Developing and Testing the Microdata Analysis System

Michael H. Freiman[1], Amy Lauger[1], Marlow Lemons[1], Bryan Schar[1], Kyle Hasenstab[2]
[1]U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233
[2]University of California, Los Angeles, 8125 Math Sciences Bldg., Box 951554, Los Angeles, CA 90095

Abstract
The Census Bureau is developing a Microdata Analysis System (MAS) that will provide custom tables and other analyses based on user specifications. The underlying microdata will be fully protected because calculations for the MAS will take place behind two firewalls and only the final output will be provided to the user. Results will be given only if the output passes a set of disclosure rules. We describe the planned capabilities of the MAS, as well as the testing and evaluation of the disclosure protection measures and some attacks that are being evaluated to ensure that the system only gives data that are safe to release. We focus on assessing and preventing one type of attack: a differencing attack in which the user produces the same table for two very similar universes and subtracts the numbers in the two tables to reveal information about an individual respondent. We consider the prevalence and risks of such an attack in the absence of protections and how the disclosure measures in place protect against an attack.

Key Words: Confidentiality, disclosure avoidance, remote access, tabulation

1. Introduction

As with other statistical agencies in the United States and around the world, the U.S. Census Bureau collects a large amount of data and wishes to disseminate those data as widely and usefully as possible. At the same time, the Census Bureau is well aware that all data are collected under a pledge that no individual respondent will be identified. In the case of the Census Bureau, the data are legally protected by Title 13, U.S. Code, the statute under which the data are collected. In addition, other laws, such as the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA), protect U.S. government data collected under a pledge of confidentiality. In order to fulfill its mission while meeting its legal and ethical obligations, the Census Bureau must use innovative methods of releasing its data.

The Census Bureau currently has several means of making data available. These dissemination methods include tables and other data products available online from the decennial census, the American Community Survey and other surveys conducted by the Census Bureau. In addition, some of our surveys and censuses include public use microdata files that give individual records of all or a sample of the surveyed units. Users who need a particular table that is not available by these means may request a special tabulation made to the user's specifications. For researchers who find these avenues insufficient—particularly those doing more involved analyses—the Census Bureau currently operates a network of physical Federal Statistical Research Data Centers (RDCs), with 24 operating as of the end of 2015 and more set to open in the near future.

At the RDCs, approved researchers may conduct research based on data from the Census Bureau, the Agency for Healthcare Research and Quality (AHRQ) and the National Center for Health Statistics (NCHS), including data from surveys co-sponsored by a number of other government agencies. The Census Bureau releases numerous other data products, which further allow a user to access data.

Each of these data access methods has limitations, however. The Census Bureau's published tables collectively have billions of estimates in them, but these estimates may not combine variables and geography in the way the user needs. The public use microdata may be coarsened, have some variables removed, have limited geographic detail or be otherwise modified to protect the data from individual identification. The methods for modification include adding noise to certain variables and using synthetic data, among others. Special tabulations and access to RDCs are available only at a substantial cost to the user, and using Census Bureau data at the RDCs further requires that the researcher's proposal is approved by the Census Bureau, that the researcher undergoes a background investigation and that Census Bureau staff check the output for compliance with applicable confidentiality rules.

Part of the solution to this set of challenges is the Microdata Analysis System (MAS), an online remote access system under development at the Census Bureau, which will allow users to request a custom tabulation of the data, free of charge. Calculations will take place behind two firewalls, and if the table is approved, it will be returned to the user. For the ACS, our intention is that the user may request a cross-tabulation of variables on a pre-specified universe, with the total number of variables used to define the universe and the table cells tentatively capped at 4. We intend to allow the user to create a universe based on both geographical restrictions—such as examining only an individual county or a collection of pre-specified counties—and non-geographical restrictions, such as considering only women age 55 or older. The geographic universe for a table must be made up of basic building block geographies the system makes available, which will vary depending on the survey. Users may also combine categories of a variable into one row/cell of the table—for example, a variable for educational attainment may separate master's degree holders and doctoral/professional degree holders into different groups, but the user may choose to combine them for the purpose of the table.

Several other agencies in the US and around the world have or are developing remote access systems; such systems include Real Time Remote Access at Statistics Canada (Simard, 2011), the Remote Access Data Laboratory (RADL) at the Australian Bureau of Statistics (2014), LISSY from the LIS Cross-National Data Center in Luxembourg (2011), ANDRE at the U.S. National Center for Health Statistics (Meyer, 2014) and the Data Analysis System at the U.S. National Center for Education Statistics (n.d.). Unlike some of the other systems, the MAS has no administrative restrictions placed on it; the system will be available to anyone with an Internet connection, with no cost or burdensome procedures to get access to the data. In addition, unlike the proprietary system Privacy-Preserving Analytics® (Sparks *et al.*, 2008), the MAS will produce fairly standard types of statistical output, such as tables. The price of this flexibility and usability is that the MAS must have certain restrictions regarding which data are allowed to be output, to ensure that the pledge of confidentiality is upheld.

2. Functionality and Disclosure Rules

Although we envision the inclusion of many surveys in the MAS, current development and testing center on the ACS. Since the ACS is a demographic survey, tables for the ACS tend to be tables of counts, which have been the focus of MAS development so far.

When a table is produced in the MAS, three main operations will be performed dynamically: a determination of whether the table is safe to output to the user, calculation of the estimates in the table and calculation of the variance of each estimate.

Dynamic disclosure avoidance will be the first set of calculations performed. The table will be checked against a set of rules to determine whether it is safe to release. Under current plans, three rules will be used to vet tables of counts, although this may change before the system is released. The mean number of unweighted observations in each interior cell of the table must be at least m, for a fixed value of m. The median number of unweighted observations in each interior cell of the table must be at least n, for a fixed value of n. Among cells with a positive number of observations, the proportion of cells with exactly one unweighted observation must be no greater than p, for a fixed value of p. The values m, n and p are still being considered to determine what values produce a safe set of tables while maximizing the utility of the system. However, we have a working set of values for these parameters, which are the basis of the analyses in Sections 3 and 5, and which are confidential. Further testing may also reveal that these rules are inadequate and need to be modified, replaced or supplemented to protect the data.

Some other disclosure rules being considered include restrictions on the marginal cell counts for multi-dimensional tables and lower bounds on the population of the area for which the table is being made. We will show in Section 4 why these tables with cells of size 1 can be problematic, leading us to try to limit their release.

The current plan is for a table on a composite geography, such as four tracts, to pass only if that table would pass for each of the individual geographies involved. Section 4 explains why this rule appears to be necessary.

The system includes further protections in addition to the disclosure rules implemented at the time of a query. Data available for tabulation in the MAS will be accessible only as a set of recodes, which will be predetermined for each variable. For categorical variables, these recodes may be identical to the original variable or may be a version with multiple categories collapsed into a single category. The latter will particularly happen in cases in which the original variable had at least a moderately large number of categories. Numerical variables will be recoded into pre-defined intervals, and only analyses on these intervals will be available. The effect of recoding will be that numerical variables are treated as de facto categorical variables for the purpose of tabulation in the system. Recodes will generally be determined so that the categories are consistent with data products that are already available, such as previously published tables. This approach to establishing recodes is partly to ensure that the recodes create categories that are salient for users and mesh well with existing estimates, and partly to prevent a malicious user from combining the MAS with another data product to isolate a small number of respondents.

For some variables, the system may provide multiple recodes at varying levels of detail, with a more detailed recode having categories nested within a less detailed recode's

categories. The different recodes will be appropriate for geographies of different sizes; a very detailed recode is unlikely to be provided for a small geography, but a less detailed recode might be. The use of several recodes accommodates multiple groups of users, who might place different relative values on geographical versus topical detail.

If a table does not pass the disclosure rules, a message will inform the user that the table could not be returned. If the table passes, the user will receive the table. The counts produced in the table will be based on the survey weights that accompany the individual observations. As is usual in the Census Bureau's published tables, results based on the unweighted survey data will not be available.

Finally, a measure of variance will be calculated and displayed. Different Census Bureau programs use different measures of variance, and the form of the measure—variance, standard error, margin of error, etc.—will agree with the form that is usually used for that survey. Different surveys also have different methodologies for calculating variance, and the MAS will calculate variance the same way as the survey in question. For the ACS, the variance measure is a margin of error with 90% confidence, which is calculated using replicate weights except in certain special cases.

The MAS will allow variance calculations directly from the microdata. In many cases, this will result in variances that have previously been unavailable, particularly when geographic regions or categories are combined. In some cases, such estimates could be found by adding the component estimates from published tables, but it was impossible to combine the variance estimates properly. The user could only approximate the variance by making an imperfect assumption of no covariance between the estimates. Since the MAS will have access to the underlying microdata, it will be able to calculate variance for these combined values.

A pilot of the MAS based on the ACS five-year microdata file will be released soon. Since the disclosure avoidance methodology is still being developed, the pilot system will be limited in scope to one- or two-way tables with a fixed set of variables. The variables have been selected so that all possible estimates in two-way tabulations are already derivable from existing Census Bureau data products, and thus no table will need to be withheld for disclosure reasons. However, the pilot will demonstrate the capabilities of the system, including margins of error for combined geographies and variable categories.

3. Utility

For the purpose of this paper, the main measure of utility of the system is the frequency with which a requested table is returned. We are considering how to incorporate broader measures of utility, taking into account that some estimates are more useful than others, based on the level of detail and other factors.

To examine how often a requested table is produced, we made one-way tables of 23 variables, two-way tables of all combinations of 16 variables and three-way tables of all combinations of 15 variables. The tables were based on the 2009-2013 ACS five-year microdata. Some of the variables were defined only on certain subpopulations, leading to a smaller collection of data on which to base the tables. We omitted tables that had "structural zeros"—table cells that must have a count of zero because one of the variables is defined only when the other takes on certain values. For example, the ACS variable

measuring whether the respondent has given birth in the past year is collected and recorded only for female respondents age 15 to 50, so a cross-tabulation of this variable with sex would by definition give a value of zero for any cell where the sex is male. We kept in the analysis tables that included cells that seemed implausible but were not structurally impossible (e.g., a person age 16 or older who is currently in kindergarten). This left us with 496 tables.

For each table shell, we ran the table for all 50 states and the District of Columbia. Figure 1 shows the frequency with which each table passed versus the number of cells in the table. The red curve was generated using a normal kernel density smoother on the points in the scatterplot, with standard deviation 3.

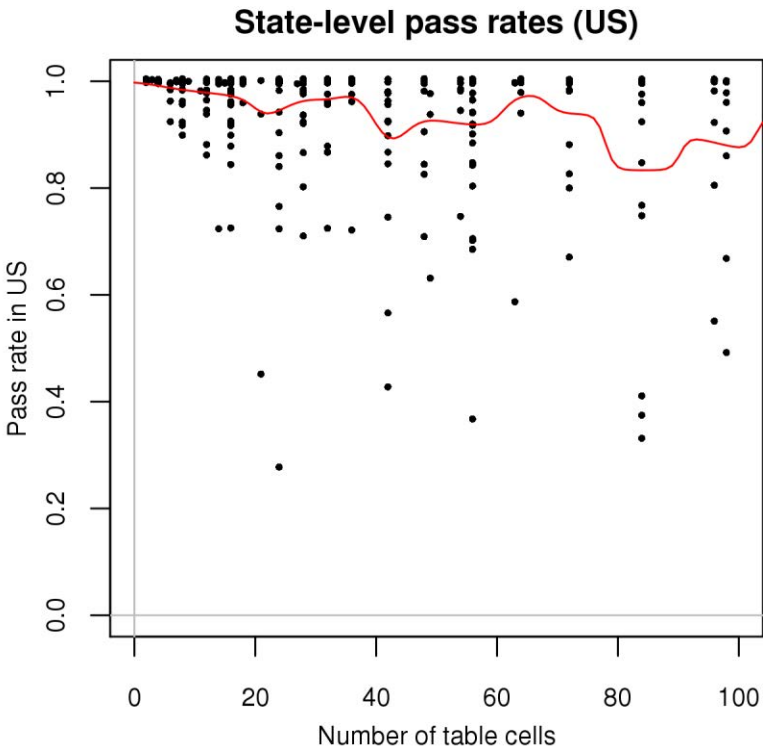

Figure 1: Proportion of states for which a table passes the proposed disclosure rules, plotted against the number of cells in the table.

The plot shows that tables with a relatively small number of cells will generally be returned for most states. As the number of cells increases, the probability of getting the table goes down but still remains fairly high. Several of the 496 tables we tried were considerably larger than 100 cells; they are not shown in this graph, or in the graphs that follow.

Figure 2 on the following page plots the same variables, but this time with each point representing the proportion of counties in which a table passed when attempted over the whole nation. The figure shows that many larger tables will not pass at the county level. Of course, pass rates will vary depending on the size of the county and the distribution of the variables. Because of the median and proportion of 1s rules, tables of variables that are skewed toward only a few categories out of many are less likely to pass the disclosure rules than tables of variables that are more evenly distributed.

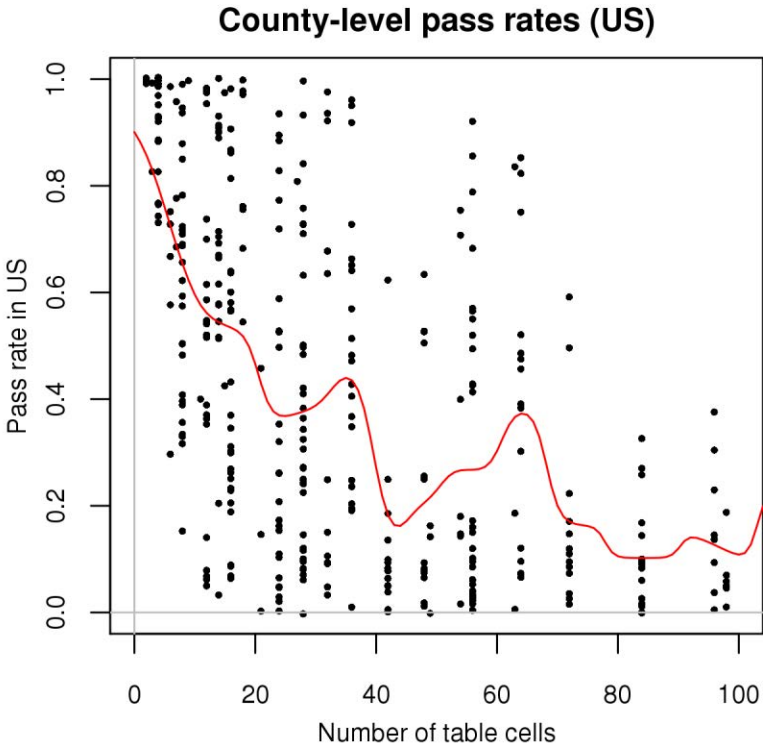

Figure 2: Proportion of counties for which a table passes the proposed disclosure rules, plotted against the number of cells in the table.

Figure 3 on the following page plots the same variables, this time with the pass rates being taken at the tract level. This figure shows that tract-level tables rarely pass the disclosure rules, except when they have very few cells.

The user may also wish to make a table for a composite geography, such as multiple tracts put together. Because a table for such a region passes the disclosure rules only if the corresponding table would pass the rules for each tract individually, it becomes increasingly difficult for a table to pass as the number of tracts increases. To examine this phenomenon further, we considered a standard recode of the age variable with 14 categories and a standard race variable with seven categories. For each variable, we found whether a one-way table of that variable passed for each of the 72,421 tracts in the United States. We also constructed at random 72,421 collections of n tracts, for n=2,3,4,5, such that each collection consisted of contiguous tracts within the same state, and examined whether the tables passed for these collections. The results are in Table 1.

Table 1: Pass rates of age and race variables for collections of one to five contiguous tracts in the same state

Number of tracts	Pass rate – age	Pass rate – race
1	97%	41%
2	94%	20%
3	92%	11%
4	90%	6%
5	88%	4%

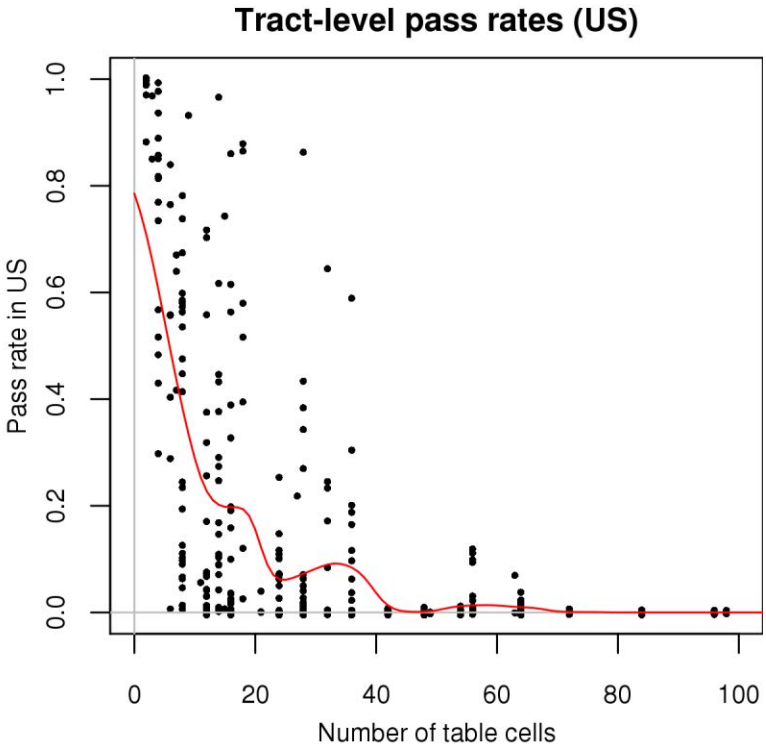

Figure 3: Proportion of counties for which a table passes the proposed disclosure rules, plotted against the number of cells in the table.

If the pass rate for a single tract is not very high, then the pass rate quickly deteriorates as the number of tracts increases, as the failure of even one tract causes the whole collection to fail. The pass rates for two to five tracts are slightly higher for contiguous tracts within the same state than they would be for an equal number of randomly selected tracts, indicating a bit of geographic clustering of passing tracts for a given query. However, this small improvement does not substantially affect the low pass rates. The rule requiring all tracts to pass in order for the composite to pass is critical for disclosure avoidance, but we are considering how to address the problems the rule creates.

We hope to consider other measures of utility beyond simply how often a table is produced or denied to the user, since not all tables or estimates are equally valuable, but the measures we have used here are sufficient to demonstrate that the system as it has been tested so far will not always be as useful as we might hope.

4. Differencing Attacks

The attack of greatest concern is differencing, where the user makes two or more tables with similar data. The user then manually subtracts the numbers in the two tables to arrive at a conclusion about an individual person, household or business.

A differencing attack could be executed as follows. A data intruder makes a cross-tabulation of two variables for a given geographic region, perhaps as small as a census tract, and finds what appears to be a unique observation with respect to these two variables. Suppose, for illustration, that the two variables are sex and a coarse recode of

age with only three categories, with the dataset containing only one woman age 65 or older:

Table 2: Sex by Age for a Hypothetical Dataset

	Age 0-17	Age 18-64	Age 65+
Male	82	68	57
Female	26	28	1

The intruder may now try to recover another variable about the person who is unique. For example, suppose the intruder wishes to recover poverty status, which could easily be determined from Table 3 if the intruder had access to it.

Table 3: Poverty status for females age 65 and older

Not in poverty	In poverty
0	1

The disclosure rules would prevent the system from giving Table 3 to the user. However, the intruder may request a table of poverty status for females in the given tract, which might pass the disclosure rules and be provided by the system. The result of this query is given in Table 4.

Table 4: Poverty status for females

Not in poverty	In poverty
39	16

The intruder may then request Table 5—a table of poverty status for the universe of females under age 65 in the tract:

Table 5: Poverty Status for females under age 65

Not in poverty	In poverty
39	15

If the system returns Tables 4 and 5, the intruder may subtract Table 5 from Table 4, taking an indirect route to deriving Table 3 and thwarting the disclosure rules. With the derived Table 3 comes the conclusion that the woman who is 65 or older is in poverty. A particular problem would arise if the intruder could repeat this approach to determine a large number of variables for the same unique person, thus constructing a large segment of a microdata record. The intruder could then attempt to match these variables to a publicly available microdata file from the Census Bureau or some other source, potentially identifying the unique respondent.

Of course, if Table 4 or Table 5 does not pass the disclosure rules, then the attempt to reveal the poverty status of the unique person would be unsuccessful, although the intruder could try to reveal other variables instead.

Most differencing attacks are not as easy to execute as the example given here. Even if a table analogous to Table 2 has a unique observation, the user will usually not be able to tell immediately by looking at the table, because almost all observations have a survey weight greater than 1.

An intruder could also undertake a differencing attack in a case where a table cell is small but is not a unique. The simplest attack of this sort occurs in a case where a cell includes two observations, one of whom is the intruder and one of whom is the target. The intruder already knows his own information and can thus manually subtract himself from any table to create a de facto cell of size 1. Again, survey weighting makes this attack not entirely straightforward, as does the fact that the intruder must be well-situated to execute the attack: in the same cell as the target and with no other sampled units in that cell.

The idea of a differencing attack also shows why we have the rule that a table passes for a composite geography only if the table would pass for each of the component geographies. Suppose an intruder is interested in a table that the system considers too risky to release for the intruder's preferred geography but that the system would release the equivalent table for some other geography. The intruder could request the relevant table for a composite of geographies for which the table passes and then request the table again for this same composite with the risky geography added in. Subtraction of these two tables reveals the risky table. The rule about composite geographies prevents this approach from being successful.

5. Study of the Prevalence of a Differencing Attack

We performed the differencing attack stated above on uniques with respect to the same two variables as above: sex and age. In this case, we used a standard recode for age that is used in the American FactFinder and elsewhere. This recode has 14 categories: 0-4, 5-9, 10-14, 15-17, 18-19, 20-24, 25-29, 30-34, 35-44, 45-54, 55-64, 65-74, 75-84 and 85 or older. To execute the attack, we found unique observations in the two-way table of sex by age. We then used the method described in the poverty example in conjunction with 23 variables in the ACS five-year dataset, using the first two (sex and age) to attempt to recover the 21 remaining variables for each of the uniques. Overall, there were 12,420 uniques in tracts where the initial two-way table passed, roughly one for every six tracts. Of these, 3,228 (approximately one per 22 tracts) had at least one of the variables recoverable.

The largest concern with a differencing attack of this sort is that the intruder might be able to construct a large portion of a confidential microdata record by repeating the attack on the same unique repeatedly, and then use the partial record for identification of a respondent. That problem did not occur in the simulated attack, as even in the 3,228 cases where a variable could be recovered, the mean number of variables recovered was 1.24. Only 90 of the uniques allowed recovery of three or four variables, with none resulting in recovery of more than four variables.

A more problematic approach is to use the same attack to leverage uniques with respect to sex and a variable called MSP, for "married, spouse present." This variable has six categories: now married, spouse present; now married, spouse absent; widowed; divorced; separated; and never married. We simulated this attack like the previous one

and tried to recover the remaining 21 variables. We found that there were 24,441 tract-level uniques with respect to these variables in tables where the initial two-way table passed—roughly one unique for every three tracts. Of these, at least one variable could be recovered for 14,148 tracts. When a variable could be recovered, there were never more than five recoverable variables, and the mean number of variables recoverable was 1.54.

6. Mitigation

Several forms of mitigation could be used to make a differencing attack less likely.

6.1 Clarifying the Disclosure Rules

In cases where a universe definition has a non-geographic aspect, the mean, median and proportion of 1s rules could be implemented in two different ways. In the case of Tables 4 and 5, the table that the user wants the system to display consists of two cells—a cell for people in the universe who are not in poverty and a cell for people in the universe who are in poverty. However, it could be argued that in this case, the universe is really part of the table being requested. By this interpretation, the user is requesting an implicit tabulation of age by sex by poverty status and electing to see only part of the resulting table. Generally, the implicit table will be less likely to pass the disclosure rules than the table based on the stated universe. We found that the attack was more frequent and more severe if the first set of universe rules was used. In the attacks that we tried (sex by age and sex by MSP), at least one of the 21 variables was recoverable for every unique, and in very unusual cases, as many as 17 of the 21 variables could be recovered. As a result, the implementation of the rules based solely on the cells in the table to be displayed—and not on the implicit table that includes any universe variables—is unworkable as a disclosure prevention method, and thus all of our analysis is based on running the disclosure rules on the implicit table.

6.2 Adjusting Recodes

In the age by sex example, three age categories accounted for the vast majority—87%—of the uniques. These were 15-17, 18-19 and 85+. The first two create lots of uniques because they are narrower than most other age categories, which have a width of five or 10 years.

One possible solution is to combine these categories with others to reduce the frequency of uniques. In this case, we can eliminate the two most problematic categories by combining the 18-19 category with the adjacent 20-24 category and combining the 85+ category with the adjacent 75-84 category.

For the attack on sex by age, this solution reduces the number of uniques from 12,420 (one per six tracts) to 2,742 (one per 26 tracts), a 78% decrease. The number of uniques where at least one variable is recoverable decreases from 3,228 (one per 22 tracts) to 672 (one per 108 tracts), a 79% decrease. The severity of the attack does not change—when at least one variable can be recovered, the average number of variables recoverable is 1.24 either way.

A side benefit of combining categories is that it may lead to fewer small cells in the table, which makes tables more likely to pass, thus potentially increasing the utility of the system for some users. However, this solution is not ideal, as it may reduce detail in a way that limits utility to the user. For example, if the 18-19 and 20-24 age categories are

combined and a particular user found the cutoff at age 20 salient, then the user may find the data to be of much diminished utility.

Another difficulty with this solution comes on the production end, as combining categories for the explicit purpose of eliminating uniques requires determining which categories contain uniques and how much the situation is improved by combining particular categories. The combination must be done separately for each set of variables, creating inefficiencies. As a result, although combination of categories may alleviate the problem, it cannot be the cornerstone of the MAS's disclosure strategy.

6.3 Data Perturbation

Data in the MAS may be further protected by modifying the system's source microdata. In the case of the ACS, the data for the MAS and all other Census Bureau data products are subjected to data swapping; see Lauger *et al.* (2014) for information on swapping at the Census Bureau, or Duncan *et al.* (2011) as a more general swapping reference. The swapping algorithm used for the ACS is targeted; households deemed to be at more risk of being identified are more likely to be swapped. Since records from group quarters are not amenable to swapping, ACS data products use partially synthetic data for this portion of the dataset (Hawala, 2008). As a result of the swapping and synthetic data, even if an intruder were able to reveal a unique record in the MAS through a differencing attack or some other means, the intruder could not be certain that the identified record corresponded to the respondent with whom it appeared to be associated.

6.4 Technical Approaches

Although there could be concerns about someone using the system as planned to come to a disclosure, the more significant problem is an intruder automating an attack by executing a large number of queries on the system. This way, even if potential successful attacks are rare, there will be a substantial possibility of the intruder's finding them. We are considering having a "throttling" feature in the system, which would prevent a single user from submitting a large number of queries in this fashion. Ideally, the threshold would be high enough so as not to frustrate legitimate users of the data.

6.5 "Other" category

Some variables—such as occupation, industry or language spoken at home—have categories that are heavily geographically clustered. For example, a language that does not have a very large number of speakers nationwide may have a substantial number of speakers on an American Indian reservation or in a community with lots of immigrants from a particular country. For example, if a user asks for a table of language spoken under the disclosure rules as they now stand, and the user hopes to find information on speakers of Mandarin in a tract where Mandarin is common, two possible problems might arise:

1) The user may have access to only a very coarse categorization of languages, where Mandarin may be grouped together with all other Asian languages, a classification that may not meet the user's needs.

2) The system may offer a more detailed categorization of languages, such that Mandarin is separated from all other languages. However, a categorization with this level of detail is likely to have a number of other detailed language categories in addition to Mandarin. Even though, for a particular tract, statistics

on speakers of Mandarin may be safe to release, the table will likely be withheld because of all of the other language categories that have minimal, if any, representation in that tract.

Since the number of people who speak Mandarin at home is a relatively small proportion of the population in the country as a whole, it would often be problematic to allow data on this subpopulation, but for this particular community, it is reasonable for a user to want data on this group. One approach we are exploring is to display data for common languages in the requested geographic region, and then combine the remaining languages under the heading "Other." This approach would allow data on Mandarin speakers in the example given here, for geographic areas where it is not problematic to give these data but where giving data about speakers of all languages might be a disclosure problem. Of course, this example extends to other cases of clustered languages, or of other geographically clustered variable values. We have not made a final decision on whether or how to include this feature, as we are still exploring whether it is safe to do so.

6.6 Modification of geographic regions

Another possibility for remedying disclosure problems is to change what geographic regions are available in the system, which could be done most simply by requiring larger geographies than what we have been planning. Although development of the system for the ACS has been done in the hope that we could provide data at the tract level, perhaps only a very limited range of basic queries will be made available for individual tracts, as we have found that once the user goes beyond these, the queries are very unlikely to pass the disclosure rules. One possibility is to group tracts to create a "super-tract" with enough sample size that data are likely to be safe to release, but still to have this geographic region be at a lower level than the Public Use Microdata Areas (PUMAs) that are the smallest region available in the ACS Public Use Microdata Sample (PUMS). (Tracts generally include about 4,000 people, although their sizes vary, while PUMAs are required to have a population of at least 100,000.) Another possibility is to use counties as the smallest allowable level of geography, except for larger counties that contain multiple PUMAs, in which case the PUMA would be the smallest level. A user who wants data at a relatively low level of geography would then have to create a geographic region from these building blocks rather than having access to the individual tracts.

7. Research in Progress on Balancing Risk and Utility

We have begun to do research on the relationship between the rate at which tables pass the disclosure rules and the rate at which uniques show up in the resulting tables as the parameters of the disclosure rules are varied. In a system where disclosure protection occurs primarily through table denial and where exposure of unique observations is a key step for an intruder, we would hope that the disclosure rules could be set in such a way that the proportion of tables produced is as large as possible, while the proportion of uniques exposed is as small as possible.

This balance can be examined by varying the parameters of the disclosure rules and comparing the number of uniques with the proportion of tables provided to the user. We could imagine grouping tables, distinguishing those that are almost always returned regardless of the parameters from those that are only returned when the parameters are set to allow almost any table, with various gradations in between. In the worst plausible case, uniques would be roughly evenly distributed among the different groups of tables. A

graph with the proportion of uniques revealed on the y-axis and the proportion of tables produced on the x-axis would then resemble a 45-degree line. However, our hope would be that under stricter disclosure rules where few tables are returned, those tables that are available would be those with relatively few uniques, while tables with more uniques would only be produced under rules that are looser. In this case, the corresponding curve would be convex up, below the 45-degree line. We could imagine a curve that is convex down, above the 45-degree line, but this situation should not happen if the disclosure rules are chosen well, as it would mean that if only relatively few tables are produced, they are more risky than if tables to be produced were selected at random.

Not every unique is a disclosure risk, nor is every disclosure risk necessarily associated with a unique; in Section 5, we saw that the attack we tried often did not reveal any additional information about a unique. However, the number of uniques provided can be a proxy for relative disclosure risk.

Results so far indicate that for some variables and geographic levels, something approximating the "worst case" described above (a 45-degree line) occurs, while in other cases, a graph of what proportion of uniques are revealed versus what proportion of tables are provided gives the convex up shape that we would hope for. The general pattern seems to be that if there are fewer uniques to be found, then the desired relationship is likely to have a convex up shape, while if there are more uniques, the relationship is likely to be closer to a 45-degree line. These preliminary findings are consistent with the hypothesis that if there are few uniques, the types of disclosure rules we have considered will release the tables containing them only if the disclosure rules are very lenient. On the other hand, it appears that if there are more uniques, the disclosure rules will release them with roughly constant frequency as the disclosure rules get more lenient. Exploration of more graphs will be necessary to determine whether this pattern holds more broadly, however.

We plan to explore this line of research more in a future paper.

8. Future Research

Research so far on disclosure in the MAS has revealed that the set of disclosure rules we have considered creates a very high rate of table denial, even if the parameters are set to make it somewhat likely that a unique will be revealed. As a result, it may be necessary to use an alternative disclosure method that focuses more on creating a dataset that is inherently protected, thus shifting some of the emphasis away from table requests being accepted or denied.

One possible approach is to use synthetic data for households as well as group quarters, resulting in a dataset that maintains the essential properties of the ACS data. The MAS would then query these data to create tables, rather than querying the data directly collected from respondents. Some disclosure rules leading to denial of tables would probably be kept in place, as the dataset might be created using a method that retains certain properties of individual respondents. Creating the dataset may be challenging if low levels of geography are involved, but even data at the tract level are more manageable to synthesize than data at smaller levels of geography.

Once we have established ACS table functionality for the MAS, we plan to extend the capabilities of the system. We would like to include other datasets, such as economic surveys and the decennial census, each of which will require their own forms of disclosure protection. We also hope to extend the analyses available in the system beyond tables, to features such as summary statistics, regression and modeling.

Acknowledgements

Laura McKenna has guided the MAS project from the beginning and provided expertise on established disclosure avoidance methods. The DataWeb and Applications Staff at the Census Bureau—particularly William Hazard, William Rankin, Romir Campbell and Tiffany Julian—have managed the technical side of the project, including implementing table generation, disclosure protection and the interface for the MAS. Thanks also to Michael Ikeda for reviewing this paper.

References

Australian Bureau of Statistics. Remote Access Data Laboratory (RADL). http://abs.gov.au/websitedbs/D3310114.nsf/home/CURF:+Remote+Access+Data+Laboratory+(RADL), 2014. Accessed September 1, 2015.

Duncan, G.T., Elliot, M. and Salazar Gonzalez, J. J. *Statistical Confidentiality: Principles and Practice*. New York: Springer-Verlag, 2011.

Hawala, S. "Producing Partially Synthetic Data to Avoid Disclosure." Proceedings of the Section on Government Statistics, American Statistical Association, 2008, 1345-1350. https://www.amstat.org/sections/srms/Proceedings/y2008/Files/301018.pdf. Accessed September 9, 2015.

Lauger, A., Wisniewski, B., and McKenna, L. *Disclosure Avoidance Techniques at the US Census Bureau: Current Practices and Research*. Technical Report 2014-02, Center for Disclosure Avoidance Research, US Census Bureau, 2014. http://www.census.gov/srd/CDAR/cdar2014-02_Discl_Avoid_Techniques.pdf. Accessed September 1, 2015.

LIS Cross-National Data Center in Luxembourg. LISSY. http://www.lisdatacenter.org/data-access/lissy/, 2011. Accessed September 1, 2015.

Meyer, P. S. "'Virtual Data Access' for Statistical and Research Purposes." Presentation to the FCSM Statistical Policy Seminar, December 15, 2014. http://www.copafs.org/UserFiles/file/2014fcsm/06_PeterMeyerFCSM%20120314.pdf. Accessed September 9, 2015.

National Center for Education Statistics. DAS Website. http://nces.ed.gov/das/, n.d. Accessed September 1, 2015.

Simard, M. Real Time Remote Access at Statistics Canada: Development, Challenges, and Issues. *Proceedings of the 58th World Statistics Congress of the International Statistical Institute*, 2011.

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