WP. 16
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Geneva, Switzerland, 9-11 November 2005)

Topic (ii): Disclosure risk, information loss and usability of data

DISCLOSURE RISK ASSESSMENT THROUGH RECORD LINKAGE

Supporting Paper

Submitted by the U.S. Bureau of the Census, United States of America[1]

[1] Prepared by Sam Hawala (sam.hawala@census.gov), Martha Stinson and John Abowd, Longitudinal Employer-Household Dynamics Project.

Disclosure Risk Assessment through Record Linkage

Sam Hawala[*], Martha Stinson[**], John Abowd[**]

[*] Statistical Research Division, U.S. Bureau of the Census, 4700 Silver Hill Road, Washington, DC 20233, USA, sam.hawala@census.gov
[**] Longitudinal Employer-Household Dynamics Project

Abstract: We present an example where we worked to assess the disclosure risk of files intended for public release or Public Use Files (PUF). These files contain synthetic data, created from a confidential data file. We used automatic record linkage experiments to assess the risk of disclosure from these files. We matched the PUF files to the confidential data file from which they were originally constructed. This example has to do with the Longitudinal Employer-Household Dynamics project (LEHD), developing data files containing linked information, matching selected worker and employer records for statistical research.

1 The Confidential File

The confidential file is a person-level file containing demographic data on individuals that were in the 1990-1996 panels of the Survey of Income and Program Participation (SIPP), linked to the Social Security Administration (SSA) administrative earning records data. The LEHD research group did the data linkage and the data cleaning for the confidential file, and is doing the preparation of the PUF planned for release. The confidential data file represents the kind of data that would be compiled for analysis by a researcher working in a protected area at either the Census Bureau or the other federal agencies that supplied the data.

The project to improve Census surveys by integrating administrative records from the Social Security Administration is a joint project between the Office of Research, Evaluation and Statistics at SSA and the LEHD Program at the US Census Bureau. It began in 2001, following the passage of a Treasury Regulation that enabled the enhanced linking. The project's goal is to develop data files containing linked information, matching selected worker and employer records for statistical research, in order to improve programs at the Census Bureau and SSA. The linked information comes from the Bureau's demographic and economic censuses and surveys, the Bureau's Employer Business Register (formerly known as the Standard Statistical Establishment List (SSEL)), which includes business tax information, and SSA's administrative records, which include information based on the IRS W-2 information return and information from SSA's various benefit programs. No information from the Business Register is included in the proposed public use file.

The Bureau obtained information from SSA's Master Earnings File (MEF), which contains Internal Revenue Service (IRS) data such as wages, tips, other

compensation and deferred wages. The MEF also contains the Social Security Numbers (SSNs) and Employer Identification Numbers (EINs), which permit the Bureau to link employee data with employer data. The IRS detailed earnings record information is provided separately for each employer of the employee. It covers all persons with wages, including non-filers and other non-covered employees; and it provides specific information on deferred compensation, such as retirement contributions.

The creation of the confidential file and the PUF is intended for communities of researchers interested in disability and retirement. The following list of variables presents an overview of the content of the confidential file. The file contains an attributed number (a person identifier) for each person and their spouse, if any. There are nine variables on the public use file that are copied exactly from the SIPP: sex, black/non-black, education (3 categories), marital status, age (3 intervals), marital status and the same variables for the spouse. There are also a host of additional SIPP variables that are subject to confidentiality protection using synthetic data methods. These include: birth date, hispanic/non-hispanic, education (5 categories), whether or not health limits the kind or the amount of work, number of children under 18, marital history, immigrant status, industry and occupation categories, total number of weeks worked in a year, annual total personal and family incomes, annual family total combined benefit dollars from government programs, total net worth, whether or not the individual is a homeowner, home equity, non-housing financial wealth, whether or not individual has a defined contribution or benefit pension plan, and a summary of the individual's annual health insurance status. Neither the confidential data file nor the proposed public use file contains Privacy Act protected identification information such as names, addresses, and social security numbers.

2 The Public Use Files

The PUF consist of several files. Each is a version ("implicate") of synthetic data constructed from the confidential data file, in the spirit of multiple imputation outlined by Rubin (1993). All SIPP variables from the confidential file are synthesized except for a few variables. For more details on the creation of the synthetic data, see papers by Abowd and Woodcock (2001, 2004), Rubin (1993), Fienberg (1994), Fienberg, Makov, and Steele (1998), Kennickell (1991, 1997, 1998, 2000), Raghunathan, Reiter, and Rubin (2003), and Reiter (2003).

Synthetic data have the advantage of making re-identification of respondents difficult while still providing analytically valid microdata to researchers in a format that they are accustomed to using. The synthetic public use data files are being prepared to closely mimic the characteristics of the confidential file. They provide analytically

useful data sets, while at the same time do not allow for re-identification of individuals in the already published SIPP public use files.

With the PUF, researchers at large, working in their own institutions, will have access to important demographic and economic information but some of the finer details of each person's record are synthesized to help preserve confidentiality. The disclosure avoidance standard is that individuals in the new PUF cannot be re-identified using the already published SIPP public use data products with a greater probability than a false re-identification. Linking of so much detailed administrative data to the SIPP necessitates this high standard of disclosure control.

In creating the synthetic data, LEHD's goal is to refrain from imposing prior beliefs about the relationships amongst variables and instead to allow the data themselves to determine the nature of these relationships. Thus, all variables can potentially be used as explanatory variables for the posterior predictive distributions of all other variables, even when such a relationship might not seem sensible to a social science researcher. In practice, due to feasibility issues, LEHD chooses some subset of variables to go on the right hand side of the predictive regressions but the goal remains to impose as few prior beliefs as possible. In this sense, the modelling done to create synthetic data is different than modelling done in order to predict future outcomes or to analyse cause and affect relationships.

Once the synthetic data are created, however, a different kind of analysis becomes necessary, where prior beliefs become important. Standard economic and demographic models must be tested using the synthetic data and analysts with experience evaluating such results must determine whether the synthetic data are statistically valid. Rubin (1996, p. 474) outlined what is meant by statistical validity:

- First and foremost, for statistical validity for scientific estimands, point estimation must be approximately unbiased for the scientific estimands averaging over the sampling and posited nonresponse mechanisms.
- Second, interval estimation and hypothesis testing must be valid in the sense that nominal levels describe operating characteristics over sampling and posited nonresponse mechanisms.

This definition should be modified to include the phrase "confidentiality protection mechanisms" wherever "nonresponse mechanisms" appears.

Thus in order to assess the quality and usefulness of the synthetic data, LEHD looks at several statistics of interest, calculates these statistics and averages them over the implicates of synthetic data, and then compares them to the best estimate of the same statistics from the confidential data. The estimates must be unbiased and the variances of the estimates must be such that inferences drawn about the estimates are similar to the inferences from the confidential data.

3 The Record Linkage Experiments

We performed the record linkage exercises using preliminary versions of the PUF. The LEHD research group continues to implement improvements for a final version. We matched each implicate PUF to the confidential file. The confidential file played the role of the already published SIPP public use data sets. We obtained similar results for each implicate, so we report on the results obtained for only one of them. We used a matching program based on the standard Census Bureau record linkage software written by Winkler *et al*. This standard software relies on the frequentist approach taken by Fellegi-Sunter (1969) to the probabilistic model of record linkage. It is used throughout the Bureau to create survey frames, to combine files, or to remove duplicates from files. Background on matching and some of the methods available in the software are described in research reports rr93/08, rr93/12, rr94/05, and rr99/04 at http://www.census.gov/srd/www/byyear.html.

The original purpose of the matching software we used was to extract plausible matching records, from a very large file A, that correspond to records in a smaller file B. The file B is assumed to fit into core memory. In our application, we treat the confidential file as file A, and one of the implicate files as file B. In our example, file A and file B are in fact the same size. Every person in the confidential file has one record in each of the implicate files. This fact further raises the bar for the disclosure testing. If file A were a large national file with many millions of records, matching to the smaller implicate files would be less successful. But the existence of the public use SIPP files for all the panels in the 1990s limits the size of file A to just under 250,000 people.

The software compares record pairs from the two files A and B when they agree on a specified blocking criterion. In order for the best matching pairs to be selected, the files must first be sorted according to this blocking criterion. The program outputs a file of records from the PUF that are plausible matches to records in the confidential file. The standard Census Bureau record linkage program features one-to-one matching that result in each record being paired with its most likely match within its blocking group. The matching program we used does not do this; rather, an output file may contain several records from the PUF that were scored as likely matches to the same record in the confidential file.

The matching process has two general steps. First, records are matched on blocking variables. The purpose of blocking variables is to identify a sub-set of records on both files A and B that share a set of common values necessary in order for matching to take place. Blocking variables make the searching more efficient by eliminating the need to compare every record in file A to every record in file B. Second, automatic searches for matches occur only within those records sharing the same values on the blocking variables. Matches agree exactly on values for the blocking variables and, additionally, they agree on values for the matching variables, although not necessarily in an identical manner. An input file to the matching software

specifies the agreement criterion for each of the matching variables. From the agreement criterion, the software computes a score, or weight. For each record in B, the program determines the matching comparison weights with records in A that share the same values of the blocking variables. If any of the comparison weights exceeds a cut-off value, the A record is written out to an output file. Another file contains the pairs of matched records. Finally the common person identifier on both the A and B file records is compared in order to determine whether the match is true or false. Thus our testing determines how many matches can be obtained by comparing the confidential and implicate files and what percentage of these matches are actually correct.

When the PUF are finally publicly released, there will be no link between the confidential data and the synthetic implicate files. However for testing purposes, we have maintained this link by keeping the common person identifier on the confidential file and all implicates of the Preliminary PUF. Thus by using this person identifier, we can check which matched record pairs with a given weight are correct matches and which are not by comparing this person identifier. When the person identifier is the same, the matching algorithm was successful in finding the person in the confidential file to whom the synthetic data record belonged. When the person identifier is different, the matching algorithm was unsuccessful. In our testing we declare any pair of records with a cumulative weight greater than 0 to be a match so all pairs with positive cumulative agreement weights are considered matches.

We report results on a matching run where we chose 5 blocking variables and 10 matching variables. The blocking variables we used were the variables that were not synthesized. They are: sex, black/non-black, education (3 categories), marital status, age (3 categories), plus these same variables for a spouse if one is present. The matching variables were all categorical except birth-date (month, year), which was converted to number of days since earliest birth-date on the confidential file. The other variables were hispanic/non-hispanic, education (5 categories), and immigration status, whether or not health limits the kind or the amount of work, whether or not the individual is a homeowner, number of children under 18, marital history, industry and occupation categories.

For all the categorical matching variables, we used the exact string comparison which assigns either the full agreement or disagreement weight based on whether the variable on the implicate file is the same or different from the variable on the confidential file. If the value of a variable is missing, the record will automatically be considered to agree on that variable. In practice this is unnecessary because the unsynthesized variables are never missing but in principle it ensures that we enable the most matches possible.

The conditional matching agreement probability is defined as the probability that the A and B values for this variable agree given that the records are truly a match. The conditional non-matching agreement probability is defined as the probability that the

A and B values for this variable agree given that the records are not a match. These probabilities are used to calculate the weights given to this variable when it agrees or disagrees. The probabilistic record linkage model defined by Fellegi and Sunter (1969) assigns the weight $\log(\frac{m_k}{u_k})$ if the records agree on the k^{th} variable, m_k being the matching agreement probability and u_k being the non-matching agreement probability. It assigns the weight $\log(\frac{1-m_k}{1-u_k})$ if the records disagree on the k^{th} variable . The software compares the values for each variable designated for matching, decides whether the values agree or not, and then assigns the appropriate weight to the variable based on the user supplied probabilities. Then a cumulative weight is calculated by summing the weights across all the variables designated for matching. This cumulative weight is the ultimate determiner of whether two records match. It is compared to the cut-off values provided by the user and if it passes the stated threshold, a match is declared. The relative matching and non-matching agreement probabilities chosen by the user control the influence of one variable relative to another on this cumulative weight. The non-matching agreement probability essentially tells how often a variable will agree at random across two files. A high value for this probability will reduce the importance of this variable in the matching by causing the agreement weight to be lower. This is desirable because if the variable is likely to agree at random, any match in values between the A and B files is less likely to signify a true match. At the same time, a high non-matching agreement probability causes the disagreement weight to be less negative or smaller, meaning that the penalty for not matching on this variable is not as high. In contrast, the relative matching agreement probability tells the importance of this variable compared to other variables in determining whether two records are a match. A high matching agreement probability means that a match on this variable is crucial to determining an overall match between 2 records. Thus a high value for m produces a high agreement weight. It also produces a more negative or higher disagreement weight, more severely penalizing non-matching in this variable. Blocking variables are essentially matching variables that have m = 1.

Except for birth-date, we assigned a conditional matching agreement probability of 0.7389 and a conditional non-matching agreement probability of 0.1 to all the other 9 categorical variables. The full agreement weight corresponding to a match on any variable besides birth-date is then log(.7389/0.1), or approximately 2. The disagreement weight is log(.2611/.9) = -1.2375. For birth-date, we use the comparison for pseudo-continuous variables that assigns the entire agreement weight if the value for birth-date is within 30 days and otherwise assigns some portion of the agreement weight depending on the difference between the confidential birth-date and the PUF birth-date. To understand how this works, consider an example of person A from the confidential file whose actual birth-date was Jan. 1, 1908. We

transform person A's birth-date into a variable (num1) that gives the number of days since the earliest birth-date on the file. The earliest birth-date is Jan. 1, 1907 so the new variable, num1, is equal to 365 for person A. Suppose that the PUF contains a synthesized birth-date for person A equal to July 1, 1912. We perform the same transformation and create a variable num2=2008 on the PUF. We then calculate a pooled standard deviation s_p of the original and synthesized birth-date variable in SAS date form. Then the comparison between the confidential and PUF birth-date takes place for person A, the weight is calculated using the following formula.

$$weight = agreewgt - (agreewgt - disagreewgt) * \frac{|num1 - num2|}{s_p}$$

The absolute value of the difference between num1 and num2 is scaled by the pooled standard deviation and then multiplied by the difference between the agreement and disagreement weights and subtracted from the agreement weight. In other words, differences between the confidential birth-date and the implicate birth-date are translated into a factor that is used to lower the agreement weight. If the disagreement between the two birth dates becomes large enough, the factor will overwhelm the agreement weight and result in a negative weight being assigned. For person A, $agreewgt = \log(\frac{.9}{.045}) = 2.9957,$ $disagreewgt = \log(\frac{.1}{.955}) = -2.2565,$ $|num1 - num2| = 1643,$ and $s_p = 9417$. Thus,

$$weight = 2.9956 - (2.9957 - (-2.2565)) * \frac{1643}{9417} = 2.0793$$

The confidential and implicate birth-date disagreement by 1643 days results in an agreement weight that is only 70% of the full agreement weight. Unlike any categorical variables that happen not to agree exactly for person A, the birth-date will make some positive contribution to cumulative weight even though it will not be the full amount possible. In cases where $|num1 - num2| \leq 30$, the birth-date is determined to agree exactly and the full agreement weight is assigned. This reflects the fact that because day of birth is not known on the public use SIPP files and on the confidential file, we create the full birth-date by randomly assigning the person a day within the reported birth month. Hence any birth-date values that agree within one month, match on all publicly available data.

4 Discussion of Results

The 5 unsynthesized variables available for all individuals (there are nine for married individuals) create 136 unique combinations (cells). There are some cells that will present disclosure problems simply by virtue of the fact that the cell contains (cell

size) only 1 or 2 individuals. Those cells where we could correctly match large numbers of records also represent disclosure problems. There were a total of 33,771 true matches and 26,174 false matches. The numbers of true and false matches vary considerably from one cell to another and do not appear to be tied to the cell size. What we consider informative is the ratio of number of true matches to false matches (tm/fm). When a cell has a tm/fm ratio much greater than 1 then the cell represents a disclosure problem. Indeed, an outside person doing the matching would obtain a total number of matches where a much higher percentage of them would be true matches. The outsider would be right much more often than they were wrong. When the ratio tm/fm is close to 1, the outsider would not be able to distinguish the true from the false matches by just guessing at random. So the ratio tm/fm was the most useful statistic for highlighting cells with problems.

For example, there are 26,590 individuals that are white males with a high school degree, married, and age 18-62. In this cell 18% (4775) matches correctly from the confidential to the implicate file. However, another 17% (4562) matches incorrectly. Thus the ratio of true to false matches is just slightly higher than 1. Hence this cell is of less concern because an outside person attempting to match the public use SIPP files to the implicate files would obtain approximately 9337 matches but only about half of them would be correct matches and the outside person would have no way to distinguish which matches were correct and which were not. Of much greater concern are black males with a college degree, married, and age 18-62. There are 658 individuals in this cell and 7% (49) are correctly matched. However only 2.7% (18) are incorrectly matched and so the ratio of true to false matches is 2.72. Thus an outside person doing the matching would obtain approximately 67 matches and a much higher percentage of them would be true matches. The outsider would still not be able to distinguish the true from the false matches but just guessing at random, the outsider would be right much more often than wrong.

Cells with a high number of true matches relative to false matches are not necessarily small cells as defined by the blocking variables. For example, a cell has 6572 individuals and also has a true to false match ratio of 3.69. Conversely for cells with under 100 people but more than 1 person, there is only one with a tm/fm ratio over 1. To make this point more clearly, Table 1 and Figure 1 summarize the findings. We sum the number of non-matches, true matches, and false matches for groups of 20 cells and plot a bar showing the percentage of each type of record. The size cut-offs of the cells are listed on the x-axis. For example, the first bar aggregates across the first twenty cells, which range in size from 1 to 9. Of the 83 total records in this grouping of cells, almost 84 percent do not match across the confidential and the implicate files. Another 12 percent match correctly and 4 percent match incorrectly. Each bar in Chart 1 represents a grouping of 20 cells except for the last bar, which contains only 16. From Table 1 and Chart 1 it is evident that larger cells have both more true and false matches but they do not necessarily have lower tm/fm ratios. Although the smallest cells have a very high number of true matches relative to false

matches (12% versus 4%, tm/fm ratio=3.33) and this same comparison is much better for the largest cells (16.3% true to 13.5% false matches, tm/fm ratio=1.21), the group with the closest number of true and false matches is the second bar (7.7% true matches versus 6.6% false matches, tm/fm ratio=1.16) whose cells range in size from 11 to 61. The third, fourth, fifth bars all have tm/fm ratios just under 1.5 and the sixth bar has a ratio just over 1.5. So there is no monotonic change in the tm/fm ratio as the cell sizes increase.

Figure 2 is a plot of number of matches vs. weights, for weights at or above 7.2684 in red for true matches and in blue for false matches. There are spikes where the weight equals 7.2684, 9.2784, 10.5059, 12.5059, and 13.7434. A cumulative weight of 7.2684 means exactly matching on two fields and within-a-month-matching on birth date. The additional spike points happen where an additional variable exactly agrees. Although the spike points account for similar overall percentages of both the true and false matches, the number of true matches relative to false matches at each one of these points is very high - somewhere between a 3 and 5 to 1 ratio. This leads us to conclude that for records with a high cumulative matching weight, the match is likely to be correct. Thus the blocking and matching variables are providing the software with some power in identifying the same person in both the confidential and implicate files

Cell Size Category	Non-matches	True Matches	False Matches	%true	%false	tm/fm
1 to 9	70	9	3	0.109756	0.036585	3
11 to 61	710	64	55	0.077201	0.066345	1.163636
65 to 208	2273	264	179	0.097202	0.065906	1.47486
242 to 501	5586	665	462	0.099062	0.068822	1.439394
607 to 918	12463	1363	947	0.092263	0.064103	1.439282
1149 to 3966	35935	6292	3782	0.136756	0.082201	1.66367
4095 to 31256	108205	25121	20746	0.163047	0.134651	1.210884

Table 1: Summary of Groups of 20 Cells

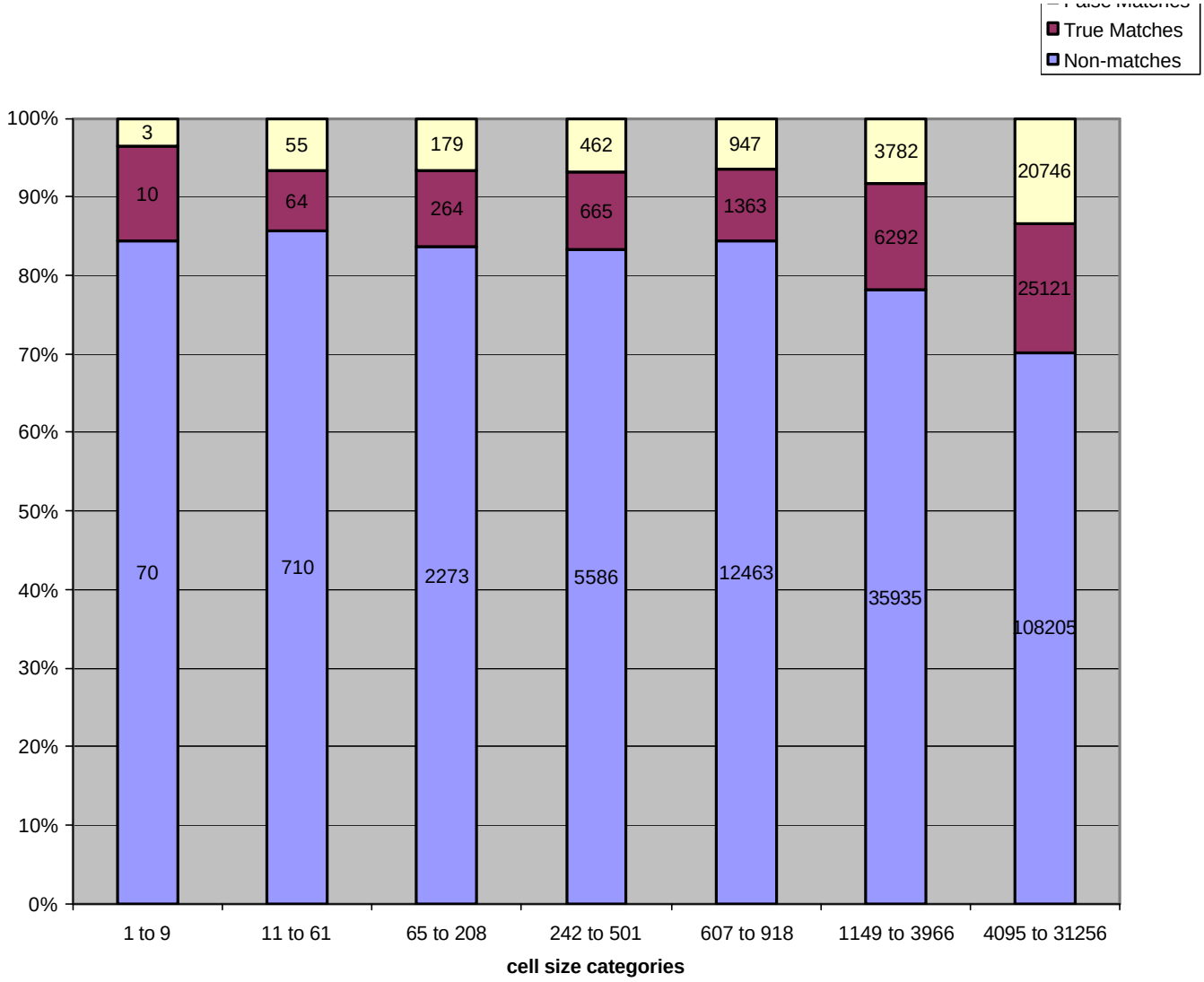

Fig 1 Percentage of Non-matches, False Matches, and True Matches

5 Conclusion

The results from the matching procedure performed on the Preliminary Public Use File give cause for some concern and some cautious optimism. There are many cells where the synthetic data are properly perturbing matches between the confidential and implicate files to the point where there are at least as many false matches as true matches. However there are also problematic cells where there are a disproportionate number of true matches. We are working on performing the best matching possible by choosing our conditional matching and non-matching agreement probabilities in an optimal manner. Any strategies employed to reduce disclosure risk will have to be measured against their effect on the analytic validity of the file. Hence at this point it is too early to make decisions about specific steps we will take to handle the problematic cells. We will repeat our matching procedure on the next version of the Preliminary Public Use File and re-evaluate how many cells have a ratio of true to false matches greater than one. At that point, different actions may be necessary to solve any remaining disclosure problems.

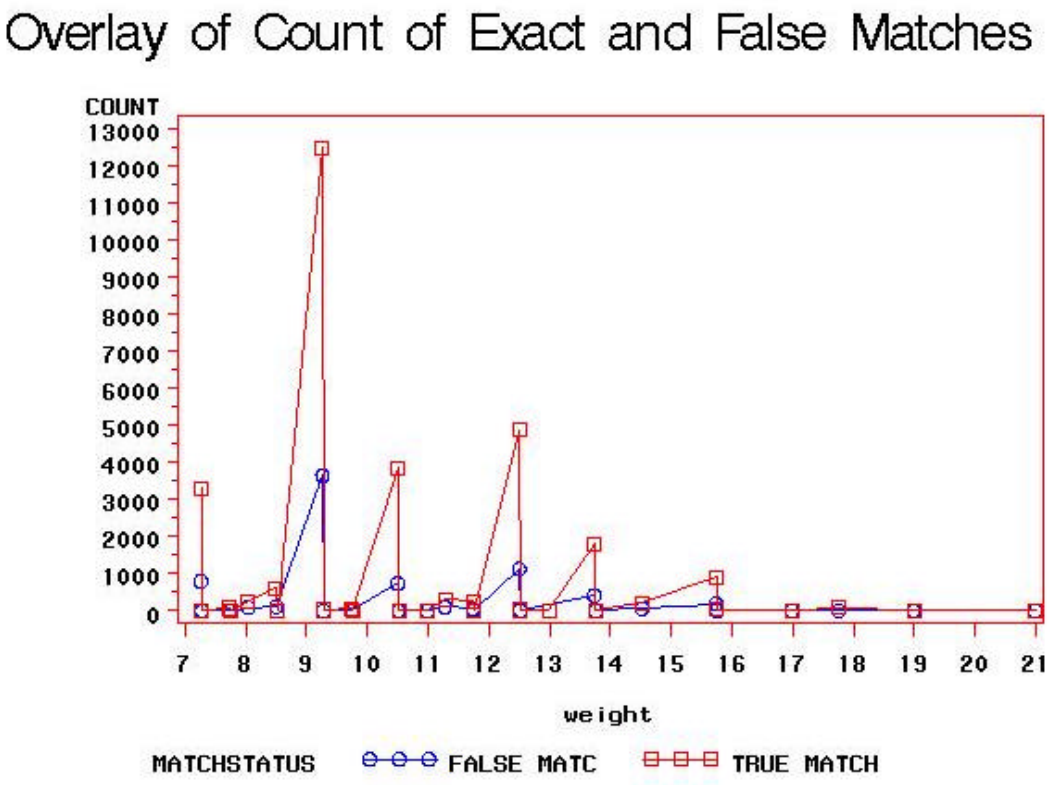

Fig 2 Plot of number of matches vs. weight.

References

Abowd, John M. and Simon Woodcock, *Disclosure Limitation in Longitudinal Linked Data, in Confidentiality, Disclosure and Data Access*: Theory and Practical Applications for Statistical Agencies, P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz (eds.), (Amsterdam: North Holland, 2001), 215-277.

Abowd, John M. and Simon Woodcock, *Multiply-Imputing Confidential Characteristics and File Links in Longitudinal Linked Data,* in J. Domingo-Ferrer and V. Torra (eds.) Privacy in Statistical Databases (New York: Springer-Verlag, 2004), pp. 290-297

Fienberg, S. E. (1994), *A radical proposal for the provision of micro-data samples and the preservation of confidentiality*. Carnegie Mellon University Department of Statistics Technical Report No. 611.

Fienberg, S.E., U.E. Makov, and R. J. Steele (1998), *Confidentiality, uniqueness, and disclosure limitation for categorical data.* Journal of Official Statistics 14(4), 485-502.

Fellegi, I. P., and Sunter, A. B., (1969), *A Theory for Record Linkage.* Journal of the American Statistical Association, 64, 1183-1210.

Kennickell, A. B. (1991), *Imputation of the 1989 Survey of Consumer Finances: Stochastic relaxation and multiple imputation.* SCFWorking Paper, prepared for the Annual Meetings of the American Statistical Association, Atlanta, Georgia, August 1991.

Kennickell, A. B. (1997), *Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances.* SCF working paper.

Kennickell, A. B. (1998), *Multiple imputation in the Survey of Consumer Finances.* SCF
working paper, prepared for the August 1998 Joint Statistical Meetings, Dallas, TX.

Kennickell, A. B. (2000), Wealth measurement in the Survey of Consumer Finances: methodology and directions for future research. SCF Working Paper, prepared for the May 2000 annual meetings of the American Association for Public Opinion Research, Portland, Oregon.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003), *Multiple imputation for statistical disclosure limitation.* Journal of Official Statistics 19, 1—16.

Reiter, J. P. (2003), *Inference for partially synthetic, public use microdata sets.* Survey Methodology 181—189.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Wiley

Rubin, D. B. (1993), *Discussion: Statistical disclosure limitation.* Journal of Official Statistics 9(2), 461-468