RESEARCH REPORT SERIES
(*Statistics #2002-08)*

Single-Ranking Micro-aggregation and Re-identification

William E. Winkler

Statistical Research Division
U.S. Bureau of the Census
Washington D.C. 20233

Report Issued: November 22, 2002

Single-Ranking Micro-aggregation and Re-identification

William E Winkler[1], william.e.winkler@census.gov 2002Oct24

Abstract

This paper shows that it is possible to create metrics for which re-identification is straightforward for situations in which continuous variables have been micro-aggregated one at a time using conventional methods.

Introduction

The purpose of this document is to provide overall justification on how micro-aggregation generally can yield public-use files in which re-identification rates are extraordinarily high. This work provides intuition and reasoning that supplements empirical evidence suggesting that re-identification rates may be high with standard micro-aggregation (see e.g., Domingo-Ferrer and Mateo-Sanz 2001, Domingo-Ferrer et al. 2002). Much of the earlier work on micro-aggregation concentrated on the degradation of analytic properties. The earlier work typically did not contain re-identification experiments.

Much of the previous micro-aggregation work (Domingo-Ferrer and Mateo-Sanz 2001, 2002; Defays and Anwar 1998) has generally concentrated on analytic properties (such as regression). The general rules are that the data values associated with individual variables in records will be put in groups of approximately size k where k is typically between 3 and 10. The original values are replaced with micro-aggregates, typically the averages within groups. As observed by Domingo et al. and others, as k increases toward 10, the analytic properties (i.e., regressions, etc.) can deteriorate severely. Generally, to reduce the deterioration of the analytic utility of the micro-aggregated data, the papers have taken k=3 or k=4. Although re-identification experiments were typically not done, the assumption was that re-identification would be more difficult as k increases. The re-identification understanding, however, was based on using variables in isolation from one another (i.e. *single-ranking* micro-aggregation). It did not consider looking at combinations of variables as might be done with nearest-neighbor matching (exceptions being the recent work of Domingo et al. 2001).

We do not consider the deterioration of analytic properties due to a combination of micro-aggregation and sampling. Domingo-Ferrer and Mateo-Sanz (2002) have shown that it is possible to micro-aggregate on several variables at once. Although the procedures are more difficult theoretically and computationally, they provide lower re-identification rates at the same values of k than the single-variable aggregation methods. The multi-variable aggregation can cause more severe deterioration of analytic properties. We do not consider multi-variable aggregation in this paper.

Basic Situation: Identifying a Micro-aggregated Database Against the Original Database

We consider a rectangular data base (table) having fields (variables) X_i i=1,…,n and value states x_{ij}, j=1,…,n_i. In many microdata confidentiality experiments, users want 10 or more variables X_i. We assume that each of the variables X_i is continuous, skewed, and not taking zero value states. The second assumption eliminates a few additional technical details. It can easily be eliminated. The third assumption is for convenience. It is not generally needed for the arguments that follow.

We begin our discussion by considering databases with 1000 or more records and situations in which micro-aggregation is on one variable at a time. Although sampling may reduce re-identification rates in some situations, it can also cause severe additional deterioration in analytic properties. We do not consider the deterioration of analytic properties due to a combination of micro-aggregation and sampling.

In this discussion, we demonstrate that micro-aggregation as currently practiced allows almost perfect re-identification with existing record linkage procedures even when k is greater than or equal 10. We can easily develop nearest-neighbor methods with similar metrics that have almost 100 percent re-identification rates.

We chose any three variables, say X_1, X_2, and X_3 that are pairwise uncorrelated (R^2 <= 0.2). Our procedure is for aggregating variables one at a time. Within each variable, sort the values and aggregate into groups of size 3 or more. Let the new micro-aggregated value-states be denoted by $a(x_{ij})= y_{ij}$, j = 1, …. k_i , i = 1, 2, 3 where a() is the aggregation function. Each (aggregated) value state is assumed three or more times (3 or more records have the same value of the y-variables). Most aggregates will be from three value-states only. In the following y_{iji} will

denote the j_i value-state of micro-aggregated variable Y_i. The micro-aggregated value y_{iji} will be a value such as the average or median. Such a value is in the range of the values being micro-aggregated. We develop new record linkage metrics (or nearest neighbor metrics) as follows. The metrics are for matching a micro-aggregated record R with the original set of data records. Let $R = (y_{1j1}, y_{2j2}, y_{3j3}) = (a(x_{1k1}), a(x_{2k2}), a(x_{3k3}))$ where y_i's are values aggregated by the aggregation operator a(.) from original values x_i's. Using the sort ordering for individual variables, for each i, let $p(y_{iji})$ be the predecessor of y_{iji} and $s(y_{iji})$ be the successor of y_{iji}. In each situation, the predecessor and the successor are distinct from the value y_{iji}. For y_{jii}, let the distance be metric dist (x, y_{iji}) be 1 if x in within distance min $(\text{abs}(y_{iji} - p(y_{iji})), \text{abs}(y_{iji}-s(y_{iji}))/2$ of y_{iji}; 0, otherwise. This allows us to match the X-variables in the original file with the Y-values in the micro-aggregated file. Suitable adjustments should be made for being at the end of the distributions (i.e., one-sided). Let N be the number of records in the original database. Then micro-aggregated record R has probability close to one of matching with its true corresponding original record. The probability is at least ((N-3)/N) on each field. It has probability close to zero of matching with any record other than its original corresponding record on each field.

We repeat the above argument. If micro-aggregated record R is matched against the original data using only variable X_1, then it can be matched against at most three records. The correct match is within the three records. Matching on variable X_1 quickly eliminates N-3 records from consideration. If we now match on variable X_2, there is a virtual certainty that we can identify the single record (of three) that R correctly matches. The intuition is that if record R matches on the first variable, then there are at most three records in the original data meeting that criterion (one of which is correct). The same thing happens on the second field; the same on the third. Typically, after two variables are compared, record R can be correctly matched. If k is increased from 3 to 10, then it is very straightforward to create new optimized metrics. Re-identification rates are still likely to be 100%.

Programming of the new metrics is exceptionally straightforward. One sorts on a variable, aggregates, and computes the new metric. The new metric is highly optimized for the given data and micro-aggregation procedure. Adaptation of the general matching (re-identification) software is also exceptionally straightforward.

First Extension: Identifying a 1% Sample of Micro-aggregated Data Against the Original Database

In this extension, we begin with a database D of 100,000 records having ten continuous variables. Again, for convenience, we assume that each of the variables X_i is continuous, skewed, and not taking zero value states. We aggregate in groups of approximately size k=3. We create a sample S containing 1% of the records. Again, we chose any three variables, say X_1, X_2, and X_3 that are pairwise uncorrelated ($R^2 <= 0.2$). Let $R = (y_{1j1}, y_{2j2}, y_{3j3}) = (a(x_{1k1}), a(x_{2k2}), a(x_{3k3}))$ where y_i's are values aggregated by the aggregation operator a(.) from original values x_i's. At this point, we use intuition from the first, much easier example. Pair record R with the approximately nine closest records in D. The pairing is according to the distance between the x_{1k1} values and y_{1j1}. Again, at least one of these nine will contain the correct match. Within the nine, compare the x_{2k2} –values with y_{2j2} to determine the plausible correct match. If the value y_{2j2} is not sufficient, use the remaining value y_{3j3}. Within three iterations (i.e., use of three variables), the correct match will be obtained. Repeat for all micro-aggregated records R until 100% of the micro-aggregated records have been correctly matched to their corresponding record in the population file D.

Second Extension: Identifying a 1% Sample of Micro-aggregated Data Against a Corresponding Database

By a corresponding database, we will mean a database D' that corresponds to D and is available to the intruder. We assume that it also contains 10 variables and that identifying information such as name is available in D'. If we can match a record in D' against a record in the micro-aggregated sample S, then a re-identification occurs. We assume that at most three variables in each record in D' have values that deviate by 30% from their corresponding values in D. We assume that the remaining variables in records deviate by at most 1-3% from the corresponding values in D. We consider restrictions similar to the previous two examples. We create a sample S containing 1% of the records. This time we use all ten variables. We only use some of the ideas from the previous example. Let $R = (y_{1j1}, y_{2j2}, \ldots, y_{10j10}) = (a(x_{1k1}), a(x_{2k2}), \ldots, a(x_{10k10}))$ where y_i's are values aggregated by the aggregation operator a(.) from original values x_i's.

For each variable X_i, i = 1, …, 10, we sequentially match record R as follows. Choose a group G_i of 360 records that agree most closely with y_{iji}. Let r' in D' be the record that matches R most closely in seven of the ten fields. By our previous reasoning, there will be a unique record in D' that agrees with R. Although record R will not agree with r' in D' on three fields, we can still find it. The redundancy of agreements allows us to overcome substantial error in three of the fields.

Discussion

More sophisticated re-identification methods than described in this document are routinely used in the record linkage of large administrative lists. A major problem with administrative lists is the amount of typographical error in name, address, date-of-birth, and fields such as income. The typographical errors that make it difficult to perform matching. Over a number of years, methods such as string comparators for text strings and optimized numeric metrics were developed for matching the lists. These methods translate naturally to the much simpler re-identification methods described in the first three technical sections of this paper.

Concluding Remarks

For researchers in methods of microdata confidentiality protection, there are two basic and complementary challenges. The first challenge is that the masked data that is created for public use should produce protected microdata that can be used for analytic purposes. There seems to be a consensus among researchers that the public-use file should allow valid approximate reproduction of means, variances, and one other statistic on a moderate number of subdomains. Single-variable micro-aggregation has sometimes been applied because it yields some analytic uses on the entire file. It does not typically allow analyses on subdomains. The second challenge is that the masked data should not allow re-identification. As we show in this note, single-variable micro-aggregation provides substantial structure for better re-identification methods. In the simpler situations, re-identification rates can be well above 20 percent.

1/ This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

References

Defays, D. and Anwar, M. N. (1998), "Masking Microdata Using Micro-aggregation," *Journal of Official Statistics*, **14**, 449-461.

Domingo-Ferrer, J. (2001), "On the Complexity of Microaggregation," presented at the UNECE Workshop On Statistical Data Editing, Skopje, Macedonia, May 2001.

Domingo-Ferrer, J. and Mateo-Sanz, J. M. (2001), "An Empirical Comparison of SDC Methods for Continuous Microdata in Terms of Information Loss and Re-Identification Risk," presented at the UNECE Workshop On Statistical Data Editing, Skopje, Macedonia, May 2001.

Domingo-Ferrer, J. and Mateo-Sanz, J. M. (2002), "Practical Data-Oriented Microaggregation for Statistical Disclosure Control," *IEEE Transactions on Knowledge and Data Engineering*, **14** (1), 189-201.

Domingo-Ferrer, J., Mateo-Sanz, J., Oganian, A., and Torres, A. (2002), "On the Security of Microaggregation with Individual Ranking: Analytic Attacks," *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems*, to appear.