file are dropped from the synthetic data, but in a way that is unknown to an intruder. Because we synthesize birthdate and then use the *synthesized* birthdate when imposing the age restriction, an intruder cannot tell for sure from examining the public use SIPP which respondents were dropped and which were not. Thus our age cutoff combined with the synthesized birthdates adds an extra layer of uncertainty to any matching exercise performed by an intruder. In our matching exercise, however, we wish to be very conservative, and so we use the full synthetic data (prior to the age cutoff) in all of our re-identification exercises. Each of these synthetic implicate files has the same sample size as the gold standard and we know that a "true match" exists in the public use SIPP.

The unsynthesized type of SSA benefit data does create some small cells. These small cells are of no direct use to the intruder because an intruder does not have access to this information. However, the intruder could conceivably use these small cells to link records between synthetic data implicates. If a record is linked across synthetic data implicates, there is some concern that averaging variables across implicates could provide better matching variables in a re-identification exercise. As a result, we perform a version of our re-identification analysis where we assume that an intruder can link every record across all 16 implicates. The first row of Table 4 shows the number of cells (as defined by gender, marital link, and type of SSA benefit) in the GSF that are small (less than or equal to 10), the average number of people in each small cell, and the total number of records contained in these small cells. The rows following contain the same numbers for the synthetic data. Here we see that despite the age cutoff, the corresponding cells in the synthetic data are almost twice as large. This is because of the data completion. Respondents who were missing the administrative data in the GSF have had their SSA type of benefit completed, adding yet another layer of uncertainty to an intruder's re-identification effort.

Overview of Re-identification exercise

In order to test the effectiveness of the synthetic data in controlling disclosure risk, we conducted minimum distance matching exercise between the synthetic data and the Gold Standard. Since the Gold Standard contains actual values of the data items as released in the original SIPP public use files, the Gold Standard variables are the equivalent of the best available information for an intruder attempting to re-identify a record in the synthetic data. Successful matches between the Gold Standard and the synthetic data represent potential disclosure risks.

It is important to remember that for an actual re-identification of any of the records that were successfully matched to an existing SIPP public use file, an additional non-trivial step is required. This additional step consists of making another successful link to exogenous data files that contain direct identifiers such as names, addresses, telephone numbers, *etc*. Hence, the results from our matching process are a very conservative estimation of re-identification risk.

The Census Bureau Disclosure Review Board has adopted two standards for disclosure avoidance in partially synthetic data. First, using the best available matching technology, the percentage of true matches relative to the size of the files should not be excessively large. Second, the ratio of true matches to the total number of matches (true and false) should be close to one-half.

Distance matching

Distance-based record linking is another common approach to estimating the risk of disclosure in micro data. In recent work, cite: domingo-ferrer-abowd-torra-2006 use distance-based methods to re-identify records on two synthetic micro-data samples. They find that distance-based metrics perform similarly to (if not better than) the more commonly used probabilistic methods. Their work suggests that re-identification exercises should also include distance based methods. The broader the selection of methods used, the more informed the analyst is of the risk of disclosure. In particular, it is important to understand which methods pose the largest threat. cite: domingo-ferrer-torra-mateo-sanz-sebe-2006 conduct similar comparisons of distance-based and probabilistic record linking methods.

Our tests consider the case of an intruder who uses distance-based re-identification to match the source records from the Gold Standard to synthetic SIPP/SSA/IRS-PUF

observations. Such re-identification methods calculate the distance between a given record in the Gold Standard and every record in the synthetic implicate. The j closest records are then declared potential candidates for a match to the source record. In our analysis we consider $j = 3$.

Our distance-based re-identification proceeds in two stages. First we split both the Gold Standard and the first synthetic implicate ($m = 1$ and $r = 1$) into groups based on the unsynthesized variables. In this case, gender and the first marital link found in the SIPP are the only unsynthesized variables. We next split each blocking group into smaller segments of approximately 10,000 observations in order to decrease the processing time, which is quadratic in the size of the largest files compared. We performed the segment split on both the Gold Standard and synthetic files so that the correct match in the Gold Standard was always in the same block and segment of the synthetic data used for comparison. In other words, we forced the segmentation of the files to guarantee that the correct match could always be found in the block/segments being compared. The segmentation of the blocks uses our prior knowledge of which records are actual matches and hence our matching results are conservative–overestimates as compared to a distance record link that could not segment the comparison files because the intruder did not have access to the true person identifier. After splitting the data into blocking groups and segments, we then calculate the distance between a given Gold Standard record and every record in the synthetic file in its corresponding blocking group and segment using a set of 163 matching variables. The three closest records are then declared possible matches. If there is a marital link, a single record contains all the data for both spouses. If there is no marital link, the record contains all the data for a single individual.

We use four distance metrics. Each metric is a special case of either Mahalanobis or Euclidian distance. Before formally defining the distance, we first define some notation. Let A and B represent the two data sets being matched. For our purposes, conceptualize the block and segment of the Gold Standard as the A file and the block and segment of the synthetic implicate as the B file. Denote α as the vector of 163 matching variables from an observation in the A file and β as the analogue for the B file. Given this notation we define the distance between a given vector α in the A file and a given vector β in the B file as follows:

$$d(\alpha, \beta) = (\alpha - \beta)'[Var(A) + Var(B) - 2Cov(A, B)]^{-1}(\alpha - \beta)$$

We consider four specific cases of the general distance. In the first case we assume that the intruder can properly calculate the $Cov(A, B)$. We denote this distance $MAHA1$, and note that it is a true Mahalanobis distance; hence we expect that this distance measure will give us the highest match rates since it uses all of the available information, including the correct covariance structure of the errors in synthesizing all 163 variables. In the second case, we assume that the $Cov(A, B) = 0$. This is equivalent to assuming that we do not know how to link the observations across the A and B files and cannot compute $Cov(A, B)$. A real intruder would not have access to $Cov(A, B)$. We denote the second distance $MAHA2$, and note that it is a "feasible" Mahalanobis distance. In the third case, we assume $[Var(A) + Var(B) - 2Cov(A, B)] = I$, where I is the identity matrix. We denote the third measure as $EUCL1$, which is a Euclidian distance with unstandardized inputs. For the fourth measure, we transform all of the matching variables in the A and B files to $N(0, 1)$ variables. Call the transformed files $\tilde{A}$ and $\tilde{B}$. We then calculate the distance using $[Var(\tilde{A}) + Var(\tilde{B}) - 2Cov(\tilde{A}, \tilde{B})] = I$. We denote this fourth metric $EUCL2$, and note that it is a standardized Euclidian distance.

The adminsitrative type of SSA benefit data are also left unsynthesized. We did not use

these data as blocking variables because an intruder would not have access to this data. Moreover, this data is missing for all SIPP respondents for whom we do not have a validated PIK. As a result, the completion of missing data provides some protection against blocking with these variables. However, if we again take a conservative approach and assume the intruder can at least use this information to match respondents across implicates, the concern arises that averaging across the synthetic implicates for a given record might provide a better set of data to use in the minimum distance matching exercise. To address this concern, we assumed the intruder could match every record across all synthetic implicates and used the average value of every matching variable to calculate the minimum distance matches.

Tables 1-2 show the results of the re-identification exercises for each of the four metrics. Table 1 shows the results using just the 1st synthetic implicate, and Table 2 shows results using the average of all 16 synthetic implicates. All measures in both tables have matching percentages around or smaller than 1% (often much smaller). Moreover, the second best match is correct about as often as the best match (by best we mean the smallest distance), and the sum of the matching rate for the second and third best matches are almost always larger than the matching rate for the best match. Surprisingly, using all 16 implicates actually performs much worse than just using one implicate. This is likely because the averaging process moves the false matches closer to the candidate record as well as moving the true match closer.

Finally, we took an extremely conservative approach and reran the reidentification exercise from table 1 adding to the list of matching variables the entire SER/DER earnings history. The results of this exercise can be found in table 3. We once again find that the results do not change much. The matching rates are still for the most part less than 1%. The only exception was the least informed strategy, the Euclidean distance matching strategy, which achieved matching rates between 4% and 8%. We should keep in mind that this is with a blocking strategy that would not be feasible for any intruder (blocking on arbitrary blocks of 10,000 observations we know to be the same). Furthermore, even these higher rates do not represent any real threat to disclosure since the intruders confidence in having a correct match would still be less than 10%.

Conclusion

We are convinced that the disclosure risk posed by the release of the 16 synthetic implicates is extremely small and hence request that the Disclosure Review Board approve this release.

```
@INCOLLECTION{domingo-ferrer-abowd-torra-2006,
  author = {Domingo-Ferrer, Josep and Abowd, John M. and Torra, Vicenc},
  title = {Using Mahalanobis Distance-Based Record Linkage for Disclosure Risk
        Assessment},
  booktitle = {Privacy in Statistical Databases},
  publisher = {Springer-Verlag},
  year = {2006},
  editor = {Domingo-Ferrer, J. and Franconi, L.},
  pages = {forthcoming},
  owner = {John Abowd},
  timestamp = {2006.11.02}
}

@INCOLLECTION{domingo-ferrer-torra-mateo-sanz-sebe-2006,
  author = {Domingo-Ferrer, Josep and Torra, Vicenc and Mateo-Sanz, J.M. and
        Sebe, F},
  title = {Empirical disclosure risk assessment of the IPSO synthetic data
generators},
  booktitle = {Monographs in Official Statistics-Work Session on Statistical Data
        Confidentiality},
  publisher = {Eurostat},
  year = {2006},
  owner = {John Abowd},
  timestamp = {2006.11.03}
}
```

Table 1			Percent Correctly Matched To:			Ratio of best to:	
	Number of Blocks	Average Block Size	Best (smallest distance)	Second Best	Third Best	Second best	Second best + Third best
MAHA1							
couples	12	9996	0.536%	0.403%	0.370%	1.33	0.69
single women	18	10185	0.135%	0.135%	0.115%	1.00	0.54
single men	16	10344	0.187%	0.155%	0.151%	1.21	0.61
MAHA2							
couples	12	9996	0.069%	0.065%	0.059%	1.06	0.56
single women	18	10185	0.058%	0.047%	0.048%	1.24	0.61
single men	16	10344	0.041%	0.047%	0.044%	0.87	0.45
EUCL							
couples	12	9996	0.239%	0.120%	0.095%	1.99	1.11
single women	18	10185	1.962%	1.198%	1.014%	1.64	0.89
single men	16	10344	1.642%	0.934%	0.685%	1.76	1.01
EUCL STD							
couples	12	9996	0.068%	0.068%	0.059%	1.00	0.53
single women	18	10185	0.076%	0.062%	0.062%	1.23	0.61
single men	16	10344	0.069%	0.086%	0.086%	0.80	0.40

Table 2			Percent Correctly Matched To:			Ratio of best to:	
	Number of Blocks	Average Block Size	Best (smallest distance)	Second Best	Third Best	Second best	Second best + Third best
MAHA1							
couples	12	9996	0.097%	0.076%	0.073%	1.27	0.65
single women	18	10185	0.046%	0.044%	0.040%	1.06	0.55
single men	16	10344	0.037%	0.033%	0.041%	1.11	0.50
MAHA2							
couples	12	9996	0.067%	0.046%	0.064%	1.45	0.61
single women	18	10185	0.036%	0.030%	0.031%	1.19	0.58
single men	16	10344	0.028%	0.029%	0.024%	0.98	0.54
EUCL							
couples	12	9996	0.038%	0.033%	0.019%	1.14	0.72
single women	18	10185	0.025%	0.028%	0.029%	0.89	0.43
single men	16	10344	0.028%	0.025%	0.018%	1.11	0.64
EUCL STD							
couples	12	9996	0.128%	0.131%	0.100%	0.98	0.56
single women	18	10185	0.085%	0.071%	0.069%	1.19	0.60
single men	16	10344	0.058%	0.070%	0.053%	0.83	0.47

Table 3

File	Couples	Single Women	Single Men
GSF	119946	183328	165502
Syn11	119841	112932	91859
Syn12	119790	112890	91638
Syn13	119763	112575	91839
Syn14	119733	112630	92534
Syn21	119762	111503	92430
Syn22	119762	112872	91467
Syn23	119786	112446	91936
Syn24	119720	113482	91304
Syn31	119799	108680	87381
Syn32	119819	108567	87646
Syn33	119820	108175	87172
Syn34	119842	108418	87356
Syn41	119854	110136	89774
Syn42	119828	110463	90225
Syn43	119840	110972	90153
Syn44	119862	110906	89836

Table 4 Unsynthesized Cells Smaller <= 10

File	Number	Average Cellsize	Total Records
GSF	144	2.77	399
Syn11	144	4.35	626
Syn12	144	4.35	626
Syn13	144	4.35	626
Syn14	144	4.35	626
Syn21	144	4.44	639
Syn22	144	4.44	639
Syn23	144	4.44	639
Syn24	144	4.44	639
Syn31	144	4.33	623
Syn32	144	4.33	623
Syn33	144	4.33	623
Syn34	144	4.33	623
Syn41	144	4.26	613
Syn42	144	4.26	613
Syn43	144	4.26	613
Syn44	144	4.26	613

Table 5			Percent Correctly Matched To:			Ratio of best to:	
	Number of Blocks	Average Block Size	Best (smallest distance)	Second Best	Third Best	Second best	Second best + Third best
MAHA1							
couples	12	9996	0.681%	0.411%	0.360%	1.66	0.88
single women	18	10185	0.614%	0.107%	0.107%	5.71	2.86
single men	16	10344	0.700%	0.178%	0.158%	3.93	2.08
MAHA2							
couples	12	9996	0.144%	0.076%	0.071%	1.90	0.98
single women	18	10185	0.564%	0.078%	0.064%	7.23	3.98
single men	16	10344	0.622%	0.105%	0.094%	5.95	3.13
EUCL							
couples	12	9996	4.219%	0.030%	0.018%	140.56	87.52
single women	18	10185	8.108%	3.501%	1.715%	2.32	1.55
single men	16	10344	7.521%	3.662%	2.033%	2.05	1.32
EUCL STD							
couples	12	9996	0.096%	0.044%	0.040%	2.17	1.14
single women	18	10185	0.228%	0.081%	0.072%	2.81	1.49
single men	16	10344	0.308%	0.069%	0.056%	4.43	2.46