On the Security of Microaggregation with Individual Ranking: Analytical Attacks *

Josep Domingo-Ferrer[1], Josep M. Mateo-Sanz[2], Anna Oganian[1], Vicenç Torra[3] and Àngel Torres[1]

[1] Universitat Rovira i Virgili, Dept. of Computer Science and Mathematics, Av. Països Catalans 26, E-43007 Tarragona, Catalonia, Spain, e-mail {jdomingo,aoganian,atorres}@etse.urv.es
[2] Universitat Rovira i Virgili, Statistics and OR Group, Av. Països Catalans 26, E-43007 Tarragona, Catalonia, Spain, e-mail jmateo@etseq.urv.es
[3] Institut d'Investigació en Intel·ligència Artificial, Campus de Bellaterra, E-08193 Bellaterra, Catalonia, Spain, e-mail vtorra@iiia.csic.es

Abstract. Microaggregation is a statistical disclosure control technique. Raw microdata (i.e. individual records) are grouped into small aggregates prior to publication. With fixed-size groups, each aggregate contains k records to prevent disclosure of individual information. Individual ranking is a usual criterion to reduce multivariate microaggregation to univariate case: the idea is to perform microaggregation independently for each variable in the record. Using distributional assumptions, we show in this paper how to find interval estimates for the original data based on the microaggregated data. Such intervals can be considerably narrower than intervals resulting from subtraction of means, and can be useful to detect lack of security in a microaggregated data set. Analytical arguments given in this paper confirm recent empirical results about the unsafety of individual ranking microaggregation.
Keywords: Microaggregation; statistical disclosure control; microdata masking; official statistics; data fusion; data security.

1 Introduction

The production of official statistics consists of three stages: data collection, data processing and data dissemination. One of the main problems in data dissemination is how to maximize the informational content delivered to the data users while preserving privacy of individuals (statistical confidentiality). A balance must be reached between the amount of information contained in the published statistical data and the desirable disclosure protection level.

* Work partly supported by the European Commission under project IST-2000-25069 "CASC"

Before releasing statistics computed on confidential data, the statistical office must make sure that identification of individuals is not easy. However, it is not possible to completely eliminate the risk of disclosure, as the released statistics must somehow reflect the reality of the population of individuals on which they have been computed. Therefore, the approach to statistical confidentiality is disclosure control rather disclosure avoidance.

Statitiscal offices release two kinds of data through their statistical databases: tabular data and microdata sets (individual respondent records). While there is a long experience in table dissemination, microdata dissemination is a much more recent activity. Following the nomenclature of [15], a *microdata set* is a set of records containing data of individual respondents, who can be persons, companies, etc. The individual records of a microdata set are stored in a *microdata file*. Each individual j is assigned a record or *data vector* V_j.

Microaggregation is a family of statistical disclosure control techniques for continuous microdata that falls into the substitution/perturbation category. This paper addresses the security of individual ranking microaggregation, a very commonly used specific type of microaggregation. Section 2 introduces basic microaggregation concepts. Section 3 analyzes the security of microaggregation when the individual ranking criterion is used. Section 4 is an analytical security study for the special case of continuous uniform data. Section 5 is a simulation study for the special case of normal data. Section 6 is a simulation study for skewed data, specifically for Weibull data. Section 7 contains the conclusions of the studies presented in the paper. Tables of results are given in the appendix.

2 Microagregation concepts

The rationale behind microaggregation is that confidentiality rules in use allow publication of microdata sets if the records correspond to groups of k or more individuals, where no individual dominates (*i. e.* contributes too much to) the group and k is a threshold value, typically between 3 and 5. Records are clustered into small aggregates or groups of size at least k. Rather than publishing an original variable X_i for a given record, the average of the values of X_i over the group to which the record belongs is published. To minimize information loss, groups should be as homogeneous as possible.

Classical microaggregation [6, 5, 1, 10] requires that all groups except perhaps one be of size k; if one group has to be of size $\geq k$, the best strat-

egy is that it be around the median of the data set. Allowing all groups to be of size $\geq k$ depending on the structure of data can be termed *data-oriented microaggregation* [11, 8, 14]. Figure 1 illustrates the advantages of variable-sized groups on a two-dimensional data set. If classical fixed-size microaggregation with $k = 3$ is used, we obtain a partition of the data into three groups, which looks rather unnatural for the data distribution given. On the other hand, if variable-sized groups are allowed then the five data on the left can be kept in a single group and the four data on the right in another group; such a variable-size grouping yields more homogeneous groups, which implies lower information loss.

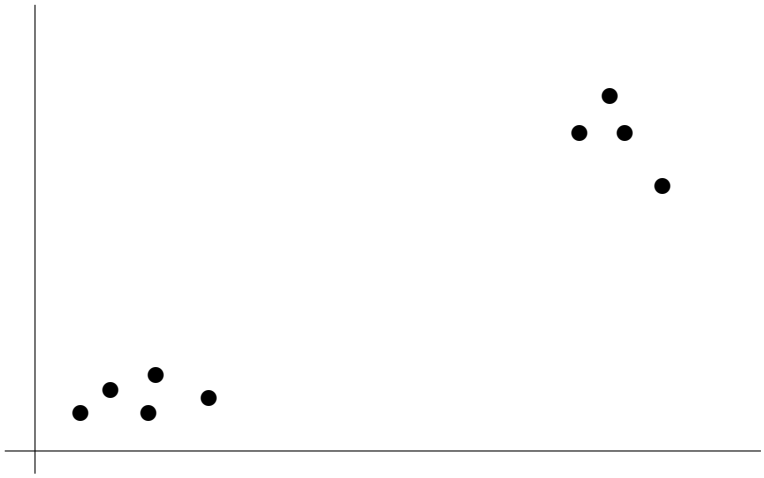

Fig. 1. Variable-sized groups versus fixed-sized groups

Without loss of generality, we assume in what follows that variables are one-dimensional; thus, "univariate" will be equivalent to "one-dimensional". Depending on whether they deal with one or with several variables at a time, microaggregation methods can be classified into univariate and multivariate:

– Univariate methods deal with multivariate data sets by microaggregating one variable at a time, *i.e.* variables are sequentially and independently microaggregated. This approach is known as individual ranking [6] and is the subject of this paper. While individual ranking causes low information loss, we will show that its disclosure risk is un-

acceptably high. Our analytical results here confirm empirical results obtained in [7] by using record linkage on two real data sets.
- Multivariate methods either map multivariate data to univariate data by projecting the former onto a single axis (*e.g.* using the first principal component, the sum of z-scores or even a particular variable[6, 5]) or directly deal with unprojected multivariate data [11, 8, 14]. Projected data can be viewed as a data set with a single variable, so that they allow univariate microaggregation to be used. When working on unprojected data, one can jointly microaggregate all variables in the data set at a time, or independently microaggregate groups of two variables at a time, three variables at a time, etc.

Exactly solving the microaggregation problem in the multivariate case without projection, *i.e.* finding a grouping where groups have maximal homogeneity and size at least k in a Euclidean space of dimension two or greater, has been shown to be NP-hard [13]. On the other hand, exactly solving the univariate microaggregation problem that appears in both the individual ranking and projected data approaches has recently been shown to be polynomially solvable as a shortest path problem [9].

Unfortunately, univariate microaggregation is not very attractive in spite of its low complexity, because it either suffers from high disclosure risk (this is what will be shown here for the case individual ranking) or from high information loss (caused by projection in the case of projected data). The empirical results reported in [7] suggest that multivariate microaggregation on unprojected data offers a much better tradeoff between information loss and disclosure risk, especially when groups or three or four variables are microaggregated at a time (rather than all at a time).

3 Security of microaggregation using individual ranking

For a microaggregation method to be called *secure*, it must not be easy to precisely estimate any value of the original data set from the microaggregated data set. Individual ranking is a popular approach to microaggregate a multivariate data set (*e.g.* see [10] and the Eurostat paper [3]). With individual ranking, each variable is considered independently. Data vectors are sorted by the first variable, then groups of k successive values of the first variable are formed and, inside each group, values are replaced by the group average. A similar procedure is repeated for the rest of variables.

Individual ranking does not partition the n data vectors in the microdata set on a data vector basis; instead, microaggregation is done for each

variable in turn so that a different partition is obtained for each variable in the microdata set.

Notation. In the rest of the paper we will use the term "element" instead of "data vector", to reflect the fact that individual ranking is a univariate procedure which deals with univariate elements (the values of one variable) rather than with the whole data vector at a time.

Individual ranking owes its popularity to its simplicity and to the fact that it usually preserves more information than one-dimensional projection (see [12, 7]). However, it makes it easier for an intruder to estimate the original data from microaggregated data. Indeed, with individual ranking any intruder knows that the real value of an element in the i-th group is between the average of the $i-1$-th group and the average of the $i+1$-th group; if these two averages are very close to each other, then a very narrow interval for the real value being searched has been determined.

Further, individual ranking is vulnerable to a less obvious attack based on the order statistics for the ranking variables. Let us assume that the data set contains at least one random variable X following a continuous distribution with cumulative distribution function $F(x)$ and density function $f(x)$. Let $X_1, \cdots, X_n$ be a random sample drawn from X. Let

$$X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$$

be the successive order statistics resulting from sorting the previous sample in ascending order. In other words, the event

$$x < X_{i:n} \leq x + \Delta x$$

for sufficiently small Δx is equivalent to saying that $i-1$ values in the sample are less than or equal to x, exactly one value lies within $(x, x+\Delta x]$, and the remaining $n-i$ values are greater than $x+\Delta x$. In general, for any Δx it is well known (*e.g.* see [2]) that

$$P(x < X_{i:n} \leq x + \Delta x)$$

$$= \frac{n!}{(i-1)!(n-i)!} \cdot [F(x)]^{i-1} \cdot [1-F(x+\Delta x)]^{n-i} \cdot [F(x+\Delta x)-F(x)] + O((\Delta x)^2) \tag{1}$$

where $O((\Delta x)^2)$ is the probability corresponding to the event "more than one sample value lies within $(x, x+\Delta x]$". Now the density function of the i-th order statistic $X_{i:n}$ can be obtained by differentiating equation (1):

$$f_{i:n}(x) = \lim_{\Delta x \to 0} \frac{P(x < X_{i:n} \leq x + \Delta x)}{\Delta x}$$

$$= \frac{n!}{(i-1)!(n-i)!} \cdot [F(x)]^{i-1} \cdot [1-F(x)]^{n-i} \cdot f(x) \tag{2}$$

where $-\infty < x < \infty$.

Thus, we can find two important properties about the i-th order statistic $X_{i:n}$:

- Its density function $f_{i:n}(x)$. This can be regarded as the prior probability density function of $X_{i:n}$ in the Bayesian sense [4].
- The interval where it lies. $X_{i:n}$ lies between the average a_{-1} of the group preceding its own group in the ranked data set and the average a_1 of the group following its own group in the ranked data set.

By combining the above information, we can derive the posterior probability density function of $X_{i:n}$ restricted to the interval $[a_{-1}, a_1]$, that is $f_{i:n}(x \mid a_{-1} \leq X_{i:n} \leq a_1)$. Now, given a confidence level α, computing a posterior probability interval

$$[x_{\alpha/2}, x_{1-\alpha/2}] \tag{3}$$

for $X_{i:n}$ from its posterior density is straightforward, since

$$\alpha/2 = \int_{a_{-1}}^{x_{\alpha/2}} f_{i:n}(y \mid a_{-1} \leq X_{i:n} \leq a_1)\, dy \tag{4}$$

$$1-\alpha/2 = \int_{a_{-1}}^{x_{1-\alpha/2}} f_{i:n}(y \mid a_{-1} \leq X_{i:n} \leq a_1)\, dy \tag{5}$$

To solve equation (4) for $x_{\alpha/2}$, the bisection method can be used, possibly combined with numerical integration. The same procedure can be used to solve equation (5) for $x_{1-\alpha/2}$. $[x_{\alpha/2}, x_{1-\alpha/2}]$ will always be narrower than the obvious interval $[a_{-1}, a_1]$, and it can be substantially narrower for i close to 1 or to n (that is, for extreme values in the original data set, see Sections 4,5 and 6).

The above attack can be further refined if one knows that $X_{i:n}$ is the smallest or largest data value in a microaggregated group. If $X_{i:n}$ is the smallest value in a group whose average is a_0, then we know that $X_{i:n}$ lies between the average a_{-1} of the group preceding its own group and the average a_0 of its own group. Likewise, if $X_{i:n}$ is the largest value in a group whose average is a_0, then $X_{i:n}$ must lie between a_0 and the average a_1 of the group following its own in the ranked data set. Note that $[a_{-1}, a_0] \subseteq [a_{-1}, a_1]$ and $[a_0, a_1] \subseteq [a_{-1}, a_1]$, so in both cases we get narrower intervals for $X_{i:n}$. Therefore,

– When $X_{i:n}$ is known to be the smallest value in a group, a posterior probability interval
$$[x'_{\alpha/2}, x'_{1-\alpha/2}] \tag{6}$$
for it can be derived from the posterior density function $f_{i:n}(x \mid a_{-1} \leq X_{i:n} \leq a_0)$ which is narrower than the posterior probability interval (3) resulting from equations (4) and (5).
– When $X_{i:n}$ is known to be the largest value in a group, a posterior probability interval
$$[x''_{\alpha/2}, x''_{1-\alpha/2}] \tag{7}$$
for it can be derived from the posterior density function $f_{i:n}(x \mid a_0 \leq X_{i:n} \leq a_1)$ and is narrower than the posterior probability interval (3) resulting from equations (4) and (5).

In practice, the attack described in this section could be mounted as follows:

1. For each individual variable, estimate the distribution of the original data based on the empirical distribution of the microaggregated data (only the latter data are assumed to be public). If individual ranking was used, then variables were microaggregated one at a time and the underlying single-variable distribution of original data is well reflected by microaggregated data.
2. For the estimated distribution, derive the expressions for $f_{i:n}(x|a_{-1} \leq X_{i:n} \leq a_1)$, $f_{i:n}(x|a_{-1} \leq X_{i:n} \leq a_0)$ and $f_{i:n}(x|a_0 \leq X_{i:n} \leq a_1)$ and the corresponding posterior probability intervals.

4 Analytical study for continuous uniform data

If X follows a continuous uniform distribution restricted to the interval $[0, 1]$ (denoted by $U[0, 1]$), the density function of $X_{i:n}$ is
$$f_{i:n}(x) = \frac{n!}{(i-1)!(n-i)!} \cdot x^{i-1} \cdot (1-x)^{n-i} \tag{8}$$
for $0 \leq x \leq 1$. Note that the density function of equation (8) is precisely the density function of the Beta distribution with parameters $(i, n-i+1)$.

Since we know the interval $[a_{-1}, a_1] \subset [0, 1]$ where $X_{i:n}$ must lie, we can derive the density function of $X_{i:n}$ restricted to that interval. For $x \in [a_{-1}, a_1]$, we have:
$$f_{i:n}(x \mid a_{-1} \leq X_{i:n} \leq a_1)$$

$$= \frac{\frac{n!}{(i-1)!(n-i)!} \cdot x^{i-1} \cdot (1-x)^{n-i}}{\int_{a_{-1}}^{a_1} \frac{n!}{(i-1)!(n-i)!} \cdot x^{i-1} \cdot (1-x)^{n-i}\, dx} = \frac{x^{i-1} \cdot (1-x)^{n-i}}{\int_{a_{-1}}^{a_1} x^{i-1} \cdot (1-x)^{n-i}\, dx} \quad (9)$$

For x not in $[a_{-1}, a_1]$, we have $f_{i:n}(x \mid a_{-1} \leq X_{i:n} \leq a_1) = 0$. Expressions analogous to (9) can be obtained when $X_{i:n}$ is known to be the smallest or the largest value of a group, by just replacing the interval $[a_{-1}, a_1]$ with $[a_{-1}, a_0]$ (for the smallest value) or $[a_0, a_1]$ (for the largest value).

The fact that $X_{i:n}$ follows a $Beta(i, n-i+1)$ leads to other results that are useful to analytically assess security when univariate microaggregation is performed on $U[0, 1]$ data. Specifically:

- The expected value for the i-th order statistic is

$$E(X_{i:n}) = E[Beta(i, n-i+1)] = \frac{i}{n+1}$$

- The j-th group is formed by $X_{kj-(k-1):n}$, $X_{kj-(k-2):n}$, $\cdots$, $X_{kj:n}$. Therefore the expected value of the average a_j of the j-th group is:

$$E(a_j) = \frac{\sum_{l=0}^{k-1} \frac{kj-l}{n+1}}{k} = \frac{\sum_{l=0}^{k-1} kj - l}{k(n+1)} = \frac{k^2 j - \frac{k(k-1)}{2}}{k(n+1)} = \frac{k(2j-1)+1}{2(n+1)}$$

- The expected difference between a_{j+1} and a_j (averages of the $j+1$-th and the j-th group) is:

$$E(a_{j+1} - a_j) = \frac{k(2(j+1)-1)+1}{2(n+1)} - \frac{k(2j-1)+1}{2(n+1)} = \frac{k}{n+1} \quad (10)$$

So the difference (10) does not depend on the particular value j being considered. Therefore, if a_{-1}, a_0 and a_1 are the averages of any three consecutive groups, we have:

$$E(a_1 - a_{-1}) = \frac{2k}{n+1}$$
$$E(a_1 - a_0) = E(a_0 - a_{-1}) = \frac{k}{n+1}$$

With the above results, the intervals (3), (6) and (7) can be analytically determined. It is thereafter straightforward to compute how much those intervals reduce on $[a_{-1}, a_1]$, $[a_{-1}, a_0]$ and $[a_0, a_1]$, respectively, where a_{-1} is the average of the group preceding the one being considered, a_0 is the average of the group being considered and a_1 is

the average of the group following the one being considered. To compare interval (3) with $[a_{-1}, a_1]$, we compute:

$$q = 100\frac{x_{\alpha/2} - x_{1-\alpha/2}}{a_1 - a_{-1}} \tag{11}$$

The value q has been computed for several values of α, k, n and for several order statistics $X_{i:n}$. The values q given in Table 1 (see Appendix) correspond to groups; for each group, the average value over the k order statistics (elements) forming the group has been taken. Taking groupwise q-values is sensible, because the intruder does not know the order of data vectors within a group (a microaggregated data file only contains the average value of each group). Table 1 contains q-values for the group $G_{(1)}$ with smallest values, and for groups ranked at percentiles 5, 15, 30 and 50 over the total number of groups (those groups are denoted by $G_{5\%}$, $G_{15\%}$, $G_{30\%}$ i $G_{50\%}$, respectively). Note that the uniform distribution being symmetrical, results for $G_{P\%}$ are analogous to results for $G_{100-P\%}$.

If $X_{i:n}$ is known to be the smallest element of its group, then we can compute the narrower posterior probability interval given by expression (6). This interval is to be compared with $[a_{-1}, a_0]$; the improvement achieved can be measured as

$$q' = 100\frac{x'_{\alpha/2} - x'_{1-\alpha/2}}{a_0 - a_{-1}} \tag{12}$$

Table 2 looks similar to Table 1, but it displays values q' corresponding to the intervals for the smallest element of each group (unlike the q in Table 1, which was an average computed over all elements in the group).

If $X_{i:n}$ is known to be the largest element of its group, then we can compute the posterior probability interval given by expression (7). This interval is to be compared with $[a_0, a_1]$; the improvement achieved can be measured as

$$q'' = 100\frac{x''_{\alpha/2} - x''_{1-\alpha/2}}{a_1 - a_0} \tag{13}$$

Table 3 looks similar to Table 1, but it displays values q'' corresponding to the intervals for the largest element of each group (unlike the q in Table 1, which was an average computed over all elements in the group.) Groups exhibiting q''-values similar to the corresponding q'-values of Table 2 have been omitted for brevity.

5 Simulation study for normal data

If X follows a $N(0,1)$ distribution, the density function of $X_{i:n}$ is

$$f_{i:n}(x) = \frac{n!}{(i-1)!(n-i)!} \cdot \left(\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\, dz \right)^{i-1} \cdot \left(1 - \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\, dz \right)^{n-i} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{14}$$

for $-\infty < x < +\infty$.

Again, we know the interval $[a_{-1}, a_1]$ where $X_{i:n}$ must lie, and we can derive the density function of $X_{i:n}$ restricted to that interval. For $x \in [a_{-1}, a_1]$, we have:

$$f_{i:n}(x \mid a_{-1} \leq X_{i:n} \leq a_1) = \frac{\left(\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\, dz \right)^{i-1} \cdot \left(1 - \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\, dz \right)^{n-i} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}}{\int_{a_{-1}}^{a_1} \left(\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\, dz \right)^{i-1} \cdot \left(1 - \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\, dz \right)^{n-i} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx} \tag{15}$$

For x not in $[a_{-1}, a_1]$, we have $f_{i:n}(x \mid a_{-1} \leq X_{i:n} \leq a_1) = 0$. Expressions analogous to (15) can be obtained when $X_{i:n}$ is known to be the smallest or the largest value of a group, by just replacing the interval $[a_{-1}, a_1]$ with $[a_{-1}, a_0]$ (for the smallest value) or $[a_0, a_1]$ (for the largest value).

To compute expressions analogous to (11), (12) and (13) for the normal case, we need to estimate the intervals $[a_{-1}, a_1]$, $[a_{-1}, a_0]$ and $[a_0, a_1]$. Unlike for the uniform case, there is no easy way to estimate those intervals without simulation, and their widths depend on the ranking of elements. Yet, the symmetry of the normal distribution implies some symmetries for interval widths:

1. $a_1 - a_{-1}$ is the same for $X_{i:n}$ and $X_{n-i+1:n}$
2. $a_0 - a_{-1}$ for $X_{i:n}$ is the same as $a_1 - a_0$ for $X_{n-i+1:n}$
3. $a_1 - a_0$ for $X_{i:n}$ is the same as $a_0 - a_{-1}$ for $X_{n-i+1:n}$

Algorithm 1 (Simulation procedure). *Consider several values of n and k (specifically $k = 3$, $k = 4$ and $k = 5$). For each combination of n and k, 50 $N(0,1)$ data sets have been generated; for each data set, confidence levels $\alpha = 0.1, 0.05, 0.01$ have been considered. In this way, given a combination of n, k and α, 50 interval widths have been obtained for each order statistic $X_{i:n}$. These 50 widths have been averaged.*

Table 4 shows how the width of interval $[a_{-1}, a_1]$ varies for several values of n and k and for several groups. The groups considered are the same of Section 4, namely $G_{(1)}$, $G_{5\%}$, $G_{15\%}$, $G_{30\%}$ and $G_{50\%}$. Note that the normal distribution being symmetrical, results for $G_{P\%}$ are analogous to results for $G_{100-P\%}$.

The following can be observed from Table 4:

- The interval width decreases as the number of elements n increases.
- The interval width increases as the size k of groups increases.
- Intervals for extreme groups (with very small or very large elements) are wider than for centered groups.

Once the intervals $[a_{-1}, a_1]$, $[a_{-1}, a_0]$ and $[a_0, a_1]$ have been estimated via simulation, it is easy to compute the posterior probability intervals (3), (6) and (7). The next step is to compare how much do the latter posterior probability intervals reduce on the former intervals. To compare interval (3) with $[a_{-1}, a_1]$, we compute for the normal case q as defined in expression (11). Table 5 is the version of Table 1 for normal data.

If $X_{i:n}$ is known to be the smallest element of its group, we can compute the interval (6), which is to be compared with $[a_{-1}, a_0]$; the improvement achieved can be measured by computing for normal data the coefficient q' defined by expression (12). Table 6 is the version of Table 2 for normal data.

If $X_{i:n}$ is known to be the largest element of its group, we can compute the interval (7), which is to be compared with $[a_0, a_1]$; the improvement achieved can be measured by computing for normal data the coefficient q'' defined by expression (13).

Table 7 is the version of Table 3 for normal data.

6 Simulation study for skewed (Weibull) data

For completeness, we include a third study on skewed data, specifically Weibull data. Continuous data, and most especially the financial data of businesses, are often quite skewed. If X follows a Weibull distribution with parameters $\alpha > 0$ and $\beta > 0$, the density function of $X_{i:n}$ is

$$f_{i:n}(x) = \frac{n!}{(i-1)!(n-i)!} \cdot (1-e^{-(\frac{x}{\alpha})^\beta})^{i-1} \cdot (e^{-(\frac{x}{\alpha})^\beta})^{n-i} \cdot \frac{\beta x^{\beta-1}}{\alpha^\beta} e^{-(\frac{x}{\alpha})^\beta} \quad (16)$$

for $x > 0$.

The density function of $X_{i:n}$ restricted to the interval $[a_{-1}, a_1]$ is

$$f_{i:n}(x|a_{-1} \leq X_{i:n} \leq a_1)$$

$$= \frac{\frac{n!}{(i-1)!(n-i)!} \cdot (1 - e^{-(\frac{x}{\alpha})^\beta})^{i-1} \cdot (e^{-(\frac{x}{\alpha})^\beta})^{n-i} \cdot \frac{\beta x^{\beta-1}}{\alpha^\beta} e^{-(\frac{x}{\alpha})^\beta}}{\int_{a_{-1}}^{a_1} \frac{n!}{(i-1)!(n-i)!} \cdot (1 - e^{-(\frac{x}{\alpha})^\beta})^{i-1} \cdot (e^{-(\frac{x}{\alpha})^\beta})^{n-i} \cdot \frac{\beta x^{\beta-1}}{\alpha^\beta} e^{-(\frac{x}{\alpha})^\beta} dx} \quad (17)$$

For x not in $[a_{-1}, a_1]$, one has $f_{i:n}(x|a_{-1} \leq X_{i:n} \leq a_1) = 0$. Expressions analogous to (17) can be obtained when $X_{i:n}$ is known to be the smallest or the largest value of a group.

To compute expressions analogous to (11), (12) and (13) for the Weibull case, we need to estimate the intervals $[a_{-1}, a_1]$, $[a_{-1}, a_0]$ and $[a_0, a_1]$. Like for the normal case, simulation is required to that end. A simulation procedure similar to Algorithm 1 has been followed, but for Weibull with $\alpha = 1$ and $\beta = 1.5$.

Table 8 is the version of Table 4 for Weibull data. The groups considered are $G_{(1)}$, $G_{10\%}$, $G_{50\%}$, $G_{90\%}$ and $G_{100\%}$ (the last group). Note that there is no distribution symmetry that can be exploited here.

The following can be observed from Table 8:

- The interval width decreases as the number of elements n increases.
- The interval width increases as the size k of groups increases.
- Weibull skewness causes that intervals for groups with very large elements are widest.

Tables 9, 10 and 11 are the versions of Table 1, 2 and 3 for Weibull data.

7 Conclusions

From the above discussion and the tables in the appendix, we can see that the order statistic attack against individual ranking microaggregation presented in this paper works well for uniform, normal and skewed (Weibull) data. However, when comparing the tables corresponding to the three distributions, it can be seen that, for normal and Weibull data, the posterior probability intervals obtained reduce more on the trivial intervals ($[a_{-1}, a_1]$,$[a_{-1}, a_0]$,$[a_0, a_1]$) than for uniform data. Therefore, the attack is more effective when original data are normal or Weibull, which on the other hand, are distributions which suit better typical continuous financial data. An additional unsafety factor for normal and Weibull data is that the trivial intervals for centered order statistics are already pretty narrow *per se*.

We next hightlight some features common to results in Sections 4, 5 and 6. Being common to uniform, normal data and Weibull data, the following remarks are likely to hold for other distributions of original data:

- The posterior probability intervals obtained for extreme order statistics are especially narrower than the corresponding trivial intervals. *This implies that the attack is most effective to estimate extreme values, which is especially worrying in the context of statistical confidentiality where microaggregation with individual ranking is usually applied*: a precise estimate of an extreme score makes it very easy to identify the individual behind that score. As an example, suppose that microaggregation using individual ranking has been applied to protect a microdata set including the variable "Age": if our attack yields a posterior probability interval $[98, 101]$ for the age value in a particular data vector, then it is easy to identify the person behind that data vector.
- For order statistics with central positions (those ranking between percentiles 15 and 85), the posterior probability intervals obtained do not improve much on trivial intervals. More specifically, the width of the posterior probability intervals for centered order statistics approaches the width of the corresponding trivial intervals times $1-\alpha$. For example, looking at Tables 1, 5, 9 for $n = 1000$, $\alpha = 0.1$ and $G_{50\%}$, we get relative widths 89.9 for uniform data, 89.5 for normal data and 89.6 for Weibull data; these values are very similar to 90.0 (relative width 100 of the trivial interval times 1-0.1).
- For fixed n, as the group size k grows from 3 to 5, the relative effectiveness (how much the width of posterior probability intervals reduces on average on the width of trivial intervals) of the attack increases. Of course, as k grows, trivial intervals tend to become larger, so the absolute effectiveness (inversely proportional to the width of posterior probability intervals) does not necessarily increase.
- As one would expect, the effectiveness of the attack decreases as α decreases. In other words, as confidence requirements grow higher, the resulting probability intervals grow wider and thus improve less on trivial intervals.

For the reasons above, it would seem reasonable to choose approaches to microaggregation other than individual ranking. For small k (between 3 and 5, as currently used in official statistics), we have justified that microaggregating one variable at a time is too transparent and does not offer enough security. For larger k, only empirical results based on record

linkage experiments are available[7]; such results show no traces of improvement as k increases.

Microaggregation on a record basis (considering all variables simultaneously by using one-dimensional projection or by dealing with unprojected data) offers a lower disclosure risk, even if it leads to higher information loss than individual ranking. In fact, empirical results in [7] indicate that projecting data results in too high an information loss: the best tradeoff between disclosure risk and information loss is reached for some variants of multivariate microaggregation.

References

1. N. Anwar, *Micro-Aggregation - The Small Aggregates Method*, Research Report. Luxembourg: Eurostat, 1993.
2. B. C. Arnold, N. Balakrishnan and H. N. Nagaraja, *A First Course in Order Statistics*. New York: Wiley, 1993.
3. Y. Baeyens and D. Defays, "Estimation of variance loss following microaggregation by the individual ranking method", in *Proceedings of Statistical Data Protection'98*. Luxembourg: Office for Official Publications of the European Communities, pp. 101-108, 1999.
4. G. E. P. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis*. New York: Wiley, 1992.
5. D. Defays and N. Anwar, "Micro-aggregation: a generic method", in *Proceedings of the 2nd International Symposium on Statistical Confidentiality*. Luxembourg: Office for Official Publications of the European Communities, pp. 69-78, 1995.
6. D. Defays and P. Nanopoulos, "Panels of enterprises and confidentiality: the small aggregates method", in *Proceedings of the 92 Symposium on Design and Analysis of Longitudinal Surveys*. Ottawa: Statistics Canada, pp. 195-204, 1993.
7. J. Domingo-Ferrer and V. Torra, "A quantitative comparison of disclosure control methods for microdata", in *Confidentiality, Disclosure and Data Access*, eds. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz. Amsterdam: North-Holland, pp. 111-133, 2001.
8. J. Domingo-Ferrer and J. M. Mateo-Sanz, "Practical data-oriented microaggregation for statistical disclosure control", *IEEE Transactions on Knowledge and Data Engineering,* vol. 14, pp. 189-201, 2002.
9. S. L. Hansen and S. Mukherjee, "A polynomial algorithm for optimal microaggregation", manuscript, 2002.
10. A. Hundepool, L. Willenborg, A. Wessels, L. van Gemerden, S. Tiourine and C. Hurkens, *μ-Argus 3.0 User's Manual*. Voorburg: Statistics Netherlands, 1998.
11. J. M. Mateo-Sanz and J. Domingo-Ferrer, "A method for data-oriented multivariate microaggregation", in *Proceedings of Statistical Data Protection'98*. Luxembourg: Office for Official Publications of the European Communities, pp. 89-99, 1999.
12. J. M. Mateo-Sanz and J. Domingo-Ferrer, "A comparative study of microaggregation methods", *Qüestiió*, vol. 22, no. 3, pp. 511-526, 1998.
13. A. Oganian and J. Domingo-Ferrer, "On the complexity of optimal microaggregation for statistical disclosure control", *Statistical Journal of the United Nations Economic Commission for Europe*, vol. 18, no. 4, 2001.

14. G. Sande, "Methods for data-directed microaggregation in one or more dimensions", in *Federal Committee on Statistical Methodology Research Conference*, Arlington VA, Nov. 14-16, 2001.
`http://www.fcsm.gov/01_papers/index.html`
15. L. Willenborg and T. de Waal, *Elements of Statistical Disclosure Control*. New York: Springer-Verlag, 2001.

Appendix

Table 1. Average percent relative width of interval (3) over $[a_{-1}, a_1]$ (uniform data)

		$k=3$				$k=4$				$k=5$			
		n				n				n			
		200	400	700	1000	200	400	700	1000	200	400	700	1000
$\alpha = 0.1$	$G_{(1)}$	81.2	81.2	81.3	81.3	76.9	77.1	77.2	77.2	72.7	73.0	73.2	73.2
	$G_{5\%}$	86.8	88.4	89.2	89.4	83.0	86.6	88.4	88.9	80.5	83.9	87.2	88.2
	$G_{15\%}$	89.0	89.5	89.7	89.8	87.9	89.0	89.5	89.6	86.1	88.3	89.1	89.4
	$G_{30\%}$	89.4	89.7	89.8	89.9	88.8	89.4	89.7	89.8	88.0	89.1	89.5	89.6
	$G_{50\%}$	89.5	89.7	89.9	89.9	89.0	89.5	89.7	89.8	88.4	89.2	89.6	89.7
$\alpha = 0.05$	$G_{(1)}$	89.5	89.5	89.6	89.6	86.1	86.2	86.3	86.3	82.4	82.7	82.8	82.8
	$G_{5\%}$	93.2	94.1	94.5	94.7	90.8	93.1	94.1	94.4	89.0	91.4	93.4	94.0
	$G_{15\%}$	94.4	94.7	94.8	94.9	93.8	94.5	94.7	94.8	92.8	94.1	94.5	94.7
	$G_{30\%}$	94.7	94.8	94.9	94.9	94.3	94.7	94.8	94.9	93.9	94.5	94.7	94.8
	$G_{50\%}$	94.7	94.9	94.9	94.9	94.5	94.7	94.9	94.9	94.1	94.6	94.8	94.8
$\alpha = 0.01$	$G_{(1)}$	97.6	97.6	97.6	97.6	96.3	96.4	96.4	96.5	94.5	94.6	94.7	94.7
	$G_{5\%}$	98.6	98.8	98.9	98.9	98.0	98.6	98.8	98.9	97.5	98.2	98.6	98.8
	$G_{15\%}$	98.9	98.9	99.0	99.0	98.7	98.9	98.9	99.0	98.5	98.8	98.9	98.9
	$G_{30\%}$	98.9	99.0	99.0	99.0	98.9	98.9	99.0	99.0	98.8	98.9	98.9	99.0
	$G_{50\%}$	98.9	99.0	99.0	99.0	98.9	98.9	99.0	99.0	98.8	98.9	99.0	99.0

Table 2. Percent relative width of interval (6) over $[a_{-1}, a_0]$ (uniform data)

		$k=3$				$k=4$				$k=5$			
		n				n				n			
		200	400	700	1000	200	400	700	1000	200	400	700	1000
$\alpha = 0.1$	$G_{(1)}$	88.0	88.0	88.0	88.0	87.0	87.1	87.1	87.1	86.0	86.1	86.1	86.1
	$G_{5\%}$	89.3	89.6	89.8	89.9	88.5	89.3	89.6	89.8	88.1	88.7	89.4	89.6
	$G_{15\%}$	89.8	89.9	89.9	89.9	89.5	89.8	89.9	89.9	89.2	89.7	89.8	89.9
	$G_{30\%}$	89.8	89.9	90.0	90.0	89.7	89.9	89.9	89.9	89.6	89.8	89.9	89.9
	$G_{50\%}$	89.9	89.9	90.0	90.0	89.8	89.9	89.9	90.0	89.6	89.8	89.9	89.9
$\alpha = 0.05$	$G_{(1)}$	93.9	93.9	93.9	93.9	93.3	93.3	93.3	93.3	92.7	92.7	92.8	92.8
	$G_{5\%}$	94.6	94.8	94.9	94.9	94.2	94.6	94.8	94.9	93.9	94.3	94.7	94.8
	$G_{15\%}$	94.9	94.9	95.0	95.0	94.7	94.9	94.9	95.0	94.6	94.8	94.9	94.9
	$G_{30\%}$	94.9	95.0	95.0	95.0	94.8	94.9	95.0	95.0	94.8	94.9	94.9	95.0
	$G_{50\%}$	94.9	95.0	95.0	95.0	94.9	94.9	95.0	95.0	94.8	94.9	94.9	95.0
$\alpha = 0.01$	$G_{(1)}$	98.8	98.8	98.8	98.8	98.6	98.6	98.6	98.6	98.5	98.5	98.5	98.5
	$G_{5\%}$	98.9	99.0	99.0	99.0	98.8	98.9	99.0	99.0	98.8	98.9	98.9	99.0
	$G_{15\%}$	99.0	99.0	99.0	99.0	98.9	99.0	99.0	99.0	98.9	99.0	99.0	99.0
	$G_{30\%}$	99.0	99.0	99.0	99.0	99.0	99.0	99.0	99.0	98.9	99.0	99.0	99.0
	$G_{50\%}$	99.0	99.0	99.0	99.0	99.0	99.0	99.0	99.0	99.0	99.0	99.0	99.0

Table 3. Percent relative width of interval (7) over $[a_0, a_1]$ (uniform data)

		$k=3$				$k=4$				$k=5$			
		n				n				n			
		200	400	700	1000	200	400	700	1000	200	400	700	1000
$\alpha = 0.1$	$G_{(1)}$	88.5	88.5	88.5	88.5	88.1	88.1	88.1	88.1	87.7	87.7	87.7	87.7
	$G_{5\%}$	89.3	89.6	89.8	89.9	88.8	89.3	89.6	89.8	88.5	88.9	89.4	89.6
$\alpha = 0.05$	$G_{(1)}$	94.2	94.2	94.2	94.2	94.0	94.0	94.0	94.0	93.7	93.7	93.7	93.7
	$G_{5\%}$	94.6	94.8	94.9	94.9	94.3	94.6	94.8	94.9	94.2	94.4	94.7	94.8
$\alpha = 0.01$	$G_{(1)}$	98.8	98.8	98.8	98.8	98.8	98.8	98.8	98.8	98.7	98.7	98.7	98.7
	$G_{5\%}$	98.9	99.0	99.0	99.0	98.9	98.9	99.0	99.0	98.8	98.9	98.9	99.0

Table 4. Width of $[a_{-1}, a_1]$ for several groups and values n, k (normal data)

	$k=3$				$k=4$				$k=5$			
	n				n				n			
	200	400	700	1000	200	400	700	1000	200	400	700	1000
$G_{(1)}$	.67	.60	.56	.53	.71	.60	.58	.58	.76	.70	.61	.59
$G_{5\%}$	.26	.26	.22	.22	.41	.34	.33	.31	.43	.38	.34	.32
$G_{15\%}$	.12	.11	.09	.09	.18	.15	.13	.12	.23	.18	.18	.16
$G_{30\%}$	.08	.06	.05	.05	.12	.08	.07	.07	.15	.11	.09	.08
$G_{50\%}$	.07	.05	.04	.03	.10	.06	.05	.05	.12	.08	.06	.06

Table 5. Average percent relative width of interval (3) over $[a_{-1}, a_1]$ (normal data)

		$k=3$				$k=4$				$k=5$			
		n				n				n			
		200	400	700	1000	200	400	700	1000	200	400	700	1000
$\alpha = 0.1$	$G_{(1)}$	67.8	66.1	68.2	65.3	62.3	61.8	61.9	60.2	56.4	56.8	58.0	56.7
	$G_{5\%}$	81.1	83.9	85.7	87.6	75.4	80.3	83.9	87.2	71.0	77.7	82.2	84.0
	$G_{15\%}$	84.4	86.9	88.5	89.3	83.1	87.2	87.5	88.8	78.2	84.5	87.3	88.0
	$G_{30\%}$	86.6	88.5	89.2	89.5	85.0	87.9	88.6	88.9	83.3	86.5	88.8	88.8
	$G_{50\%}$	88.2	89.0	89.4	89.5	85.8	88.4	88.8	89.1	84.1	86.9	88.5	88.5
$\alpha = 0.05$	$G_{(1)}$	77.2	75.0	77.1	74.2	71.4	70.7	70.8	69.0	65.1	65.6	66.7	65.4
	$G_{5\%}$	88.8	91.0	92.3	93.6	84.3	88.3	90.8	93.4	80.2	86.4	89.8	91.2
	$G_{15\%}$	91.4	92.9	94.2	94.6	90.1	93.4	93.5	94.3	86.4	91.3	93.4	93.9
	$G_{30\%}$	92.9	94.2	94.5	94.7	91.8	93.8	94.2	94.4	90.8	92.9	94.3	94.3
	$G_{50\%}$	94.0	94.4	94.7	94.7	92.3	94.1	94.3	94.5	91.3	93.2	94.2	94.1
$\alpha = 0.01$	$G_{(1)}$	90.3	87.7	89.3	86.8	85.0	84.0	84.1	82.4	78.9	79.6	80.4	79.2
	$G_{5\%}$	96.7	97.8	98.3	98.7	94.8	96.9	97.4	98.6	92.4	96.2	97.5	98.0
	$G_{15\%}$	98.0	98.3	98.8	98.9	96.9	98.6	98.7	98.9	95.6	97.8	98.6	98.7
	$G_{30\%}$	98.5	98.8	98.9	98.9	98.1	98.7	98.8	98.9	97.9	98.5	98.9	98.9
	$G_{50\%}$	98.8	98.9	98.9	98.9	98.3	98.8	98.9	98.9	98.0	98.6	98.8	98.8

Table 6. Percent relative width of interval (6) over $[a_{-1}, a_0]$ (normal data)

		$k=3$				$k=4$				$k=5$			
		n				n				n			
		200	400	700	1000	200	400	700	1000	200	400	700	1000
$\alpha = 0.1$	$G_{(1)}$	78.5	75.1	78.3	75.7	74.0	73.8	73.7	72.4	69.3	70.1	70.9	70.3
	$G_{5\%}$	86.5	88.7	88.0	89.2	85.2	87.7	87.4	89.3	83.0	86.7	88.0	88.3
	$G_{15\%}$	87.5	89.2	89.6	89.7	87.4	89.2	89.2	89.6	86.4	88.1	89.2	89.4
	$G_{30\%}$	89.0	89.7	89.7	89.9	88.6	89.5	89.5	89.6	88.0	89.0	89.5	89.6
	$G_{50\%}$	89.4	89.7	89.7	89.8	88.3	89.6	89.7	89.7	88.5	89.0	89.6	89.6
$\alpha = 0.05$	$G_{(1)}$	87.0	83.6	86.1	83.8	82.6	82.4	82.0	81.4	78.3	79.0	79.7	79.2
	$G_{5\%}$	92.8	94.3	93.8	94.5	92.1	93.7	93.3	94.6	90.5	93.1	93.8	94.0
	$G_{15\%}$	93.5	94.5	94.8	94.9	93.3	94.5	94.5	94.8	92.8	93.8	94.6	94.7
	$G_{30\%}$	94.4	94.8	94.9	94.9	94.2	94.7	94.7	94.8	93.9	94.4	94.8	94.8
	$G_{50\%}$	94.7	94.8	94.9	94.9	94.0	94.8	94.8	94.8	94.1	94.5	94.8	94.8
$\alpha = 0.01$	$G_{(1)}$	96.2	94.2	95.0	93.1	93.5	93.2	92.1	93.1	90.9	90.9	91.3	90.9
	$G_{5\%}$	98.5	98.8	98.7	98.9	98.3	98.7	98.6	98.9	97.8	98.5	98.7	98.8
	$G_{15\%}$	98.6	98.9	98.9	99.0	98.5	98.9	98.9	99.0	98.5	98.7	98.9	98.9
	$G_{30\%}$	98.9	99.0	99.0	99.0	98.8	98.9	98.9	99.0	98.7	98.9	98.9	99.0
	$G_{50\%}$	98.9	99.0	99.0	99.0	98.7	98.9	99.0	99.0	98.8	98.9	99.0	98.9

Table 7. Percent relative width of interval (7) over $[a_0, a_1]$ (normal data)

		$k=3$				$k=4$				$k=5$			
		n				n				n			
		200	400	700	1000	200	400	700	1000	200	400	700	1000
$\alpha = 0.1$	$G_{(1)}$	81.1	80.6	81.3	79.9	79.0	78.2	77.7	78.2	74.6	74.6	74.9	74.8
	$G_{5\%}$	86.3	88.9	88.3	89.3	85.9	87.7	87.7	89.4	84.4	86.8	88.0	88.4
$\alpha = 0.05$	$G_{(1)}$	89.0	88.6	88.8	87.5	87.2	86.6	85.7	86.8	83.6	83.3	83.6	83.5
	$G_{5\%}$	92.7	94.4	94.0	94.6	92.5	93.6	93.5	94.7	91.6	93.2	93.8	94.1
$\alpha = 0.01$	$G_{(1)}$	97.2	97.1	96.8	95.7	96.3	96.0	94.6	96.3	94.7	93.9	94.1	94.0
	$G_{5\%}$	98.4	98.9	98.8	98.9	98.4	98.7	98.6	98.9	98.1	98.6	98.7	98.8

Table 8. Width of $[a_{-1}, a_1]$ for several groups and values n, k (Weibull data)

	$k=3$				$k=4$				$k=5$			
	n				n				n			
	200	400	700	1000	200	400	700	1000	200	400	700	1000
$G_{(1)}$	.07	.05	.03	.02	.10	.06	.04	.03	.11	.07	.05	.03
$G_{10\%}$	.05	.02	.01	.01	.06	.03	.02	.01	.08	.04	.02	.02
$G_{50\%}$	.04	.02	.01	.01	.06	.03	.02	.01	.08	.04	.02	.01
$G_{90\%}$	.13	.07	.04	.03	.19	.09	.06	.04	.23	.12	.08	.05
$G_{100\%}$	.68	.66	.69	.59	.74	.68	.67	.66	.77	.75	.69	.66

Table 9. Average percent relative width of interval (3) over $[a_{-1}, a_1]$ (Weibull data)

		$k=3$				$k=4$				$k=5$			
		n				n				n			
		200	400	700	1000	200	400	700	1000	200	400	700	1000
$\alpha = 0.1$	$G_{(1)}$	73.2	70.8	73.5	75.8	68.5	69.1	69.2	70.1	65.8	65.8	64.4	65.0
	$G_{10\%}$	83.7	87.3	88.1	88.2	80.5	85.3	87.9	87.5	77.2	82.8	85.2	86.6
	$G_{50\%}$	87.3	88.5	88.9	89.6	85.0	87.8	89.2	89.2	84.8	86.2	88.4	89.0
	$G_{90\%}$	83.4	86.9	87.9	88.6	78.3	85.6	87.8	87.8	75.8	83.2	84.3	87.3
	$G_{100\%}$	65.0	63.0	61.6	67.1	60.4	61.7	59.8	61.0	55.7	56.0	54.9	55.8
$\alpha = 0.05$	$G_{(1)}$	82.2	79.8	82.4	84.3	74.8	78.2	78.1	79.1	75.1	75.0	73.7	74.3
	$G_{10\%}$	90.7	93.4	93.9	94.0	88.4	92.0	93.8	93.5	85.7	90.4	91.8	92.9
	$G_{50\%}$	93.4	94.2	94.4	94.8	91.9	93.7	94.5	94.6	91.7	92.6	94.1	94.5
	$G_{90\%}$	90.5	93.1	93.8	94.2	86.6	92.2	93.7	93.7	84.4	90.5	91.3	93.4
	$G_{100\%}$	74.3	71.8	70.6	76.3	69.8	71.0	68.6	70.2	64.5	64.9	63.8	64.3
$\alpha = 0.01$	$G_{(1)}$	93.6	91.6	93.4	94.5	87.8	90.6	90.2	91.2	88.4	88.2	87.3	87.7
	$G_{10\%}$	97.7	98.6	98.7	98.8	96.8	98.2	98.7	98.6	95.6	97.7	98.0	98.5
	$G_{50\%}$	98.6	98.8	98.9	98.9	98.2	98.7	98.9	98.9	98.1	98.4	98.8	98.9
	$G_{90\%}$	97.5	98.5	98.7	98.8	96.0	98.3	98.7	98.7	94.5	97.7	97.9	98.6
	$G_{100\%}$	87.7	84.5	84.0	89.3	84.2	85.2	82.1	84.2	78.7	79.2	78.6	78.0

Table 10. Percent relative width of interval (6) over $[a_{-1}, a_0]$ (Weibull data)

		$k=3$				$k=4$				$k=5$			
		n				n				n			
		200	400	700	1000	200	400	700	1000	200	400	700	1000
$\alpha = 0.1$	$G_{(1)}$	82.5	80.3	82.9	83.1	75.8	80.3	80.1	81.0	79.4	78.4	78.1	79.2
	$G_{10\%}$	86.8	89.2	89.5	89.2	86.5	88.7	89.4	89.2	86.0	88.1	88.4	88.9
	$G_{50\%}$	89.3	89.7	89.6	89.8	88.7	89.3	89.7	89.8	88.6	88.9	89.5	89.7
	$G_{90\%}$	88.2	89.1	89.4	89.6	87.3	89.0	89.3	89.3	85.2	88.0	88.5	89.4
	$G_{100\%}$	80.3	77.4	79.4	81.9	78.4	79.8	75.6	78.4	73.7	74.4	76.1	73.6
$\alpha = 0.05$	$G_{(1)}$	90.0	88.1	90.2	90.1	84.3	88.2	87.9	88.8	87.7	86.6	86.6	87.5
	$G_{10\%}$	93.0	94.6	94.7	94.7	92.9	94.3	94.7	94.6	92.6	93.9	94.0	94.4
	$G_{50\%}$	94.6	94.8	94.8	94.9	94.3	94.6	94.9	94.9	94.2	94.4	94.7	94.8
	$G_{90\%}$	94.0	94.5	94.7	94.8	93.4	94.4	94.6	94.6	91.8	93.8	94.1	94.6
	$G_{100\%}$	88.2	85.4	87.3	89.5	86.9	88.0	84.0	86.9	82.2	83.2	85.2	82.5
$\alpha = 0.01$	$G_{(1)}$	97.5	96.6	97.6	97.1	94.4	96.7	96.5	97.1	96.7	95.7	96.3	96.5
	$G_{10\%}$	98.5	98.9	98.9	98.9	98.4	98.8	98.9	98.9	98.4	98.7	98.8	98.9
	$G_{50\%}$	98.9	99.0	99.0	99.0	98.8	98.9	99.0	99.0	98.8	98.9	98.9	99.0
	$G_{90\%}$	98.8	98.9	98.9	99.0	98.6	98.9	98.9	98.9	98.0	98.7	98.8	98.9
	$G_{100\%}$	96.7	94.7	96.0	97.3	96.4	96.9	94.1	96.4	92.6	93.9	95.7	93.9

Table 11. Percent relative width of interval (7) over $[a_0, a_1]$ (Weibull data)

		$k=3$				$k=4$				$k=5$			
		n				n				n			
		200	400	700	1000	200	400	700	1000	200	400	700	1000
$\alpha = 0.1$	$G_{(1)}$	82.9	80.5	83.3	82.5	75.8	80.3	80.1	81.0	79.6	78.1	78.9	79.8
	$G_{10\%}$	88.6	89.2	89.5	89.5	87.1	88.6	89.5	89.3	86.4	88.0	88.7	89.2
	$G_{50\%}$	89.1	89.3	89.7	89.9	88.3	89.5	89.8	89.8	88.3	88.9	89.6	89.7
	$G_{90\%}$	87.4	88.8	89.3	89.6	85.2	88.5	89.5	89.3	85.0	87.8	88.3	89.2
	$G_{100\%}$	75.5	72.2	72.9	78.0	78.4	79.8	75.6	78.5	69.4	70.2	70.3	68.8
$\alpha = 0.05$	$G_{(1)}$	90.3	88.1	90.6	89.6	84.3	88.2	87.9	88.8	88.0	86.5	87.3	88.0
	$G_{10\%}$	94.2	94.6	94.7	94.7	93.2	94.2	94.7	94.6	92.8	93.9	94.2	94.6
	$G_{50\%}$	94.5	94.6	94.8	94.9	94.0	94.8	94.9	94.9	94.0	94.4	94.8	94.9
	$G_{90\%}$	93.4	94.3	94.6	94.8	91.9	94.1	94.7	94.6	91.8	93.7	94.0	94.6
	$G_{100\%}$	84.0	80.5	81.7	86.4	86.9	88.0	84.0	86.9	78.2	79.1	79.8	77.7
$\alpha = 0.01$	$G_{(1)}$	97.7	96.4	97.7	96.6	94.4	96.7	96.5	97.1	96.9	95.7	96.5	96.9
	$G_{10\%}$	98.8	98.9	98.9	98.9	98.5	98.8	99.0	98.9	98.4	98.7	98.8	98.9
	$G_{50\%}$	98.9	98.9	99.0	99.0	98.8	98.9	99.0	99.0	98.8	98.9	99.0	99.0
	$G_{90\%}$	98.6	98.9	98.9	99.0	98.2	98.8	98.9	98.9	98.1	98.7	98.8	98.9
	$G_{100\%}$	94.2	91.2	92.6	95.7	96.4	96.9	94.1	96.4	89.8	91.0	92.7	90.0