A Generalized Domain Size Threshold for Analysis Restrictions with Remote Analysis Servers

Avinash C. Singh, Joshua M. Borton, and Allan M. Crego

NORC at the University of Chicago, Chicago, IL 60603
singh-avi@norc.org, borton-joshua@norc.org, crego-allan@norc.org

Abstract

We consider cost and time saving alternatives to the practice of direct access to microdata under licensing agreements. The traditional alternative of public use files (PUFs) does not allow for a rich set of analytic variables (AVs) with high analytic utility due to confidentiality concerns. The main reason is the difficulty in creating disclosure-safe microdata at the unit level in view of the potential public availability of more and more personal information now or in future. Instead we consider transforming the problem of disclosure treatment of unit level microdata to that of the relatively simpler problem for aggregate level macrodata whose solution comprises disclosure treatment of output data for analysis domains via restrictions. We propose imposing analysis restrictions through a checklist of screened AVs which gives rise to a query-based public use file (Q-PUF) as it requires neither user authorization for indirect query-access to the raw data nor output disclosure-safety review even for complex analyses. The premise is that it is conceptually simpler to impose analysis restrictions in terms of AVs than analysis outputs as many different analyses can be conducted with the same set of AVs. Q-PUF provides a simplified option for remote analysis servers as it is free from any problems of repeated queries because of checklist restrictions on AVs. In particular, perturbation of estimates is not required because analysis domains via AVs cannot be arbitrarily specified by the user. For both aggregate level and unit level outputs, we propose, based on estimating functions, a criterion termed generalized domain size threshold for analysis restrictions (GDSTAR or GD*) for constructing a checklist of AVs for disclosure-safe output which is a generalization of the usual condition of minimum domain sample size for estimating domain counts. If the GD* criterion is not satisfied by an AV, it is either dropped or transformed through hierarchical collapsing of categories, if possible, to satisfy GD*. For AVs in the checklist, the GD* criterion allows for exact results for aggregate level output but only approximate results for unit level output although perhaps with little practical consequence. The GD* criterion is not a serious restriction because it is also needed for reliability or precision of estimates besides confidentiality. An illustrative example of GD* application based on the 2010 PUF for National Survey of Drug Use and Health (NSDUH) is also provided.

Key Words: GDSTAR; Generalized or g-domain sample size; Input and Output Disclosure Treatment; Query-based PUF; Remote Analysis Server

1. Introduction

There are two broad types of disclosure treatment: first, input treatment in that the unit-level microdata (or very low aggregate-level macrodata such as tabular) are treated for statistical disclosure limitation; second, output treatment in that analysis results (which are generally at high levels of aggregation in the interest of reliability) are treated. Public use files (PUFs) are based on input treatment while query-based systems for remote analysis servers are based on output treatment. Both input and output treatments comprise some forms of perturbation and/or suppression: the input treatment involves perturbation in the form of recoding, substitution, noise addition etc., and suppression via blanking-out sensitive fields or even the whole record; while the output treatment involves suppression such as imposition of analysis restrictions through minimum size of analysis domains, limited inclusion of model covariates, and restrictions on distributional or model diagnostics; and perturbation such as introducing uncertainty through subsampling in estimating totals to thwart differencing attacks under repeated queries, and use of synthetic residuals for model diagnostics; Gomatnam et al. (2005) and Lucero et al. (2011). For the traditional method of creating public use files (PUFs) using input disclosure treatment of microdata, it has become essentially impossible to build in high analytic utility in PUFs via inclusion of a rich set of analytic variables (AVs; e.g.,

categorical AVs with finer categories and their products or interactions) while controlling disclosure risk at a reasonable level; the main reason being the ever-increasing potential of public availability of personal information now or in future. The problem considered in this paper arose in the context of creating PUFs from CMS Medicare Claims data—a large administrative database, and because of this problem, the goal was modified to creating a disclosure-treated file but with controlled use by housing it in a secure data enclave environment; see Borton et al. (2011). Compared to the usual untreated controlled use file (such as research identifiable files and limited data sets) in the data enclave, it is expected that due to the initial disclosure treatment of microdata, significant cost and time can be saved by having much simplified processes of user authorization and output disclosure-safety review before exporting analysis results output from the data enclave.

Nevertheless, the traditional PUF dissemination model for microdata is appealing because the user does not require any authorization for direct access to the treated microdata, and is not subject to any restrictions on types of analyses except for low utility of analyses due to availability of limited AVs and lack of precision in available AVs. On the other side of the spectrum of data dissemination models are remote analysis servers where users have no direct access to microdata but only indirect access through analysis queries; see Keller-McNulty and Unger (1998) for an early formulation of the problem. This approach involves output disclosure treatment as mentioned above. Remote analysis servers are appealing in terms of resulting high analytic utility but the main problem is to protect against the potential of differencing attacks by repeated queries especially under complex analyses; this problem may lead to requirements for additional security in terms of user authorization (and possibly data use agreements--DUAs) before gaining access, and of output disclosure-safety review; Sparks et al. (2011).

The above restrictions with remote analysis servers are not likely to be appealing to users in general because of inherent cost and time burden compared to traditional PUFs. In practice, somewhat analogous to freely accessible PUFs, it would be desirable for remote analysis servers to neither require user authorization for query-access nor approval for exporting complex analysis results; i.e., they behave like PUFs except for direct access—we will term such an option as query-based PUF or Q-PUF. To this end, we propose imposing analysis restrictions through a checklist of screened AVs as it would provide a systematic and conceptually simpler way to impose analysis restrictions in terms of AVs than analysis outputs since many different analyses can be conducted with the same set of AVs. Being in the family of query-based access systems, Q-PUF is similar to other methods currently being developed for remote analysis servers except that it is free from any problems of repeated queries because of checklist restrictions on AVs. In particular, perturbation of estimates is not required because analysis domains cannot be arbitrarily specified by users.

Unlike the high level of aggregation (i.e., at analysis domain level) used for disclosure treatment in the proposed Q-PUF, a low level of aggregation (i.e., building blocks or micro-groups (MG) used as proxy for unit level) with input of MG counts and micro-means (i.e., means of AVs for each MG) was recently proposed by Singh and Borton (2012) where the input data of MG counts is disclosure-treated by perturbations via subsampling but without any suppression. This method, termed Aggregate Level PUF, complements Q-PUF in that it allows for direct access for preliminary analysis purposes while Q-PUF provides final analysis. It may be remarked that maintaining analytic utility of traditional PUFs is challenging because disclosure treatment is required at the unit level in view of the requirement that the treated data be disseminated at the unit level. Both Q-PUF and aggregate level PUF, however, exhibit departures from unit level microdata by transforming it to aggregate level macrodata for disclosure treatment. Incidentally, the basic idea of aggregate level PUF comes from commonly used aggregate level modeling for small domain estimation and is expected to yield higher analytic utility and data confidentiality in comparison to unit level PUFs in view of the limitations mentioned above. However, Q-PUF is expected to yield even higher utility for analyses for screened AVs in the checklist as they are performed on the original untreated microdata without any perturbations of output, and higher confidentiality because disclosure-treatment is analysis output-specific and not generic performed on the source input data at the unit level.

The focus of this paper is to provide the option of Q-PUF via a checklist of screened AVs. The screened AVs satisfy a criterion of minimum generalized domain size threshold for analysis restrictions (GDSTAR or GD*for short; see Section 2 for its definition) in that the number of cases in the sample belonging to the domain of interest for each of zero and nonzero values of the AV (which need not be categorical; if categorical then we have 0/1 values of categorical indicators) is above a threshold such as 50. The choice of the threshold as 50 is somewhat ad hoc but is guided by the commonly used rule of thumb for sample size of 30 adjusted by an approximate design effect of 1.75

for complex surveys resulting in the effective sample size of about 50. For protecting confidentiality, we could have chosen a much smaller value of c_0 such as 10, but the choice of a larger value of 50 is based on considerations of good finite sample properties of summary statistics toward the goal of protecting analytic utility. If it is not satisfied by an AV, a hierarchical collapsing (which may involve combining domains of zero and nonzero values of the AV; see Section 3) is performed to meet the threshold. The GD* criterion based on estimating functions (see Section 4 c and d) provides a unified approach to protecting confidentiality of output from different types of analysis, and is a generalization of the commonly used cell size restriction in output treatment. Moreover, it is not deemed to be a serious analytic limitation as it is also needed to ensure reliability of estimates. In other query-based systems currently under development through remote analysis servers, restrictions are placed directly on analysis specifications and their output instead of AVs but the user is free to specify any function of AVs available in the dataset for analysis; this is unlike Q-PUF where AVs are restricted to the checklist. Remote analysis servers form a very active research area with important contributions from the Census Bureau on Microdata Analysis System (MAS; Steel and Reznek 2005; Lucero et al. 2011), NCHS on ANalytical Data for Research by Email and Web (ANDREW; Gambhir and Harris, 2006), NCES on Data Access System (DAS; Seastrom and Kaufman, 2003, and Russell and White, 2011), and CSIRO (Australia) on Privacy Preserving Analytics (PPA; Sparks et al., 2008); see also Rowland (2003) for a good review of early developments.

The organization of this paper is as follows. Section 2 provides a heuristic motivation of the proposed criterion of GD* for output treatment under Q-PUF followed by its description in Section 3; the description is provided in terms of an illustrative example based on the 2010 PUF from the National Survey on Drug Use and Health (NSDUH) data. In Section 4, we consider issues of analytic utility and confidentiality of results from both descriptive and analytic inferences and in Section 5, comparison of GD* with Census Bureau's MAS method along with others is discussed. Summary and remarks are presented in Section 6.

2. Heuristic Motivation of the Proposed Criterion for Output Treatment under Q-PUF

The current approach to confidentiality protection of analysis results from remote analysis servers comprises restrictions on analysis specifications and resulting output needed for descriptive and analytic inferences about study variables. The study variable may be categorical or numeric including continuous or semi-continuous; i.e., product of continuous and categorical variables. There are two types of output: aggregate level and unit level. Under descriptive inference, aggregate level output deals with summary statistics of means, totals, and ratios (i.e., point estimates as well as variance and interval) for study variables for analysis domains defined by auxiliary variables; here values of estimates do not in general (except when the domain is rare) have useful information about realized or observed values of the study variable or functions of it. Besides aggregate level output, unit level output under descriptive inference deals with estimation of other distributional characteristics such as quantiles where estimates may have useful information about observed values of the study variable—a potential problem for disclosure. On the other hand, under analytic inference, aggregate level output deals with estimation of model parameters in the mean (function of covariates or auxiliary variables) and model diagnostics including measures of goodness-of-fit, while unit level output deals with prediction of the study or dependent variable for a given set of covariates, and additional model diagnostics such as residual plots, and detection of extreme observations.

For imposing analysis restrictions on aggregate level output under Q-PUF, we propose to use the strategy of creating a checklist of screened analytic variables (AVs) based on desired analyses such that they satisfy the criterion (termed GD*) mentioned in the introduction and to be defined below. Consider an analytic variable (AV) which may be elementary (such as AVs comprising the source microdata set) or composite (such as AVs obtained as functions of elementary AVs; e.g., product of categorical indicators of different AVs or product of continuous and categorical AVs). In any analysis output, AVs of interest are typically composite, to be denoted by u, which are products of study variables (y) and auxiliary variables (x); composite AVs clearly encompass elementary AVs. We now define certain terms such as generalized (or g-) domains corresponding to AVs of interest which are needed to deal with protecting confidentiality and reliability of domain totals when AVs are numeric or continuous study variables. These are also applicable to the case of analytic parameters because estimating functions for model parameters can be regarded as domain totals involving given values of unknown parameters and essentially composite AVs; see Section 4(c, d).

Domains, Generalized or G-Domains, and their Sample Sizes: Each domain is defined by a categorical AV, and signifies a subpopulation of individuals taking either values of 1 or 0 corresponding to the binary indicator of the categorical AV (elementary or composite), thus dividing the population into two parts corresponding to zero and nonzero (actually 1) values of the AV. As an example, binary indicators could be categories of gender, age, and income, or products of their indicators. The AV may be intrinsically categorical such as gender, or a categorized version of a numeric or continuous variable such as age or income. Next, each g-domain is defined by an AV which may or may not be categorical, and signifies a subpopulation of individuals taking either zero or nonzero values of the AV, e.g., products (u) of income (a continuous variable) and gender (categorical variable). Clearly, domain is a special case of g-domain when u is categorical. In defining g-domains, we are implicitly categorizing u as a binary indicator u^* with values of 1 for individuals in the category of nonzero values of u and values of 0 for individuals in the category of zero values of u, and then the definition of g-domain follows from the usual definition of domain. The domain sizes are simply the number of individuals with the corresponding categorical AV values of 1 or 0. The g-domain sizes are naturally the number of individuals with nonzero contributions of u (i.e., corresponding to $u^*=1$) and the number with zero contributions; i.e., corresponding to $u^*=0$.

DSTAR (D*) and GDSTAR (GD*) Criteria: These criteria are needed to screen AVs for the checklist. If the domain size meets a pre-specified threshold (denote by c_0 which can be set at 50—this choice is data-specific) for a categorical AV (which defines the domain), then we say that the AV satisfies the domain size threshold for analysis restrictions (D* for short) criterion. Similarly for a g-domain corresponding to a general AV (not necessarily categorical), if the g-domain size meets the threshold, we say the AV satisfies GD*. Clearly D* is a special case of GD* when the AV is categorical. Additionally, the above requirement of GD* compliance for any AV to belong to the checklist is not marginal, but joint with respect to existing AVs in the checklist in that all g-domains (corresponding either directly to AVs or derived indirectly as linear combination (e.g., complementary domains) meet the threshold. For example, a new elementary AV may satisfy GD* but when it is multiplied by an existing AV in the checklist, the resulting composite AV may not satisfy GD*. If an AV is not compliant, hierarchical collapsing (described in the next section) is performed in a systematic manner such that analysis results remain meaningful and useful in practice.

We remark that if a domain or g-domain count is zero either as a structural zero or random zero, the corresponding AV in order to be GD* compliant will undergo hierarchical collapsing so that it will manifest itself after being collapsed with some other domain related to the same AV or in combination with some other AV in the checklist. Thus, with Q-PUF, there is no disclosure problem for null query sets because the output would not reveal whether the collapsing was made due to null query set or the g-domain size being too small. It is also remarked that the GD* criterion stipulates much higher threshold than the usual value (5 or 10) for count data because unlike the case of raw input data of cell counts, analysis output domains are not of much interest if they are very small in the interest of reliability of estimates. For unit level output from screened AVs, additional analysis restrictions are imposed such that the output is modified to an aggregate level in terms of summary statistics satisfying the GD* criterion although with a smaller threshold c_0' (such as 10) because of concerns here only about confidentiality and not reliability. Thus for output treatment under Q-PUF, we propose analysis restrictions by restricting users to a checklist of AVs which are pre-screened by the GD* criterion along with further restrictions on the unit level output.

3. Q-PUF Output Treatment by GD*: Description

We will describe various kinds of output treatment under Q-PUF in terms of an example based on the 2010 NSDUH PUF data. For simplicity, we consider only three AVs (age, gender, and cocaine use) and the corresponding aggregate level output of unweighted domain counts as shown in the three-dimensional Table 1 based on 57,873 observations. The counts are not disclosure-treated and using the GD* criterion of 50, there are several domains or cells such as (age A1, male, cocaine use) which are at risk of disclosure.

3.1 Descriptive parameters such as domain counts and proportions: The output of estimates of these parameters is needed for comparison of domain or subdomain proportions which may be defined by one factor (e.g., AV category indicators), two factors (e.g., product of two AV category indicators) or more based on several categorical AVs. The shaded domains in Table 2 show an alternative but equivalent representation of cell counts of Table 1 in terms of a set of hierarchical vectors of domain counts. It is hierarchical in that the set starts with domains defined by constant or null AV followed by one AV, two AVs, and so on. There are a total of 40 linearly independent

domains defined by products of indicators of 10 age, 2 gender and 2 cocaine use categories. It is easily seen from these domain counts which ones are not safe (i.e., below the specified threshold of 50) and whether their suppression requires complementary suppression in order to avoid re-engineering of suppressed domain counts.

The nonstandard tabular representation of cell counts in Table 2 is useful in screening AVs for disclosure safety of corresponding domain counts and if necessary transforming their categories by collapsing to meet the threshold. Category collapsing is also performed in a hierarchical manner which is driven by analyst needs and meaningful practical interpretation in that it should allow an AV to have finer categories for lower order factor effects than for higher order factor effects, and AV categories in higher order effects are either same or coarsened versions of AVs in lower order effects. It follows how suitable AVs or their functions (such as products of category indicators from different AVs) could be defined using substantive considerations such that corresponding domains meet the GD* threshold. Incidentally, in Table 2, domains are defined in the usual sense and not as g-domains because the study variable (cocaine use) is categorical.

Table 3 shows a checklist of screened AVs for the above simple example such that the GD* criterion is satisfied by all listed AVs. In particular, for the age AV, one variable factor levels are 10, two variable factor levels are 20 with gender, 14 with cocaine use because of collapsing of age categories A1 to A3, and A9 to A10, and three variable factor levels are 28 with gender and cocaine use. Similarly, for other AVs of age and cocaine use, one, two, and three variable factor effects are defined such that GD* is met. Table 3 also shows rows for noncategorical variables such as # days smoked in the past 30 days where the concept of g-domain (the superscript * on AVs is used to denote g-domains) is required to check GD*; the example considered here, however, doesn't have any noncategorical AV. Once the checklist is specified, disclosure-safe aggregate level output for analysis domains defined by screened AVs become available under Q-PUF as shown in Table 4 in a nonstandard tabular format of a set of hierarchical vectors of domain counts. Notice that the number of linearly independent domains is reduced from a maximum of 40 to 34 due to the GD* restriction on AVs. With screened AVs used in Table 4, no domain count (both in the selected linearly independent set and the remainder set) is below the specified threshold of 50.

Before we consider other kinds of analysis restrictions for output disclosure treatment under Q-PUF, it would be useful for the sake of understanding to introduce a general notation for domain count vectors. Let $\boldsymbol{z}_{D\times 1}$ denote a *D*-vector of cell counts $col_{1\le i\le I}col_{1\le j\le J}col_{1\le k\le K}\{z_{ijk}\}$ stacked in a nested manner from a 3-dimensional table (e.g., Table 1) where *I* is the number of rows (e.g., 10 age categories), *J* is the number of columns (e.g., 2 gender categories), *K* is the number of layers (e.g., two cocaine use categories), and *D* equals *IJK* —the total number of cells; i.e., 40. Here z_{ijk} denotes the count for cell *ijk*, and *col* denotes the column formation operator. Thus the set of hierarchical vectors of linearly independent domain counts (ref. Table 2) can be defined as $\{z_{+++}\}$ for the null variable or the constant factor with only one domain, $col_{1\le i\le I-1}\{z_{i++}\}$ for the single age variable or factor with the number of linearly independent domains being *I-1*, $col_{1\le j\le J-1}\{z_{+j+}\}$ for the gender factor with the number of independent domains being *J-1*, $col_{1\le k\le K-1}\{z_{++k}\}$ for the cocaine use factor with the number of independent domains being *K-1*, $col_{1\le i\le I-1}col_{1\le j\le J-1}\{z_{ij+}\}$ for the age by gender factor with the number of independent domains being *(I-1)(J-1)*, $col_{1\le i\le I-1}col_{1\le k\le K-1}\{z_{i+k}\}$ for the age by cocaine use factor with the number of independent domains being *(I-1)(K-1)*, $col_{1\le j\le J-1}col_{1\le k\le K-1}\{z_{+jk}\}$ for the gender by cocaine use factor with the number of independent domains being *(J-1)(K-1)*, and $col_{1\le i\le I-1}col_{1\le j\le J-1}col_{1\le k\le K-1}\{z_{ijk}\}$ for the age by gender by cocaine use factor with the number of independent domains being *(I-1)(J-1)(K-1)*, where '+' in the subscript is the traditional notation for summing over the subscript to obtain marginal counts. Note that the total number of linearly independent domains in the set of hierarchical vectors (Table 2) is *1+(I-1)+(J-1)+(K-1)+(I-1)(J-1)+(I-1)(K-1)+(J-1)(K-1)+(I-1)(J-1)(K-1) or IJK; i.e., D* as expected. To have a more general and compact notation applicable to domain count vectors from higher dimensional tables, we will denote the hierarchical vectors of domain counts as follows.

$t_{(x0)} = \{z_{+++}\}$,
$\boldsymbol{t}_{(x1)} = col_{1\le i\le I-1}\{z_{i++}\}$,
$\boldsymbol{t}_{(x2)} = col_{1\le j\le J-1}\{z_{+j+}\}$,
$\boldsymbol{t}_{(x3)} = col_{1\le k\le K-1}\{z_{++k}\}$,
$\boldsymbol{t}_{(x1*x2)} = col_{1\le i\le I-1}col_{1\le j\le J-1}\{z_{ij+}\}$,

$\boldsymbol{t}_{(x1*x3)} = col_{1\le i\le I-1} col_{1\le k\le K-1}\{z_{i+k}\}$,
$\boldsymbol{t}_{(x2*x3)} = col_{1\le j\le J-1} col_{1\le k\le K-1}\{z_{+jk}\}$, and
$\boldsymbol{t}_{(x1*x2*x3)} = col_{1\le i\le I-1} col_{1\le j\le J-1} col_{1\le k\le K-1}\{z_{ijk}\}$

where x_0 denotes the whole population domain indicator or the constant factor taking the value of 1 for all observations; x_1 (age), x_2 (gender), x_3 (cocaine use) denote domain indicators corresponding to the three one factors; x_1*x_2, x_1*x_3, and x_2*x_3 are products of domain indicators corresponding to two factors; $x_1*x_2*x_3$ is the product of all the three domain indicators corresponding to three factors, and the vector $\boldsymbol{t}_{(x1)}$, for example, is a column of totals for category indicators of x_1. Now, to meet the GD* criterion (i.e., to make any x-variable part of the screened AV checklist as shown in Table 3), consider transformed versions of x-variables (to be denoted by $\tilde{x}$) whenever necessary, and then the output treated hierarchical vectors of domain counts of Table 4 can be expressed as a column vector of dimension $\widetilde{D}$ (=34 in our example) from the set

$\{t_{(x0)}, \boldsymbol{t}_{(\widetilde{x1})}, \boldsymbol{t}_{(\widetilde{x2})}, \boldsymbol{t}_{(\widetilde{x3})}, \boldsymbol{t}_{(\widetilde{x1}*\widetilde{x2})}, \boldsymbol{t}_{(\widetilde{x1}*\widetilde{x3})}, \boldsymbol{t}_{(\widetilde{x2}*\widetilde{x3})}, \boldsymbol{t}_{(\widetilde{x1}*\widetilde{x2}*\widetilde{x3})}\}$.

Standard error and interval estimates of domain counts and proportions can also be output under Q-PUF using available software.

3.2 Descriptive Parameters such as Domain Totals and Means: The output of domain totals and means is needed for comparison of means of numeric or continuous study variables (y) over domains defined by categorical variables x_1, x_2, and x_3. For g-domains defined by products and x-variables, consider hierarchical collapsing of categories, if necessary, to satisfy the GD*criterion (i.e., the g-domain count for each of zero and nonzero contributions of products of y and x is at least c_0 as explained in Section 2), and the output treated hierarchical vectors of domain totals under Q-PUF can be expressed as a column vector from the set

$\{t_{y(x0)}, \boldsymbol{t}_{y(\widetilde{x1})}, \boldsymbol{t}_{y(\widetilde{x2})}, \boldsymbol{t}_{y(\widetilde{x3})}, \boldsymbol{t}_{y(\widetilde{x1}*\widetilde{x2})}, \boldsymbol{t}_{y(\widetilde{x1}*\widetilde{x3})}, \boldsymbol{t}_{y(\widetilde{x2}*\widetilde{x3})}, \boldsymbol{t}_{y(\widetilde{x1}*\widetilde{x2}*\widetilde{x3})}\}$,

where $\boldsymbol{t}_{y(\widetilde{x1})}$, for example, is the column vector of totals of the product $y(\widetilde{x1})$ variable for domains defined by $\widetilde{x1}$. Standard error and interval estimates of domain totals and means are also output under Q-PUF.

3.3 Analytic Parameters under Linear Regression Models: The output of estimates of regression of continuous study variables (y) over auxiliary variables (x--categorical or continuous) is needed to analyze trend over a given x holding others fixed. Since the y and x AVs are from the checklist, they satisfy GD* in that counts of zero and nonzero values of products of y and x variables over the whole dataset are at least c_0. So for y and x AVs in the checklist, regression coefficient estimates along with their variance and interval estimates can be output under Q-PUF.

3.4 Analytic Parameters under Nonlinear Regression Models: For nonlinear regression such as logistic regression of binary study variables (y) over auxiliary variables (x), point, variance, and interval estimates of model parameters can be output under Q-PUF if y and x variables satisfy the GD* criterion as in the case of linear regression models.

3.5 Unit level Output: So far we considered aggregate level output under Q-PUF. For unit level output such as quantiles and Q-Q plots for distributional diagnostics, additional analysis restrictions in terms of suppression of minimum and maximum observations, and requiring bin size between successive quantiles to be at least c_0' (less than c_0, e.g.,10) are placed. Regression residual diagnostics are performed using aggregates over small sub-domains or groups in order to satisfy GD* and regression predictions are performed over analysis domains satisfying GD* where the threshold for GD* is set at a smaller value like 10.

3.6 Other Restrictions: Some other analysis restrictions that are not too limiting for analysts but important for thwarting intruders comprise allowing only standard transformations of analytic variables such as logit, probit, complementary log-log for binary variables, and log, square-root, inverse, and $\sinh^{-1}$ for other variables. It is known (see e.g., Gomatnam et al., 2005) that with certain nonstandard transformations, it may be possible to disclose information at the unit level—known as threat due to transformation attacks. Also, parameter estimates based on

saturated models are not allowed to protect disclosure of original microdata. Additional restrictions with regard to providing only a lower confidence bound of R^2 (squared multiple correlation coefficient) in regression modeling when it is very close to 1 may be needed because an intruder having the knowledge of covariates about the target used in the model and regression coefficients can predict the true value of the study variable quite well. However, such an intrusion scenario is not very likely because the intruder is likely to know only coarsened versions of the covariates and not the more precise ones used in the model.

3.7 Summary of Q-PUF Output Disclosure Treatment: Since Q-PUF deals with analysis restrictions, we give below a summary of information provided under the output treatment to help users make meaningful interpretations.

Module I. The data producer makes an initial list of all AVs deemed to be of interest to the analyst based on the data codebook and in consultation with subject matter expertise. The AVs could be functions of elementary AVs where elementary AVs, for example, are individual indicators (demographic--age or gender) or cocaine use.

Module II. From the initial list of Module I, a checklist of screened AVs satisfying the GD* criterion is created. It should be noted that it is not necessary to develop in advance an exhaustive list of screened AVs. A reasonable comprehensive initial list can be updated as more AVs are required to meet user needs.

Module III. Examples are provided to users for formulating analysis problems in terms of AVs (study variables and covariates) from the checklist for descriptive and analytic inferences. A list of allowed AV transformations in the codebook for users to consult with is also included.

Module IV. Guidelines for users are provided on following the protocol to submit web-based queries about AVs in the checklist for indirect access to the raw microdata in a data enclave environment. Program codes for queries are based on customization of available analysis software such that the unit level output can be suitably modified (see Module V).

Module V. For unit level output, additional restrictions are specified to modify it to an aggregate level such that GD* with a smaller threshold is satisfied. User is informed about such modifications in that while the aggregate level output for any analysis is automatically protected from disclosure because AVs are from the checklist, the unit level output is nonstandard and has an aggregate level form for disclosure-safety.

4. Theoretical Properties of Q-PUF

We will consider properties of Q-PUF with respect to confidentiality and reliability of analysis output after the output treatment under GD*. It is seen that GD* provides a unified approach to disclosure-safety and reliability for a wide class of analyses.

4.1 Confidentiality: We list here the main observations.

(a) For domain counts, disclosure risk can be made negligible with a large threshold for GD*. In practice, a working value of the threshold c_0 can be specified (e.g., $c_0 = 50$) although it would depend on the sensitivity of information in the dataset and the need for also protecting analytic utility as mentioned in the introduction. Moreover, since AVs for small domain counts are not included in the checklist and therefore such counts are not released on their own under Q-PUF but only as part of collapsed domains, any attempt by an intruder to reverse-engineer small domain counts from other allowed domain counts containing small cell counts is not likely to be very successful. The reason for this is that the Fréchet bounds (tightest possible; see Dobra and Fienberg, 2000) constructed from marginal counts become quite loose if the threshold c_0 is sufficiently large.

More specifically, for Q-PUF results (Table 4) for the simple 2010 NSDUH PUF example, consider domain counts defined by cocaine use and age where age categories A1-A3 and A9-A10 are collapsed to meet GD*. In the hierarchy of count vectors, the lowest level requiring collapsing is the 2-dimensional count vector for cocaine use by age. For the 2x2 Table 5.1 of counts defined by (A9, A10) and cocaine use, let $\{z^*_{11}, z^*_{12}, z^*_{21}, z^*_{22}\}$ denote the four unknown counts as they are not released to the public due to z^*_{21} being too small, but the margins, $z^*_{1+} = 8376$, $z^*_{2+} = 5667$, $z^*_{+1} = 169$, $z^*_{+2} = 13874$, and $z^*_{++} = 14043$, are known to the public as they can be derived from the treated output set of hierarchical domain count vectors (Table 4; shaded domain counts are released to the public).

The sharp Fréchet bounds for any of the cell counts z^*_{jk} in Table 5.1 are given by (see inequality [3] in Dobra and Fienberg, 2000),

$$\max\{z^*_{j+}+z^*_{+k}-z^*_{++},\, 0\} \leq z^*_{jk} \leq \min\{z^*_{j+}, z^*_{+k}\}.$$

To see why above bounds can be very loose for sufficiently large c_0, suppose $\min\{z^*_{j+}, z^*_{+k}\}$ is z^*_{j+} without loss of generality. Then the difference between upper and lower bounds clearly equals z^*_{j+} if $\max\{z^*_{j+}+z^*_{+k}-z^*_{++},\, 0\}$ is 0, and equals $z^*_{+k'}$ if $\max\{z^*_{j+}+z^*_{+k}-z^*_{++},\, 0\}$ is not zero because $z^*_{j+} - (z^*_{j+}+z^*_{+k}-z^*_{++})$ is $z^*_{+k'}$ where $k \neq k'$. The desired claim about the gap between upper and lower bounds follows since all the marginal are at least as large as the threshold c_0.

For the 3x2 Table 5.2 with uncollapsed domains defined by (A1, A2, and A3) and cocaine use, one can also show bounds for each cell are quite loose for large c_0. Table 4 also contains collapsing of domains defined by the three variables. It easily follows from the difficulty in obtaining useful bounds for uncollapsed domain counts released as collapsed ones in the 2-dimensional table defined by Cocaine Use and Age, that bounds for uncollapsed domains in the 3-dimensional table defined by cocaine use, gender and age would be even looser because it would also depend on the precision of estimates (in terms of bounds) for uncollapsed domains defined by the lower dimensional table for cocaine use and age.

(b) For tables of magnitude data; i.e., cell totals of numeric AVs, disclosure-safety of cells with small counts of nonzero values of the numeric AV is automatically protected under the GD* criterion. The reason is that the bounds for g-domain sizes from collapsed g-domains under GD* would again be rather loose in view of the above result in (a). In fact this property of GD* for disclosure-safety of any g-domain total (of binary, or numeric, or continuous or semicontinuous) holds as long as each of the g-domains defined separately by zero and nonzero contributions of the product of the study variable (y) and auxiliary variables (x--defining domains) satisfy the threshold. Note that for a sample of size n, a domain total in general is defined as $\sum_{k=1}^{N} x_k y_k w_k \delta_k$ where w_k is the sampling weight for unit k, δ_k is the sample inclusion indicator taking the value of 1 if the unit k is selected in the sample and 0 otherwise, and the population total parameter is $\sum_{k=1}^{N} x_k y_k$. It should also be noted that any disclosure risk due to outliers in domain totals is considerably diffused when a high threshold is used under GD*.

(c) For analytic parameters such as the linear regression coefficient parameter β in a model with a single covariate and no intercept, the estimate is obtained as a solution of the estimating function given by $\sum_{k=1}^{N} x_k(y_k - \beta x_k) w_k \delta_k$ which is set to 0 to obtain the usual estimate of the slope β as $\hat{\beta} = \sum_{k=1}^{n} x_k y_k w_k / \sum_{k=1}^{n} x_k^2 w_k$. It is observed that the estimating function behaves like a domain total and is equal to zero at the estimate $\hat{\beta}$. If the number of nonzero terms $x_k(y_k - \beta x_k) w_k \delta_k$ in the estimating function evaluated at $\hat{\beta}$ (assuming that the model is unsaturated, else all terms will be zero) is at least c_0, the output of the regression parameter estimate can be considered disclosure–safe. The reason for this is that if the number of nonzero values of the product $x_k y_k$ is very small, for instance, in the above example of regression problem with a single covariate and no intercept, the estimated regression parameter might disclose the value of the dependent variable. It is remarked that it is sufficient to check for products $x_k y_k$ to satisfy the GD* criterion as their compliance automatically implies that x_k-variables satisfy GD* and hence the products $x_k(y_k - \beta x_k) w_k \delta_k$ evaluated at $\hat{\beta}$, assuming that the model is unsaturated; a saturated model is, of course, not of interest in practice. It is also remarked that in practice often the estimating function for β may involve a covariance matrix due to heterogeneity of observation errors. However, for disclosure-safety, it is sufficient to work with the identity matrix as a working covariance matrix; i.e., it is sufficient to check the GD* criterion for the products $x_k y_k$.

(d) When dealing with nonlinear regression models, in particular generalized linear models with canonical parameters, estimating functions take the same form as in the case of linear regression with uncorrelated errors and constant variance; i.e., $\sum_{k=1}^{N} x_k(y_k - \mu_k(\beta, x_k)) w_k \delta_k$, and so it is sufficient to check the GD* criterion for the products $x_k y_k$. For more general nonlinear models, we can always use the identity matrix as a working covariance to simplify checking of the GD* criterion in terms of $x_k y_k$.

(e) For confidentiality of unit level output, use of an analysis output filter to modify it to an aggregate level in order to satisfy GD* with a smaller threshold provides the necessary protection. In particular, regression

predictions are not released at unit levels but for analysis domains satisfying GD*. Similarly, for regression diagnostics based on residuals, if the analysis goals are at higher levels of aggregation as is often the case, it may be sufficient to perform these diagnostics at suitable low levels of aggregation as in the case of commonly used aggregate level models. In other words, g-domains can be defined as small groups or sub-domains (these are not like regular analysis domains but small ones created typically by finer categories of geo-demographics in population surveys) with residual analysis based on group means. For example, for the gth group, the total residual r_g and the corresponding standardized residual r_g^* are given by $r_g = \sum_{k\ in\ g}((y_k - \mu_k(\boldsymbol{x}_k, \hat{\beta}))w_k$, $r_g^* = r_g/se(r_g)$, where μ_k is, in general, a nonlinear mean function for the kth observation y_k with the corresponding predicted value $\mu_k(\boldsymbol{x}_k, \hat{\beta})$, and w_k is the sampling weight.

4.2 Analytic Utility: The main observations are given below**.**

(a) The reliability of domain counts and hence their analytic utility is ensured by making the GD* threshold c_0 large enough such as 50. It also applies to domain totals because the corresponding g-domain size will be large enough. Note that the output restriction for domains with g-domain size below the threshold is not a serious restriction because of instability of corresponding domain totals.

(b) For analytic parameters, the corresponding estimating functions are like domain totals for given parameter values. They are unbiased estimates of known zero totals, and their reliability or precision depends on the g-domain sample size. If the estimating function is precise as an estimate of zero, then its solution or the parameter estimate obtained by setting it equal to zero is also precise; this follows easily from the sandwich form of estimate's variance-covariance matrix. Note that the above discussion in terms of estimating functions applies to both frequentist and Bayesian approaches to inference since screening of AVs in the frequentist approach for reliability and confidentiality can also serve as a screening check of AVs for taking the Bayesian approach.

(c) For regression model diagnostics, use of standard residual diagnostics at suitable low levels of aggregation (i.e., small subdomains as in diagnostics for aggregate level models) may often be adequate in practice since the analysis goals are typically at higher levels of aggregation. The residual plot would, of course, not show inadequacy of the model for unit level predicted values and extreme observations. However, if the model behaves well at lower levels of aggregation, it may certainly be deemed adequate for higher aggregate level predictions keeping in mind that no model is perfect.

(d) For allowable AVs in the checklist, Q-PUF provides exact results for aggregate level output—i.e., with very high analytic utility. However, for unit level output, a compromise is made by transforming the output to an aggregate level but its utility may still be adequate. For instance, regression predictions for individual observations are not provided, but are provided for aggregate levels of analysis domains as long as they satisfy GD* --here we require only a smaller threshold because it is driven by confidentiality and not reliability concerns. Note that with the output of regression coefficients, an intruder can predict the target's y-value if x-values are known although it is unlikely for the intruder to know x-values precisely as mentioned in subsection 3.6.

(e) Under Q-PUF, variance estimation of estimates for both descriptive and analytic inferences do not pose any problems as standard software can be used to compute these estimates including adjustments for taking account of the sampling design because they can be viewed as aggregate level output based on squares and cross-products of AVs.

5. Comparison of Q-PUF with Alternative Methods for Remote Analysis Servers

The topic of output disclosure treatment methods for remote analysis servers is a very active research area and as mentioned in the introduction, several methods are currently under development. One of the main methods is the Microdata Analysis System (MAS) being developed at the Census Bureau; Lucero et al. (2011). We discuss below the MAS in comparison to Q-PUF.

(a) While the MAS allows for the flexibility of user-specified AVs (although somewhat limited due to data-driven restrictions in any particular application), Q-PUF does not. Instead it restricts users to a checklist of screened AVs based on the GD* criterion. It is for this reason Q-PUF is not subject to differencing attacks by repeated queries

because AVs cannot be modified from the checklist. However, the MAS addresses this problem by a clever approach based on random subsampling termed 'drop q' rule where the number q of observations to be sampled out is chosen at random from a uniform distribution over 2 to an upper bound; the interval is chosen such that the disclosure risk for certain types of differencing attack for a given dataset is small. In repeated queries for the same domain, the same subsample is used by retaining the random seed.

(b) The analysis results in the MAS are not exact as they are perturbed leading to a slight loss of efficiency. In Q-PUF, there is no perturbation of aggregate level output and the results are exact for AVs from the checklist. However, for unit level output, results are modified by aggregation to satisfy GD*. In particular, regression residual diagnostics and regression predictions are performed only at an aggregate level. The MAS, on the other hand, uses synthetic residuals for regression diagnostics—a novel approach due to Reiter (2003) in which a synthetic dataset is created for each set of AVs used in modeling.

(c) The MAS employs the usual condition of cell (or domain or universe) sample size threshold for confidentiality of count data (termed universe gamma rule where gamma denotes the threshold), and also another related rule termed 'no marginal of 1 or 2'. If the threshold is not met, the corresponding output is suppressed. This is unlike Q-PUF where there is no suppression but hierarchical collapsing of domains is performed leading to a nonstandard table of counts which can be expressed as a set of hierarchical domain count vectors. The specification of hierarchical collapsing is based on a meaningful practical interpretation for analysts. In the case of regression models, the MAS also utilizes collapsing to deal with sparse categories of x-variables but it is not hierarchical as in the case of Q-PUF.

(d) In the MAS, although there is quite a bit of flexibility in user specification of AVs, it allows only a limited range of analysis; in particular, up to three way interactions with categorical data and up to 20 auxiliary variables (not counting interactions) in regression analysis. Also, exploratory analysis is not allowed because it is difficult to anticipate and protect against possible differencing attacks as there is no control on user-specified AVs. However, with the concepts of g-domain and GD* criterion based on estimating functions, Q-PUF uses a unified and systematic approach to screening and listing of AVs allowing for many different types of analysis or queries which by themselves (i.e., without mapping to AVs and any control through AVs) may be prohibitively large to list. For AVs in the checklist, all types of analysis including exploratory and complex can be conducted under Q-PUF. On the other hand, the MAS is currently designed for simple analyses, and can serve as a precursor to performing complex analyses with Census Bureau's Research Data Centers which require both user authorization and output disclosure-safety review.

We now briefly discuss other important contributions in the area of indirect query-based access through remote analysis servers. The PPA system of CSIRO, Australia (Sparks et al., 2008) proposes innovative ideas for disclosure-safe analysis such as use of correspondence analysis for table of counts, robust regression to lessen the influence of outliers, and parallel box plots instead of the usual residual plots—use of parallel box plots is somewhat analogous to aggregation of residual output under Q-PUF. Unlike MAS, no subsampling is performed to deal with the differencing attack scenario, but limitations on too many closely related queries are placed by keeping track of user requests. The ANDREW system of NCHS (see e.g., Gambhir and Harris, 2006) is under development for an automated web-based access to microdata requiring user authorization for analysis proposal, and review of program codes for automated output restrictions by disallowing certain commands. Research is also underway to find a suitable solution to the problem of differencing attack. The DAS system of NCES (see e.g., Russell and White, 2011) is one of the early online analysis tools for table generation as an alternative to licensing for restricted use datasets. DAS for tabular data puts several restrictions such as analysis up to 4-way tables, and proportion estimates with 30 or more cases in the denominator and 3 or more in the numerator. Other online analysis tools include PowerStats for more complex tables and regression. The differential privacy methodology of Dwork (2006) approaches the problem in a completely different way but is attractive as it provides guarantees for disclosure-safety of output under the methodology in general by using a key noise parameter epsilon without imposing any data-specific analysis output restrictions as in earlier methods. It imposes a stringent criterion on privacy and requires that output of summary statistics be insensitive to presence or absence of any one observation. In an empirical study, Fienberg et al. (2010) provide an evaluation of differential privacy in the context of risk-utility trade-off for tabular data, and found problems with a suitable specification of the epsilon parameter in the interest of analytic utility.

6. Summary and Remarks

We proposed a new criterion, termed GD*, for disclosure-safety of analysis output from a query-based access to microdata through remote analysis servers. The proposed criterion is defined in terms of AVs (elementary or composite), and for each domain (defined by 0/1 values of the categorical study variable and other categorical auxiliary variables) or g-domain (defined by zero and nonzero values of a numeric or continuous study variable and other categorical/continuous auxiliary variables), it requires a minimum threshold for each count of zero and nonzero values of the corresponding AV for individuals in the domain or g-domain. It generalizes the usual domain sample size threshold criterion for count data, and provides a unified approach, based on estimating functions, to confidentiality and reliability of output for a wide class of analyses—both descriptive and analytic. For AVs satisfying GD*, the aggregate level output (such as totals, means, ratios, and regression parameters) are automatically disclosure-safe, but for unit level output (such as quantiles and regression predictions), additional disclosure treatment is performed by transforming the output to an aggregate level so that they satisfy GD* but with a lower value of the threshold suitably chosen for protecting only confidentiality as output reliability is not an issue here.

With GD* criterion to screen AVs, it was feasible to list allowable AVs needed for various kinds of analysis and thus the concept of pre-constructed checklist was introduced to help users choose suitable AVs in advance for their problem specification. While listing all potential allowable queries may be prohibitive, the checklist in terms of allowable AVs is feasible as many different analyses can be conducted with the same set of AVs. The checklist is data-specific, is based on commonly used AVs from the dataset, and can be updated over time to incorporate more AVs depending on user needs. So it is not expected to be too limiting for the analyst although the analyst is restricted to choose AVs from the checklist. Specification of screened AVs is somewhat analogous to coarsening of identifying and sensitive variables in the creation of traditional PUFs based on common analytic needs, where users do not have any freedom to define their own coarsening or specification of AVs.

Screening of AVs for the checklist is performed in a hierarchical manner in that domains or g-domains defined in order by one AV, two AVs, and more are checked to meet the GD* criterion. If at any stage, GD* is not satisfied by a domain (or g-domain), then collapsing of categorical AVs defining that domain is also performed in a hierarchical manner so that collapsed categories of an AV at lower levels of hierarchy are consistent with higher level collapsing and guided by commonly used analysis practices. The above hierarchical collapsing in the construction of checklist avoids suppression of domains but they contribute instead to the output in combination with other domains. This feature is highly desirable and conducive for analysis as it leads to practically meaningful and analytically useful output. In particular, it leads to a nonstandard tabular representation (i.e., as a set of hierarchical vectors of domain counts) of disclosure-safe output of counts. It was shown that the reconstruction of uncollapsed small domain (or cell) counts from collapsed domains (or margins) cannot be very successful because of loose Fréchet bounds (although tightest possible) if GD* threshold is not too small. This observation carries over to output for numeric or continuous AVs by considering g-domain counts.

The GD* criterion used for constructing a checklist of AVs leads to Q-PUF because by restricting analysis output to those corresponding to AVs in the checklist, there is no need for user authorization or output disclosure review even for complex or exploratory analyses. There is also no problem of differencing attack by repeated queries because users are not free to specify AVs arbitrarily. In addition, under Q-PUF, there is no need to perturb aggregate level output for AVs in the checklist and thus aggregate level analysis results are exact. A theoretical comparison of Q-PUF with alternative methods such as the MAS at the Census Bureau currently under development was also made. In particular, unlike Q-PUF, the MAS allows for user-specified AVs but analysis output is perturbed by subsampling to thwart differencing attacks, and is restricted to simple types of analysis which are not conducive for exploratory analyses.

Acknowledgments

The research in this article was supported in part by the Centers for Medicare and Medicaid Services under contract number 500-2006-00007I/#T0004 for the Medicare Claims CER Public Use Data Pilot Project. The views expressed in this article are those of the authors and do not necessarily reflect the views of the U.S. Department of Health and Human Services or the Centers for Medicare and Medicaid Services. We would like to thank Chris Haffer of CMS for

his support and encouragement, and Mike Davern and Dan Kasprzyk of NORC for helpful comments. Thanks are also due to Laura Zayatz, Phil Steel, and Michael Freiman of Census Bureau for information about the MAS, Peter Meyer of NCHS about ANDREW, Marilyn Seastrom and Andy White of NCES about DAS, and Bill Winkler of Census Bureau for his useful comments as a discussant of the paper presented at the 2012 FCSM Conference, January 10-12.

References

Borton, J.M., Yu, A.T.-C., Crego, A.M., and Singh, A.C. (2011). Evaluation and Limitations of Disclosure-Treated Health Data Using Random Substitution and Sub-sampling. *Proceedings of Survey Research Methods Section*, American Statistical Association.

Dobra, A. and Fienberg, S.E. (2000). Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Proceedings of the National Academy of Sciences*, 97, 11885-11892.

Dwork, C. (2006). Differential privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) *ICALP 2006. LNCS,* Springer :Heidelberg, Vol. 4052, pp 1-12.

Fienberg, S.E., Rinaldo, A., and Yang, X. (2010). Differential privacy and the risk-utility trade-off for multidimensional contingency tables. In Domingo-Ferrer, J., and Magkos, E. (Eds). *Privacy in Statistical database*, LNCS 6344, Springer: Heidelberg, pp. 187-199.

Gambhir, V., and Harris, K.W. (2006). ANalytical Data Research by Email and Web (ANDREW). Monograph of Official Statistics, UN/ECE Work Session on Statistical Data Confidentiality, Geneva, Nov 9-11(2005), 53-56.

Gomatnam, S, Karr, A.F., Reiter, J.P, and Sanil, A.P. (2005). Data Dissemination and Disclosure limitation in a World without Micro-data: A risk-Utility Framework for Remote Access Analytic Servers. *Statistical Science*, Vol 20, 163-177.

Keller-McNulty, S. and E. Unger (1998). A database system prototype for remote access to information based on confidential data. *Journal of Official Statistics*, 14:347–360

Lucero, J., Freiman, M., Singh, L., You, J., DePersio, M. and Zayatz, L. (2011). The Microdata Analysis System at the U.S. Census Bureau, SORT Special issue: Privacy in statistical databases, 2011, 77-98

Reiter, J. P. (2003). Model diagnostics for remote-access regression systems. *Statistics and Computing*, 13, 371-380.

Rowland, S. (2003). An examination of monitored remote microdata access systems. In *National Academy of Sciences Workshop on Data Access*, Washington, DC, October 16-17.

Russell, J.N. and White, A.A.(2011). The NCES data analytic system: public access to restricted-use data. Paper presented at the *American Statistical Association Joint Statistical Meetings*, Miami, FL.

Seastrom, M. and Kaufman, S. (2003). NCES disclosure risk procedures. Paper presented at the *American Statistical Association Joint Statistical Meetings*, San Francisco, CA.

Singh, A.C. and Borton, J.M. (2012). Aggregate Level PUF as a new alternative to the traditional unit level PUF for improving analytic utility and data confidentiality. Technical Report, Center for Excellence in Survey Research, NORC at the University of Chicago, (to be presented at JSM, San Diego)

Sparks, R., C. Carter, J. Donnelly, C. O'Keefe, J. Duncan, T. Keighley, and D. McAullay. (2008). Remote access methods for exploratory data analysis and statistical modeling: Privacy-Preserving Analytics. *Computer Methods and Programs in Biomedicine*, 91(3):208–222, 2008. ISSN 0169-2607.

Steel, P. and A. Reznek. (2005). Issues in designing a confidentiality preserving model server. *Monographs of Official Statistics*, 9:29, 2005.

Table 1: Untreated Output of Counts for Domains defined by Cocaine Use by Age and Gender
(Based on unweighted 2010 NSDUH PUF; Ten Age categories show a granular level of available categorization)

Age	Cocaine User		Margin Age x Coc	Not Cocaine User		Margin Age x NonCoc	Row Marginal Total
	Male	Female		Male	Female		
A1 (12-13)	3	7	10	3012	2957	5969	5979
A2 (14-15)	18	21	39	3095	3040	6135	6174
A3(16-17)	63	58	121	3268	3072	6340	6461
A4(18-19)	116	97	213	2453	2535	4988	5201
A5 (20-21)	163	72	235	2204	2420	4624	4859
A6(22-23)	144	78	222	2049	2387	4436	4658
A7(24-25)	128	61	189	2020	2385	4405	4594
A8(26-34)	110	58	168	2605	3131	5736	5904
A9(35-49)	83	56	139	3753	4484	8237	8376
A10 (50+)	23	7	30	2488	3149	5637	5667
Column Marginal Total	851	515	**Subtotal** 1366	26947	29560	**Subtotal** 56507	**Grand Total** 57,873

Table 2: Alternative but Equivalent Representation of Table 1 of Untreated Counts
(Set of hierarchical domain count vectors)

Constant (null AV) & One AV	**Product of Two AVs (or 2-factor)** (Domains defined by products of categoriy Indicators)		**Product of Three AVs (or 3-factor)** Coc, Gender, and Age	
Constant	Coc and Gender		Gender and Age given Coc User	
57873(Total)	851 (C,M)	26947(NC,M)	3 (C,M,A1)	7 (C,F,A1)
Coc	515 (C,F)	29560(NC,F)	18 (C,M,A2)	21 (C,F,A2)
1368 (C)	Coc and Age		63 (C,M,A3)	58 (C,F,A3)
56507(NC)	10 (C,A1)	5969 (NC,A1)	116 (C,M, A4)	97 (C,F, A4)
Gender	39 (C,A2)	6135 (NC,A2)	163 (C,M,A5)	72 (C,F,A5)
27798(M)	121 (C,A3)	6340 (NC,A3)	144 (C,M, A6)	78 (C,F, A6)
30075(F)	213 (C,A4)	4988 (NC,A4)	128 (C,M, A7)	61 (C,F, A7)
Age	235 (C,A5)	4624 (NC,A5)	110 (C,M, A8)	58 (C,F, A8)
5979 (A1)	222 (C,A6)	4436 (NC,A6)	83 (C,M,A9)	56 (C,F,A9)
6174 (A2)	189 (C,A7)	4405 (NC,A7)	23 (C,M,A10)	7 (C,F,A10)
6461 (A3)	168 (C,A8)	5736 (NC,A8)	Gender and Age given Not Coc User	
5201 (A4)	139 (C,A9)	8237 (NC,A9)	3012 (NC,M,A1)	2957 (NC,F,A1)
4859 (A5)	30 (C,A10)	5637 (NC,A10)	3095 (NC,M,A2)	3040 (NC,F,A2)
4658 (A6)	Gender and Age		3268 (NC,M,A3)	3072 (NC,M,A3)
4594 (A7)	3015 (M, A1)	2964 (F, A1)	2453 (NC,M, A4)	2535 (NC,F, A4)
5904 (A8)	3103 (M, A2)	3061 (F, A2)	2204 (NC,M,A5)	2420 (NC,F,A5)
8376 (A9)	3331 (M, A3)	3130 (F, A3)	2049 (NC,M, A6)	2387 (NC,F, A6)
5667 (A10)	2569 (M, A4)	2632 (F, A4)	2020 (NC,M, A7)	2385 (NC,F, A7)
	2367 (M, A5)	2492 (F, A5)	2605 (NC,M, A8)	3131 (NC,F, A8)
	2193 (M, A6)	2465 (F, A6)	3753 (NC,M,A9)	4484 (NC,F,A9)
	2148 (M, A7)	2446 (F, A7)	2488 (NC,M,A10)	3149 (NC,F,A10)
	2715 (M, A8)	3189 (F, A8)	*C: Cocaine user; NC: Not Cocaine User*	
	3836 (M, A9)	4540 (F, A9)		
	2511 (M, A10)	3156 (F, A10)		

Note: Ten detailed age categories are--A1: 12-13, A2: 14-15, A3: 16-17, A4: 18-19, A5: 20-21, A6: 22-23, A7: 24-25, A8: 26-34, A9: 35-49, A10: 50+. There are a total of 40 linearly independent domains (an example is shown by shaded domains) in the above nonstandard table. The non-shaded domain counts can be derived from the shaded ones. The domain vectors are hierarchical in that categories at higher factor levels respect the categories at lower factor levels.

Table 3: Q-PUF Checklist of Screened AVs for the NSDUH Example

Analytic Variables (AVs)		**One Variable or Factor** (g-domains defined by category* indicators)	**Two Variables or Factors** (g-domains defined by products of two category or category* indicators)	**Three Variables or Factors** (g-domains defined by products of three category or category* indicators)
Categorical	**Age**	Age (10)	Age (10) x Gender (2) Age (7) x Coc (2)	Age (7) x Gender (2) x Coc (2)
	Gender	Gender (2)	Gender (2) x Age (10) Gender (2) x Coc (2)	Gender (2) x Age (7) x Coc (2)
	Coc Use	Coc (2)	Coc (2) x Gender (2) Coc (2) x Age (7)	Coc (2) x Gender (2) x Age (7)
	Categorized Version of Numeric ($\tilde{y}$)	$\tilde{y}$	$\tilde{y}$ x Age $\tilde{y}$ x Gender Etc.	$\tilde{y}$ x Age x Gender Etc.
Noncategorical (numeric or continuous)	**y (e.g., # days smoked in the past 30 days)**	y*	y* x Age y* x Gender Etc	y* x Age x Gender Etc.

Note: The underscored AV in each row represents the AV under consideration for one, two or higher order factor effects. The 10 age categories are A1-A10 as in Table 1. The coarsened 7 age categories are A1+A2+A3, A4-A8, and A9+A10. The notation category* is used to define g-domains by categorizing numeric variables in two categories—zero and nonzero values. Numeric AVs are screened by checking the GD* criterion on g-domains and can be included in the checklist as shown in the last row—these AVs are not needed though for the simple NSDUH example considered here.

Table 4: Treated Output Set of Hierarchical Vectors of Domain Counts under Q-PUF
(Nonstandard Tabular Representation of Cell Counts; GD* criterion of 50)

Constant (null AV) & One AV	Product of Two AVs (or 2-factor) (Domains defined by products of category Indicators)		Product of Three AVs (or 3-factor) Coc, Gender, and Age	
Constant	Coc and Gender		Gender and Age given Coc User	
57873(Total)	851 (C,M)	26947(NC,M)	84 (C,M,A1+A2+A3)	86 (C,F,A1+A2+A3)
Coc	515 (C,F)	29560(NC,F)	116 (C,M,A4)	97 (C,F,A4)
1368 (C)	Coc and Age		163 (C,M,A5)	72 (C,F,A5)
56507(NC)	170 (C,A1+A2+A3)	18444(NC,A1+A2+A3)	144 (C,M,A6)	78 (C,F,A6)
Gender	213 (C,A4)	4988 (NC, A4)	128 (C,M,A7)	61 (C,F,A7)
27798(M)	235 (C,A5)	4624 (NC, A5)	110 (C,M,A8)	58 (C,F,A8)
30075(F)	222 (C,A6)	4436 (NC,A6)	106 (C,M,A9+A10)	63 (C,F,A9+A10)
Age	189 (C,A7)	4405 (NC,A7)	Gender and Age given Not Coc User	
5979 (A1)	168 (C,A8)	5736 (NC, A8)	9375 (NC,M, A1+A2+A3)	9069 (NC,F, A1+A2+A3)
6174 (A2)	169 (C,A9+A10)	13874 (NC, A9+A10)	2453 (NC, M,A4)	2535 (NC,F,A4)
6461 (A3)	Gender and Age		2204 (NC, M,A5)	2420 (NC,F,A5)
5201 (A4)	3015 (M, A1)	2964 (F, A1)	2049 (NC, M,A6)	2387 (NC,F,A6)
4859 (A5)	3103 (M, A2)	3061 (F, A2)	2020 (NC, M,A7)	2385 (NC,F,A7)
4658 (A6)	3331 (M, A3)	3130 (F, A3)	2605 (NC,M, A8)	3131 (NC,F, A8)
4594 (A7)	2569 (M, A4)	2632 (F, A4)	6241 (NC,M,A9+A10)	7633 (NC,F, A9+A10)
5904 (A8)	2367 (M, A5)	2492 (F, A5)		
8376 (A9)	2193 (M, A6)	2465 (F,A6)		
5667 (A10)	2148 (M, A7)	2446 (F, A7)	*C: Cocaine user; NC: Not Cocaine User*	
	2715 (M, A8)	3189 (F, A8)		
	3836 (M, A9)	4540 (F, A9)		
	2511 (M, A10)	3156 (F, A10)		

Note: There are a total of 34 linearly independent domains (an example is shown by shaded domains) in the above nonstandard table. The non-shaded domain counts can be derived from the shaded ones. The domain vectors are hierarchical in that categories at higher factor levels respect the categories at lower factor levels or are collapsed versions of them. Moreover, the order in which AVs are selected for collapsing and choice of categories for an AV to be collapsed in order to meet the GD* criterion follows a

prespecified hierarchy based on substantive considerations. In the above example, neighboring age categories are collapsed first before collapsing of gender and cocaine use categories if needed.

Table 5.1 Sub-table of Cocaine Use by Age (A9, A10)
(An illustration of attempted re-engineering of small cell counts)

	Cocaine User	**Not Cocaine User**	**Age Margin**
Age (A9)	139 (z^*_{11}) (0, 169)	8237 (z^*_{12}) (8207, 8376)	8376 (z^*_{1+})
Age (A10)	30 (z^*_{21}) (0, 169)	5637 (z^*_{22}) (5498,5667)	5667 (z^*_{2+})
Coc Use Margin	169 (z^*_{+1})	13874 (z^*_{+2})	14043

Note: Underlined cell counts appear as collapsed counts in the output treated Table 4 under Q-PUF due to small cell count of 30. The intervals in each cell show Fréchet bounds—the interval length in this example turns out to be common and equal to 169.

Table 5.2 Sub-table of Cocaine Use by Age (A1, A2, A3)
(Another illustration of attempted re-engineering of small cell counts)

	Coc User	**Not Coc user**	**Age Margin**
Age (A1)	10 (0, 170)	5969 (5809, 5979)	5979
Age (A2)	39 (0, 170)	6135 (6004, 6174)	6174
Age (A3)	121 (0, 170)	6340 (6291, 6461)	6461
Coc Use Margin	170	18444	18614

Note: Underlined cell counts appear as collapsed counts in the output treated Table 4 under Q-PUF due to small cell counts of 10 and 39. Interval lengths in this example also turn out to be common and equal to 170.