ICPSR
A PARTNER IN SOCIAL SCIENCE RESEARCH

INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH
UNIVERSITY OF MICHIGAN

Working Paper

Disclosure Risk Components of Contextualized Microdata: Identifying Unique Geographic Units and the Implications for Pinpointing Survey Respondents

Kristine M. Witkowski
Inter-university Consortium for Political and Social Research,
University of Michigan

ICPSR Working Paper Series
Working Paper No. 3

June 2008

Disclosure Risk Components of Contextualized Microdata: Identifying Unique Geographic Units and the Implications for Pinpointing Survey Respondents

Kristine M. Witkowski

Inter-university Consortium for Political and Social Research, University of Michigan

To safely respond to increased demand for microdata that contain contextual information, producers ought to consider how this data may be used to identify the location of survey respondents. This study informs the design of these datafiles with its hierarchical matching algorithm and discussion of associated methodological concerns. Compiling nearly 15,000 test datasets composed of person-records, I assess three determinants of "locational" risk, that of identifying the location of survey respondents whose contextual characteristics: (1) are rarely found among the total population of geographic units; (2) are rarely found within a survey; and (3) pose no disclosure risk given the protection offered by the area's dense population. Using the "datafile" as my unit of analysis, the proportion of survey respondents whose locations are easily-reidentified as the outcome of interest, and indicators of different components of this risk, I detail the complexity of reidentification patterns that emerge when constructing public-use files that provide contextual data.

Key Words: confidentiality, dissemination

Acknowledgements: Research support from the National Institute of Child Health and Human Development (NICHD), Grant 5 P01 HD045753 as a supplement to the project Human Subject Protection and Disclosure Risk Analysis, is gratefully acknowledged. Special thanks are also given to Myron Gutmann for his thoughtful comments.

Contact Information: Please address all correspondence to Kristine M. Witkowski, Inter-University Consortium for Political and Social Research (ICSPR), Institute for Social Research (ISR), The University of Michigan, P.O. Box 1248, Ann Arbor, Michigan 48106-1248; Email: kwitkow@umich.edu; Telephone: 734-763-7102.

Disclosure Risk Components of Contextualized Microdata: Identifying Unique Geographic Units and the Implications for Pinpointing Survey Respondents

Kristine M. Witkowski

Inter-university Consortium for Political and Social Research, University of Michigan

To safely respond to increased demand for microdata that contain contextual information, producers ought to consider how this data may be used to identify the location of survey respondents. This study informs the design of these datafiles with its hierarchical matching algorithm and discussion of associated methodological concerns. Compiling nearly 15,000 test datasets composed of person-records, I assess three determinants of "locational" risk, that of identifying the location of survey respondents whose contextual characteristics: (1) are rarely found among the total population of geographic units; (2) are rarely found within a survey; and (3) pose no disclosure risk given the protection offered by the area's dense population. Using the "datafile" as my unit of analysis, the proportion of survey respondents whose locations are easily-reidentified as the outcome of interest, and indicators of different components of this risk, I detail the complexity of reidentification patterns that emerge when constructing public-use files that provide contextual data.

Key Words: confidentiality, dissemination

Acknowledgements: Research support from the National Institute of Child Health and Human Development (NICHD), Grant 5 P01 HD045753 as a supplement to the project Human Subject Protection and Disclosure Risk Analysis, is gratefully acknowledged. Special thanks are also given to Myron Gutmann for his thoughtful comments.

Contact Information: Please address all correspondence to Kristine M. Witkowski, Inter-University Consortium for Political and Social Research (ICSPR), Institute for Social Research (ISR), The University of Michigan, P.O. Box 1248, Ann Arbor, Michigan 48106-1248; Email: kwitkow@umich.edu; Telephone: 734-763-7102.

Draft: Please do not cite without author's permission.

1. Introduction

Many problems in contemporary social science lend themselves to an analysis in which the individuals under study are placed in their context, especially a context that can be defined spatially, as street, block, town, county, or some other spatial unit. Data producers have found two ways of providing this information, either identifying the spatial unit (so that the data user can link the appropriate contextual data herself), or merging the contextual data, effectively adding the characteristics of the spatial unit in which the subject lives. In this second case, the record for a given individual includes that person's characteristics (e.g., age of respondent) as well as those where they live (e.g., proportion of population in respondent's neighborhood that is poor).

One reason for providing the contextual data themselves, rather than the identity of the spatial unit, is that doing so makes it more difficult to identify the spatial unit in which the survey respondent lives (Armstrong, Rushton, and Zimmerman 1999). However, it is possible that the contextual data themselves constitute

enough information to be a geographical unique. If that's the case – for example if the combination of contextual information about a given spatial unit is rare among spatial units of that type – then identification is more likely, rather than less (Saalfeld, et. al, 1992). Care must then be taken to modify these data to maintain their confidentiality and their statistical properties, while at the same time ensuring that the data have the maximum analytic value for the broadest user group.

Two recent studies followed this practice of adding contextual data to their analytical files. In producing their public-use files for their Residential Energy Consumption Survey, the Energy Information Administration perturbed temperature data to mask the location of weather stations (Subcommittee on Disclosure Limitation Methodology 2005;http://www.eia.doe.gov/emeu/recs/recs2001/codebook82001.txt). And in a study of discrepancies between official votes and exit polls in the 2004 presidential election, official tallies of the proportion of Kerry votes were blurred for a sample of Ohio precincts; thereby concealing the identity of these controversial voter locations (Kyle et al 2007). Although they address confidentiality issues stemming from contextual data, these studies do not detail the likelihood of reidentifying these locations and associated determinants.

In earlier work (Witkowski 2007), I have conducted reidentification experiments to assess how easy it is to pinpoint locations within the total population of counties, tracts, and blockgroups, when only aggregate information is available. That work provides estimates of the amount of identification risk associated with the spatial scale of contextual measures; the identification of division, state, and MSA-status; and the number and coarseness of contextual variables provided in a dataset. This earlier work is limited to the risk of identifying geographic units. In other words, if we know the percent minority and the percent employed in the civilian labor market (or other characteristics); can we know which county we're talking about?

Knowing the county (or census tract, or blockgroup) doesn't mean that we can identify individuals. It's just the starting point. The next question is to ask about the implications of disclosure risk of geographic units for survey respondents. How does the amount of risk associated with contextual design elements change when we consider the dispersion of people across space? To answer these questions, we must then consider reidentification mechanisms that translate risk from geographic units to individuals.

Because microdata files typically consist of both individual and contextual measures, a full assessment of risk requires a nested approach to its disclosure analyses that incorporates identifying characteristics of both survey respondents and their locations. This study helps lay the groundwork for such an evaluation by: (1) constructing a hierarchical matching algorithm; (2) discussing methodological concerns that arise from this type of assessment; and (3) providing estimates for three key components of locational risk derived from persons nested within geographic contexts.

With an analytical approach bridging two levels, my current study informs the design of public-use datafiles composed of person-records containing contextual measures at three spatial scales: (1) census blockgroups, (2) census tracts, and (3) counties. Compiling nearly 15,000 test datafiles, I assess three determinants of "locational" risk associated with contextualized microdata, that of: (1) identifying survey respondents whose contextual characteristics are rarely found among the total population of geographic units (i.e., contextual population uniques); (2) identifying survey respondents whose contextual characteristics are rarely found within a survey (i.e., contextual sample uniques); and (3) identifying survey respondents whose unique contextual characteristics pose no disclosure risk given the protection offered by their location's dense population.

In doing so, I provide estimates for these risk components and illustrate how they vary across the spatial scale of contextual information while controlling for other pertinent design elements, such as: (1) the identification of division, state, and MSA-status; and (2) the number and coarseness of contextual variables provided in a dataset. Using the "datafile" as my unit of analysis, the proportion of survey respondents whose locations are easily-reidentified as the outcome of interest, and indicators of different

components of this risk, I detail the complexity of reidentification patterns that emerge when constructing public-use files that provide contextual data.

2. Empirical Approach to Disclosure Analyses of Contextual Data

In this section, I outline the analytical steps involved in evaluating disclosure risk for contextual data (Armstrong, Rushton, and Zimmerman 1999; De Waal and Willenborg 1995, 1996; Interagency Confidentiality and Data Access Group 1999; Subcommittee on Disclosure Limitation Methodology 2005; United States General Accounting Office 2001; Zayatz 2005).

Widely-available summary files of census data identify all counties, tracts, and blockgroups in the United States and provide measures describing the characteristics of the population located within these geographies. Trying to safely meet user demands for geographically-rich information, a producer can attach these contextual data to their survey respondents' locations, without directly identifying the geography. But before these microdata files can be released publicly, the producer must assess the likelihood that a survey respondent's location can be correctly reidentified by an intruder using this contextual data.

An intruder may identify geographies by conducting an experiment that matches contextual information provided in the survey's public-use file with data available from the original tabulation files for the full population of geographies. Searching within the survey file, the intruder identifies sampled locations that share a specific set of contextual characteristics. Using the same contextual indicators, the intruder then identifies locations in the population file that have the same characteristics and compiles identifying information (i.e., county name, tract, and blockgroup identifiers) for these "population matches".

A sampled location having a single match, defined as a "unique", is unequivocally reidentified when the external database represents the full population. A location with a small number of matches, defined as a "rarity", also faces a significant risk of being reidentified. Conversely, a location is said to be adequately obscured only when there is a sufficiently large number of population matches. In turn, the producer must decide upon a match-threshold that defines the upper bound of risk. If the number of matches falls below this threshold, then risk of reidentification is considered intolerable and therefore the contextual data is not safe for release. The producer considers information sensitivity and intruder behavior when defining this match-threshold, incorporating statistical inference arguments to justify their selection. After defining what constitutes anonymity for geographic units, the producer then assesses the amount of risk associated with a public-use file's contextual information by calculating the proportion of locations that are easily reidentified.

Using this measure to gauge changes in risk, the producer can then modify the composition of their public-use file to meet their goal of maximizing the utility of contextual data while minimizing the chances that geographic units are reidentified. Data utility, as well as the ability to reidentify locations, is enhanced by releasing geographic identifiers and a large number of contextual variables. To effectively design their public-use dataset, the producer needs to set priorities regarding (1) the release of geographic identifiers, (2) the scope of geographic identifiers, (3) the scale of contextual variables, and (4) the number of contextual variables. The producer can then better select disclosure limitations methods (DLMs) that offset the risk associated with these contextual data.

To reduce disclosure risk in public-use microdata files, agencies often only apply nonperturbative methods in order to maintain the statistical properties of the original data, thus maximizing its utility for widely disparate and largely unknown applications (Subcommittee on Disclosure Limitation Methodology 2005; Zayatz 2005). Consequently "global recoding" and "local suppression" are important techniques to be considered for statistical disclosure control (De Waal and Willenborg 1995, 1996). Aggregating continuous measures into various levels of coarseness decreases the likelihood that locations are reidentified (i.e., global recoding). But for geographic units that are still easily pinpointed, their contextual characteristics are not to be released on a microdata file (i.e., local suppression); where the proportion of

sampled locations with suppressed contextual data is represented by the aggregate measure of disclosure risk. Because fewer measurement categories and high suppression rates heighten the amount of information-loss, a producer needs to consider how data utility varies with these methods.

After assessing the likelihood of pinpointing locations among geographic-unit populations, the producer needs to assess how this "locational" risk transfers to their survey's respondents. Respondent identity is believed to be protected when there are an adequate number of persons – in the survey – sharing the same contextual characteristics (i.e., sample unique; k-anonymity; Sweeney 2002; Subcommittee on Disclosure Limitation Methodology 2005). Like finding a needle in a haystack, living in a highly populated area offers another safeguard for respondents, regardless of their location's unique contextual characteristics (Subcommittee on Disclosure Limitation Methodology 2005). Given these reidentification mechanisms, the producer must then establish how many sampled individuals – sharing the same contextual characteristics – are needed to assure anonymity. Furthermore they must decide upon the minimal population size of a geographic unit that adequately obscures the identities of its residents, contemplating how the utility of data may be affected by the disproportionate suppression of contextual information for low-density areas.

Finally, the producer must assess the number of persons sharing the same identifying personal characteristics for individuals within the population, other external datasets, and within the survey dataset following similar steps used in estimating the disclosure risk of geographic units. In doing so, they need to define what constitutes anonymity for individuals nested within contexts as dictated by their personal characteristics. The risk associated with the personal characteristics of survey respondents is heightened under a hierarchical framework of reidentification. When contextual data are not included on a file, the search for person-level matches is open to the full population. However risk would generally increase when we limit the matching process to persons living in similar contextual environments, especially if these contexts are fairly uncommon.

From this empirical approach, I develop a hierarchical matching algorithm composed of different components of risk stemming from the contextual and personal characteristics of survey respondents. In the following section, I describe my measures of these risk components and discuss the methodological issues that arise when implementing this type of disclosure analysis. For the rest of the paper, I conduct an experimental study that: (1) utilizes a limited version of this algorithm and (2) systematically varies the characteristics of contextual data that are associated with a single synthetic sample of persons and their locations. Assessing the marginal contributions of three components of locational disclosure risk, I illustrate how the translation of locational risk to survey respondents influences the design of contextualized datasets.

3. Assumptions and Hypotheses

3. a. Hierarchical Matching Algorithm

Indicative of the risk associated with a public-use datafile, "Aggregate Disclosure Risk" is the fraction of person-records that an intruder can confidently reidentify, calculated as the proportion of survey respondents who have unique characteristics (i.e., AR). For microdata that characterize individuals and their geographic contexts, estimating the likelihood of reidentifying survey respondents (i.e., $R_{I, J}$) depends on studying two populations: persons and locations. Presented in Equations 1 to 5, I develop a hierarchical matching algorithm that considers the geographic distribution of survey respondents as well as their personal and contextual characteristics, formulating five components of risk. The hierarchical assessment of risk is a recursive process in that disclosure analysis should be performed in four stages, where each step builds upon and reciprocates with another.

Setting the stage, the number of persons who share the same personal (i) and contextual (j) characteristics is a function of the dispersion of persons across space (i.e., $Y_{I, J \mid i, j}$), as defined by the

survey's sampling methodology (T). Identifying geography- and person- matches within the population and sample also depends on four characteristics of a dataset: the spatial scale of contexts (S), geographic identifiers (G), the number of identifying personal and contextual keys (K), and the coarseness of their measures (M).[1] It is the interplay between these distributions – persons nested within geography and variation in measurement – that determine the likelihood that a survey respondent is reidentified and, ultimately, the amount of risk associated with a dataset's identifying information (i.e., $Z_{i,j}$).

Given these distributions and unperturbed contextual data provided from census tabulations, the first stage in the hierarchical assessment of risk requires that the data producer appraise how easy it is to reidentify a geographic unit given perfectly accurate information about its characteristics (Lambert 1993; Duncan and Lambert 1989). Using simple combinations of key variables, the probability of a geographic unit being reidentified depends on the number of its matches found in the populations of counties, tracts, and blockgroups; where "matches" are those units sharing the same set of characteristics (i.e., A_j). A set of geographic units – having the same characteristics – are considered anonymous when there are at least T_1 in the set, where T_1 represents a selected "geography population unique" threshold.

After distinguishing sets of geographic units resembling one another (i.e., contextual matchsets), it is then necessary to identify individual-level characteristics that are rarely found in the general population for each contextual matchset. Because an intruder may know about a person's participation in a survey, one must also identify respondents having unique characteristics (Subcommittee on Disclosure Limitation Methodology 2005; Sweeney 2002). So for the second stage of the analysis, the producer assesses the number of survey respondents who share the same identifying personal and contextual characteristics within the population (i.e., $B_{i,j}$) and the sample (i.e., $C_{i,j}$). An individual is considered anonymous when at least T_2 people in the population and T_3 survey respondents have the same characteristics, where T_2 and T_3 represent selected "person population unique" and "person sample unique" thresholds. Estimating the joint probability that a respondent has individual and contextual variables that do not uniquely identify them within the population and the survey, the search for person-matches is constrained to people sharing common contexts. The risk associated with personal characteristics generally increases when we limit the search for person-matches to specific contexts (i.e., $\Pr[B_{i,j}] \leq \Pr[B_i]$ and $\Pr[C_{i,j}] \leq [C_i]$).

Utilizing data constructed from the above evaluations in a third set of analyses, the producer considers another piece of identifying information for geographic units that is based on the nested nature of a respondent's contextual and personal characteristics. An approximated set of contextual measures can be derived from the personal characteristics of survey respondents who share the same contexts. For instance, a new "age" composition measure characterizing an unidentified context is derived from the proportion of respondents living in "Context A" who are age 65 and over. "Context A" is described as a set of contexts that have 80 to 85 percent of their population living in poverty (as identified by census tabulations). Let's take as an example the 20 respondents living in 3 sampled tracts having 80 to 85% of their population in poverty, where 3 of these persons are age 65 and over. The intruder can then limit his search within an external database to geographic units sharing this contextual trait (i.e., tracts with 80 to 85% persons in poverty, "Context A"), identifying geographic units who resemble each other in terms of the new composition measure (i.e., "Context A" tracts with 15% of their population age 65 and over). Research by Ragunathan, et al. (2007; need citation) have indicated that this type of disclosure (i.e., $D_{i,j}$) poses significant risk. A set of geographic units – having the same approximate contextual characteristics – are considered anonymous when there are at least T_4 in the set, where T_4 represents a selected "approximate population unique" threshold.

And lastly, the producer can consider the protective roles of a geographic unit's population size (i.e., $E_{i,j|J}$). Even when a location may be pinpointed within the population using extant contextual measures (i.e., A_j) and approximated contextual measures (i.e., $D_{i,j}$), the ability to identify a specific respondent is negligible when she lives in a highly-populated area. However this protection is only offered to respondents assured k-anonymity (i.e., $C_{i,j}$). An individual is considered anonymous when at least T_3 survey respondents have the same identifying personal and contextual characteristics, and when they live

in an area with a population of T_5 or more (where T_5 represents a selected "geographic unit population-size" threshold).

$$AR = \Sigma (Pr [R_{I, J}]) \quad (1)$$

$$Pr [R_{I, J}] = 1 - (Pr [Y_{I, J | i, j}] \times Pr [Z_{i, j}]) \quad (2)$$

$$Pr [Z_{i, j}] = Pr [(A_j \cap B_{i, j} \cap C_{i, j} \cap D_{i, j}) \cup (C_{i, j} \cap E_{i, j | J})] \quad (3)$$

$$Pr [Y_{I, J | i, j}] = f (T) \quad (4)$$

$$Pr [Z_{i, j}] = f (T, S, G, K, M) \quad (5)$$

Where:

AR = Proportion of survey respondents who are easily reidentified

$R_{I, J}$ = Survey respondent (I) in sampled geographic unit (J) is easily reidentified, given identifying "personal" variables (i), "contextual" variables (j), and "approximate contextual" variables (i, j)

$Y_{I, J | i, j}$ = Survey respondent (I) in sampled geographic unit (J) has identifying "personal" variables (i), "contextual" variables (j), and "approximate contextual" variables (i, j)

$Z_{i, j}$ = Identifying "personal" variables (i), "contextual" variables (j), and "approximate contextual" variables (i, j) are safe for release

A_j = Combination of values of "contextual" variables (j) is considered safe if combination occurs at least T_1 times among geographic unit population

$B_{i, j}$ = Combination of values of identifying "personal" variables (i) and "contextual" variables (j) are considered safe if combination occurs at least T_2 times among person population

$C_{i, j}$ = Combination of values of identifying "personal" variables (i) and "contextual" variables (j) are considered safe if combination occurs at least T_3 times among survey respondents

$D_{i, j}$ = Combination of values of identifying "personal" variables (i) and "contextual" variables (j) are considered safe if combination of values of "approximate contextual" variables (i, j) occurs at least T_4 times among geographic unit population ["approximate contextual" variables (i, j) derived from identifying "personal" variables (i) among survey respondents with "contextual" variables (j)]

$E_{i, j | J}$ = Combination of values of identifying "personal" variables (i) and "contextual" variables (j) are considered safe when the population size of sampled geographic unit (J) is T_5 or more persons

T = Sampling methodology

S = Spatial scale of contextual data

G = Identified geography

K = Number of identifying "personal" or "contextual" keys

M = Masking technique

Assumption-1: T_1 = "Geography Population Unique" Threshold
Assumption-2: T_2 = "Person Population Unique" Threshold
Assumption-3: T_3 = "Person Sample Unique" Threshold
Assumption-4: T_4 = "Approximate Population Unique" Threshold
Assumption-5: T_5 = "Geographic Unit Population-Size" Threshold

The contribution of each component to aggregate risk depends on another, indicated by the covariation between (A_j and $B_{i,j}$), ($B_{i,j}$ and $C_{i,j}$), ($C_{i,j}$ and $D_{i,j}$), and ($C_{i,j}$ and $E_{i,j|J}$). Summarizing these relationships, the size of contextual matchsets (i.e., the number of geographic units resembling each other) determines the likelihood of identifying persons, nested therein, who have similar individual characteristics. If a set of personal and contextual characteristics are found to be unique within the population, then they are also unique within a survey's sample. Large numbers of survey respondents sharing common contexts enhances the estimation of approximate contextual variables, with small contextual matchsets narrowing the search for uniques within the geographic-unit population. Finally the protection offered by dense populations can be offered to more survey respondents when more sets of personal and contextual characteristics are rare within the general population, but are commonly found within the sample.

3. b. Current Study's Matching Algorithm: Modifications and Implications

Because of these joint distributions, it is useful to study each component's marginal contribution to aggregate risk. In doing so, we can assess the largest contributors and develop disclosure limitation methodologies that best minimize reidentification. The spatial distribution of respondents, $Y_{I,J|-,j}$, and geography population uniques, A_j, are fundamental structures that determine the efficiencies of all other components. In turn, this study seeks to inform the design of confidential datasets by assessing these key determinants of risk and how they relate to two modified components, $C_{-,j}$ and $E_{-,j|J}$. Building upon these baseline estimates by assessing person-level characteristics along with contextual information (i.e., $Y_{I,J|i,j}$), the ultimate level of aggregate risk would then a function of the full formulation of the studied components (i.e., $C_{i,j}$ and $E_{i,j|J}$) and the assessment of those unobserved (i.e., $B_{i,j}$ and $D_{i,j}$).

As described in this study's limited hierarchical matching algorithm (Equations 6 to 10), I conduct disclosure analyses of geographic units, measuring the degree to which locational risk translates to survey respondents. I assess the likelihood of identifying unique locations among the total population of geographies and among my survey respondents. I also appraise how much locational risk declines when I assume that high-density areas adequately obscure respondent identity, regardless of location reidentification. However, I do not estimate the amount of disclosure risk resulting from individual-level characteristics of survey respondents. A full-assessment requires confidential information gathered from a real survey of persons, or at least a synthesized set of person-level characteristics, going beyond the scope of this study.

In turn, I focus my analyses on identifying problematic sets of contextual information (i.e., $Z_{-,j}$) and the implications for reidentifying individuals (i.e., $R_{I,J}$). The distribution of contextual variables characterizing these survey locations (i.e., $Y_{I,J|-,j}$) is determined by my experiment's sampling methodology. Indicative of the locational risk associated with a set of individual survey respondents, I then define "Locational Disclosure Risk", my outcome of interest, as the fraction of person-records that an intruder can confidently pinpoint their location. For each dataset, I calculate this as the proportion of survey respondents residing in an easily reidentified geographic unit (i.e., LR).

The proportion of survey respondents who are easily re-identified from contextual variables is derived from three components of risk: A_j, $C_{-,j}$, and $E_{-,j|J}$. Providing estimates for these components of locational risk, I tally the number of contextual-matches found in the population of geographies (i.e., A_j)

as well as those associated with my survey respondents (i.e., $C_{-,j}$) and whether these matches are high-density areas (i.e., $E_{-,j|J}$), incorporating a "geography population unique" threshold of 20 (T_1), a "person sample unique" threshold of 3 (T_3), and a "geographic unit population-size" threshold of 100,000 (T_5).

$$LR = \Sigma (Pr [R_{I,J}]) \quad (6)$$

$$Pr [R_{I,J}] = 1 - (Pr [Y_{I,J|-,j}] \times Pr [Z_{-,j}]) \quad (7)$$

$$Pr [Z_{-,j}] = Pr [(A_j \cap C_{-,j}) \cup (C_{-,j} \cap E_{-,j|J})] \quad (8)$$

$$Pr [Y_{I,J|-,j}] = f (T) \quad (9)$$

$$Pr [Z_{-,j}] = f (T, S, G, K, M) \quad (10)$$

T = Sampling methodology

S = Spatial scale of contextual data

G = Identified geography

K = Number of identifying "personal" or "contextual" keys

M = Masking technique

Where:

$R_{I,J}$ = Survey respondent (I) in sampled geographic unit (J) is easily reidentified, given "contextual" variables (j)

$Y_{I,J|-,j}$ = Survey respondent (I) in sampled geographic unit (J) has "contextual" variables (j)

$Z_{-,j}$ = Identifying "contextual" variables (j) are safe for release

A_j = Combination of values of "contextual" variables (j) is considered safe if combination occurs at least T_1 times among geographic unit population

$C_{-,j}$ = Combination of values of "contextual" variables (j) are considered safe if combination occurs at least T_3 times among survey respondents

$E_{-,j|J}$ = Combination of values of "contextual" variables (j) are considered safe when the population size of sampled geographic unit (J) is T_5 or more persons

Assumption-1: $Pr [Y_{I,J|i,j}] = Pr [Y_{I,J|-,j}]$
Assumption-2: $Pr [C_{i,j}] = Pr [C_{-,j}]$
Assumption-3: $T_1 = 20$, "Geography Population Unique" Threshold
Assumption-4: $T_3 = 3$, "Person Sample Unique" Threshold
Assumption-5: $T_5 = 100{,}000$, "Geographic Unit Population-Size" Threshold

In that I do not have the joint distribution of personal and contextual characteristics for a set of actual survey respondents, I make a simplifying assumption in my assessment of locational risk (i.e., $Pr [C_{i,j}] = Pr [C_{-,j}]$). Among those drawn from similar contexts, I assume that there are sufficient numbers of respondents with the same personal identifiers; and that risk only accrues when few respondents share contexts. As an illustrative example, "Context A" is described as a set of tracts that have 80 to 85 percent of their population living in poverty (as characterized by census tabulations) while "Persona B" is described as "white, male, 40 years-old". There is no disclosure risk when: (1) three people live in "Context A" and (2) all three people have "Persona B". If the first stipulation is met, I assume that the second is always true.

When personal information from a real survey of individuals is incorporated into the disclosure analysis, the risk from "person sample uniques" should generally be higher (i.e., $\Pr[C_{i,j}] \leq \Pr[C_{-,j}]$). The degree to which my modified estimate increases is a function of the real survey's sampling design that selects geographic units and individuals therein as well as its provision of identifying person-level information for respondents.

My study's synthetic sample of persons is (in essence) a random sample of individuals scattered across the United States, reflecting the areal distribution of the population (i.e., $Y_{I,J \mid i,j}$). This sampling approach maximizes the dispersion of respondents across space and purposely does not heap respondents within any particular geographic unit. Taking a bottom-up approach in my contextualization of my microdata files, I take my set of respondents and build contexts around them, instead of selecting geographic units first and then sampling for persons. Because my sample is drawn to identify a broad set of contexts that are most likely represented in surveys, my estimates of risk associated with "geography population uniques" (i.e., $\Pr[A_j]$) should be robust.

However my sampling methodology does not reflect how survey data is typically collected. In reality, surveys are likely to have a clustered design where a stratified sample of persons is drawn from primary sampling units, defined by geographic location. The higher the sampling rates of persons within geographic units, the more likely we can find adequate numbers of respondents (within shared contexts) having the same personal characteristics. Consequently, a clustered design would reduce risk associated with sample uniques in that more persons are drawn from a limited number of geographies (i.e., $\Pr[C_{i,j}] \approx 1$).

The degree to which this component converges to unity is a function of the concentration persons sampled within contextual matchsets. Primary sampling units may take the form of blended geographies, where persons are sampled across administrative units. If these straddled administrative units do not have exactly the same contexts, then a clustered sampling design may fail to produce an adequate number of respondents with shared personal and contextual characteristics. Minimizing the role of person-identifiers by requiring only three persons be in a contextual matchset for k-anonymity, I provide a baseline assessment of how the identification of sample uniques depends on the variability in contextual measures among geographic units and how this translates to survey respondents given the likelihood that they are residing in any particular location.

In turn, the risk posed by contextual sample uniques ($C_{-,j}$) is primarily an issue for persons residing in marginally common contexts (i.e., number of population matches equals t_1 where $[20 \leq t_1 < n]$ and n is an unknown small value) having relatively little chance of being represented in a dispersed sample (i.e., number of sample matches < 3). Given a sample that captures the contours of the population distribution outside administrative units, the chance of a respondent being in any particular contextual matchset is then a function of their location's area size and variability in contextual characteristics.

The spatial scale of geographic units operates in two diametric ways when assessing uniques within a sample of persons. Carving up finite space into units of various sizes influences disclosure risk by (1) setting the number of persons within geographic boundaries and (2) determining the number of geographic units in the population. Covering more land area and a higher absolute number of persons, large scale geographic units are more likely to have a respondent falling within its boundaries, given a randomly distributed sample of persons. In contrast, the relatively high number of population units for small scale geography increases variability in contextual characteristics, promoting the identification of uniques. Given the extremely large number of blockgroups, I expect that these small scale contexts are most susceptible to sample uniques, beaconing the need for a refined sampling methodology.

Because counties are relatively large areas that are few in number, I expect that my ad hoc estimates of "needle-in-the-haystack" protection are accurate (i.e., $\Pr[C_{i,j} \cap E_{i,j \mid J}] = \Pr[C_{-,j} \cap E_{-,j \mid J}]$). Most

sampling designs should be able to draw adequate numbers of respondents (sharing similar personal characteristics) from these relatively homogeneous administrative units.

3. c. Statistical Properties and Expected Outcomes

Laws of probability predict that the likelihood of identifying a unique is negatively associated with the number of units within the total population. Consequently, disclosure risk rises with the spatial scale of a geographic unit because of declining population size and the possibility of locating matches. In other words, small geographic units – that are large in number – offer more opportunities to locate matches, thereby ensuring confidentiality. Hence risk should be lowest for blockgroups, given the high probability of finding multiple matches. Applying the same logic to counties, risk of disclosure should be highest for these large-scale geographic units.

However, these probabilistic patterns are offset when we consider the dispersion of persons across geographic units. The likelihood of having more than one respondent living in similar contexts depends on the number of respondents sampled from each geographic unit. Because relatively few people are sampled from any set of "look-alike" blockgroups, the chance of achieving k-anonymity for these small-scale locations is low. Besides promoting k-anonymity, large counties also compensate for the small number of geographic units themselves in that people living in highly populated areas are considerably less likely to be reidentified. Given these population-dispersion factors, disclosure risk should actually decrease with the spatial scale of contextual data.

Besides these scale and dispersion factors, the ability to reidentify geographic units varies with the scope of the study. Directly identifying state/regional location and MSA-status limits the size of the population that is matched upon by confining the disclosure assessment to units within these areas, thereby increasing the likelihood of locating uniques. In turn, risk should generally be higher with the release of state, regional, and population density variables. Furthermore risk of reidentification should be highest when state-location is known because they are precise geographies covering the smallest land area.

The ability to identify population uniques is further enhanced when more information is provided about a given location. Consequently geographic units are generally more easily reidentified in datasets with relatively large numbers of contextual measures. The amount of risk resulting from these keys depends on the coarseness of their measurement.

Expected Disclosure Risk Associated with the Translation of Locational Risk to Survey Respondents and Design Elements of Contextual Data

	Translation of Locational Risk to Respondents				Design Elements	
Spatial Scale (S)	Population Uniques (A)	Sample Uniques (C)	Population Density (E)	Risk	Identified Geography, Number & Coarseness of Contextual Variables (G, K, M)	Risk
Counties	+++	No Change	+	+	MSA-Status, 1-Key, or Top, Bottom-25% Categories	+
Tracts	++	No Change	No Change	++	Region, 3-Keys, or 10%-Categories	++

Blockgroups	+	+++	No Change	+++	State, 5-Keys, or 1%-Categories	+++

3. d. Specific Considerations

To broadly inform the design of survey datasets, I construct a moderately-sized, cross-sectional sample of individuals that is randomly drawn to reflect each state's population distribution. In selecting my reidentification thresholds (i.e., T_1, T_3, and T_5), I take into account these survey characteristics as well as my project's need to concurrently study a variety of contextual data – whereby I consistently apply these parameters across all datasets.

Using statistical inference, survey respondents can refute an intruder's assertion that she has correctly reidentified their location. But to provide convincing evidence, researchers must certify that the chances of an intruder's match being correct is no better than if the intruder chose it at random from the other matches in the set. Hence, the task at hand is to avoid type I errors where we have mistakenly rejected the null hypothesis (H_0) that the intruder's locational-match is correct. Assuming that the significance level (or p-value) of 0.05 provides a sufficient test of this hypothesis, there is a 5-in-100 chance of wrongly rejecting the intruder's claim, if it is in fact true. If we are unable to reject the null hypothesis, it only suggests that there is not sufficient evidence against the intruder even though the match may still be incorrect. By correctly rejecting the null hypothesis, we can assuredly deny the intruder's claim. This is discussed in greater detail in VanWey, et al. (2005).

Although the p-value of 0.05 has been established as the standard for correctly rejecting null hypotheses (i.e., avoiding type I errors), this significance-level may be inadequate when applied to the disclosure of geographic units. Intruder costs – associated with verifying reidentified locations – are significantly affected by a dataset's scope which concentrates the geographic dispersion of matches. The question then becomes: Given the close proximity of matches, do intruders gather additional information that increases our chances of incorrectly rejecting their claim? If this is true, researchers may want to lower their significance level to offset reduced intruder costs associated with a dataset's release of geographic identifiers; whereby smaller p-values indicate stronger evidence for rejecting the null hypothesis. Further research is needed to inform this decision.

When deciding what constitutes adequate protection, researchers must also contemplate how the utility of their data will be affected. Given the small number of large geographic units, counties are generally easier to reidentify – based solely on the number of potential matches, without regards to the number and coarseness of contextual measures. Suppressing data for "at-risk" geographic units, contextual data would primarily be released for an extremely homogeneous set of counties; thereby reducing variation in these measures to the point where they become analytically useless. This suppression bias is exacerbated with increased confidence threshold-levels.

Finally when choosing a "geography population unique" threshold (i.e., T_1), researchers need to consider the spatial scale of contexts and whether the dataset will provide identifying geographic information. Reducing the number of matches, the ability to provide convincing evidence is significantly lowered when a dataset increases its contextual scale or limits its scope of study.

While I am unable to address variability in intruder costs, I do incorporate a consistent definition of risk that maximizes data utility across all test datasets. In turn, I choose the reidentification confidence level of $p>0.05$, where "at-risk" geographic units are those with 1 to 19 matches. This reidentification confidence level conversely translates into a confidence threshold value of 20, where contextual data are considered safe for release when geographic units have twenty or more matches (i.e., $T_1=20$).

Besides intruder costs, researchers also need to consider the characteristics of their sampling methodology and the spatial scale of contexts when selecting a "person sample unique" threshold (i.e., T_3). While it may be more desirable to have five respondents in the sample with the same contexts, achieving this threshold may not be feasible given a moderately small sample of persons thinly distributed across numerous geographic units. Incorporating a threshold commonly used in federal survey research (Subcommittee on Disclosure Limitation Methodology 2005) and addressing my need to study blockgroups, I consider contextual data safe for release when three or more survey respondents have exactly the same key-values (i.e., T_3=3).

When choosing a "geographic unit population-size" threshold (i.e., T_5), researchers need to consider such survey characteristics as its provision of longitudinal information and detailed data on social networks; as well as the sampling rate of persons within geographic units. The more people drawn from an area, the less affective the protection offered by its large population. It is also relatively easy to reidentify persons in a high-density area when intruders can utilize information describing a person's life over an extended period of time or their family members.

The U.S. Census Bureau has set disclosure standards allowing for the release of geographic information onto microdata files (5% sample) for areas with 100,000 or more persons (U.S. Census Bureau 2003). For panel surveys containing detailed information, the Census Bureau has an alternative approach. For instance, the Survey of Income and Program Participation does not identify states and metropolitan areas with populations less than 250,000 (U.S. Census Bureau 2001). Given the characteristics of my hypothetical survey, I apply the processing rule where survey respondents – living in counties with 100,000 or more persons – cannot be reidentified from their contextual characteristics (i.e., T_5=100,000).

4. Assessing Disclosure Risk of Masked Contextual Data

As the empirical basis of my study, I conduct experiments to assess the amount of disclosure risk associated with a dataset's contextual data (Domingo-Ferrer and Torra, 2001a). Using the test microdata file as my unit of analysis, the proportion of survey respondents whose locations are easily reidentified as the outcome of interest, and associated experimental traits, I produce descriptive analyses to test my hypotheses. In doing so, I followed six methodological steps:

- select contextual measures and identify population of base variable sets;
- draw a sample of base variable sets;
- construct test datasets that vary in spatial scale of contextual measures, identified geography, number of keys, and masking method, holding constant sample of base variable sets;
- construct microdata files composed of a single synthetic sample of survey respondents, attaching test datasets of contextual information to these person-level records;
- reidentify locations of survey respondents, using available geographic identifiers and contextual data for counties, tracts, and blockgroups; and
- calculate aggregate disclosure risk for each test microdata file as proportion of survey respondents whose geographic location is easily reidentified, identifying datasets having risk-levels falling below 5%.

4. a. Sources, Measures, and Sampled Datasets of Contextual Data

My sources of contextual data are summary files tabulated from the 2000 U.S. Census of Population and Housing (U.S. Department of Commerce 2000a, 2000b, 2000c). These summary files are prominent public-use databases within the social sciences, providing a diversity of measures and a range of geographic detail. Contextual data are compiled from published tabulations for all blockgroups, tracts, and counties in the United States.

To identify which measures would be of most interest to researchers, I draw on the sociological literature on stratification, residential segregation and mobility, and labor markets. I limit my test datasets to those having subsets of the seventeen concepts (i.e., base variables) listed below.

- Race/Ethnic Composition
  - % Persons, Non-Hispanic White
  - % Persons, Non-Hispanic African-American
  - % Persons, Non-Hispanic Asian or Pacific Islander
  - % Persons, Non-Hispanic Other Race
  - % Persons, Hispanic
  - % Persons, Foreign-Born
  - % Foreign-Born, Naturalized Citizens
  - % Households, Linguistically Isolated

- Socioeconomic Status
  - % Persons, In-Poverty
  - % Households, With Wage Income
  - % Households, Receiving Public Assistance
  - % Persons Age 25+, College Degree

- Social Context
  - % Families, Female-Headed
  - % Persons Age 16-19, Neither Enrolled nor Graduated from High School
  - % Housing Units, Owner-Occupied

- Labor Market
  - % Persons Age 16+, Civilian Labor Force
  - % Civilian Labor Force, Unemployed

The conceptual content of datasets varies with (1) the number of base variables (i.e., k = 1, 2, 3, 4, 5) and (2) which base variables are included in the sets. All possible combinations of base variables (i.e., base variables sets) are constructed for sets containing one, two, three, four, and five concepts (N_k = 17; 136; 680; 2,380; 6,188; respectively). I compile datasets – holding constant a specific base variable set – by varying its spatial scale, identified geography, and masking technique. Consequently, I can better clarify risk factors associated with spatial scale and geographic relationships and avoid confounding my results with varying sub-domains.

Relationships between variables within a dataset have implications for reidentification and vary with the geographic scale of the measures. Large amounts of variation within measures (i.e., wide range of values) increase the likelihood of identifying uniques and, therefore, disclosure risk. However large amounts of collinearity among measures may allow producers to release more variables within a dataset without drastically increasing disclosure risk. To assure an unbiased selection of measures that are representative of the degrees of variance and collinearity among all possible datasets, I sample 137 sets of base variables composed of one to five concepts. All seventeen base variable sets with a single concept are included in my study. Thirty datasets are randomly sampled from each stratum of the multiple-concept base variable sets. An assessment of the effectiveness of this sampling approach is presented in a working paper that is available upon request.

4. b. Experimental Traits

Given a finite number of sampled sets of contextual concepts (n=137), datasets ($C_{b,s,g,m}$) are varied along the [B x S x G x M] matrix of: (1) base variable sets (b = *1 to 137, described above*); (2) spatial scales of contextual data (s = *1, 2, 3 = counties, tracts, blockgroups*); (3) identified geographies (g = *1, 2, 3, 4, 5, 6 = none, population density, division, state, population density and division, population density and state*); and (4) masking techniques (m = *1, 2, 3, 4, 5, 6 = 1%, 5%, 10%, 15%, 20%, and Top and Bottom-25% categories*). Consequently, 14,796 datasets (= 137 x 3 x 6 x 6) are compiled and assessed.

Illustrating this nomenclature, two datasets are compiled using one of the sampled set of concepts consisting of five base variables (b=137): (1) % Persons, Non-Hispanic White; (2) % Persons, Foreign-Born; (3) % Households, Receiving Public Assistance; (4) % Housing Units, Owner-Occupied; and (5) % Civilian Labor Force, Unemployed. Test dataset ($C_{137,1,1,1}$) contains these five contextual variables measured at the county-level (s=1) without any geographic identifiers (g=1), masked into 1% categories (m=1). In comparison, test dataset ($C_{137,3,3,3}$) contains these five contextual variables measured at the blockgroup-level (s=3) along with the identification of division (g=3), masked into 5% categories (m=3).

Presented in Appendix Table A-1, equal proportions of datasets across categories of spatial scale, geographic identifiers, and masking techniques reflect my experimental design. Taking a base variable set of concepts, I compile datasets that systematically vary across a matrix of these experimental traits. However, the proportions of datasets across categories of the number and conceptual composition of contextual variables reflect the random sampling of base variables sets, stratified by key-sets (i.e., including all 17 sets with 1 key; including 30 sets each with 2, 3, 4, and 5 keys). Every conceptual domain (i.e., base variable) is represented in my test datasets. Fifteen of the seventeen concepts were included in 15 to 24% of the datasets. However, "% Persons, Non-Hispanic Asian or Pacific Islander" was least likely to be represented (7% of datasets); while "% Persons, Non-Hispanic White" was most often represented (31% of datasets). These inconsistencies are strictly random artifacts.

Given these base variable sets, I constrain my matching process by geographic identifiers released in the dataset. A dataset can directly identify the state, division, and population density of respondent location. U.S. Census geographic divisions categorize states into seven regional groups of (1) New England, (2) Middle Atlantic, (3) East North Central, (4) West North Central, (5) South Atlantic, (6) East South Central, (7) West South Central, (8) Mountain, and (9) Pacific. Population density is defined by three categories of MSA-status: (1) MSA 1-million or more, (2) MSA less than 1-million, and (3) Non-MSA (Sources: U.S. Census Bureau, 2002, 2006a, 2006b). Measured at the county-level, these data are also used to characterize the population density of tracts and blockgroups.

Finally I systematically vary the amount of measurement detail across my experimental datasets; thereby assessing how rates of local suppression fluctuate with global recoding schema. After top-coding and bottom-coding my continuous variables to conceal outliers, I recode contextual measures into six grades of coarseness (i.e., 1%, 5%, 10%, 15%, 20%, and Top and Bottom-25% categories). Outliers were identified as those within the top and bottom 0.5% of each variables distribution (Zayatz, 2005), given geographically-specific distributions defined by each dataset's identified geography. Contextual variables are recoded into aggregated categories based on their absolute values (i.e., absolute recoding). For example, let us consider a county having 72% of its population that is non-Hispanic White. Coarsening the measure into 10%-categories, the county would be characterized as having an absolute value that falls between 70% and 80%.[2]

Global Recoding of Contextual Variables

Coarseness of Contextual Variables	Metric Spaces	Absolute Values	Coarseness of Contextual Variables	Metric Spaces	Absolute Values
1%-Categories	100	0%, 1%, 2% . . . 98%, 99%, 100%	15%-Categories	7	0 - 14%, 15 - 29%, . . . 75 - 89%, 90 -100%
5%-Categories	20	0 - 4%, 5 - 9%, . . . 90 - 94%, 95 - 100%	20%-Categories	5	0 - 19%, 20 - 39%, . . . 60 - 79%, 80 - 100%
10%-Categories	10	0 - 9%, 10 - 19%, . . . 80 - 89%, 90 - 100%	Top, Bottom-25% Categories	3	Top-25%, Bottom-25%, Other

4.c. Synthetic Sample of Survey Respondents and their Geographic Locations

I derive synthetic person-records with attached contextual data from a sample of blocks represented in the 2000 U.S. Census of Population and Housing (U.S. Department of Commerce 2000a). A stratified sample of blocks is drawn to reflect the areal distribution of the U.S. population across states. The block is chosen as my sampling unit because it most closely approximates the residential location of our theoretically ideal sample of individual survey respondents (i.e., persons). Being the foundational spatial unit from which all geographies are built upon, blocks also pinpoint various contexts to a single location. In turn, tabulations from identified counties, tracts, and blockgroups, which overlap with my sampled blocks, are included in my study as contextual data. These contextual data are then represented in a synthetic dataset of person-records.

Fifty state-specific block samples (excluding the District of Columbia) are drawn with probability-proportional-to-size without replacement (PPS). Each block within a state has a probability of selection that is proportional to its population density, defined as the total number of persons per square meter of block area. Whereas I am not studying how individual-level characteristics shape disclosure risk, I purposely set the number of person-records to one per sampled block-unit. By sampling one person per block, I disperse my survey respondents as thinly as possible over space. Presented in Appendix Table A-3, 11,475 blocks are sampled, representing 11,475 synthetic persons dispersed across approximately 5% of all blockgroups, 14% of all tracts, and 57% of all counties in the U.S.

My dispersed sampling approach broadens the selection of locations for study; thereby enhancing the assessment of risk associated with contextual data. It also furnishes the most cautious (or highest) estimates of locational risk resulting from sample uniques. Representing the antithesis of modern survey collection methodology, my sample provides a baseline for future work that explores disclosure risk associated with clustered and hierarchical designs.

4. d. Locational Disclosure Risk Associated with Test Datasets

The foundation of my first component of locational disclosure risk (i.e., A_j) is the confidence-level of correctly identifying the location of a survey respondent among a population of geographic units. Given the p-value of .05, I assume that I can strongly refute an intruder's claim of correctly reidentifying a location (H_0) when there is a 5-in-100 (or less) chance of wrongly rejecting this hypothesis (VanWey, et. al, 2005). Taking the inverse of this confidence-level, a geographic unit is considered easily reidentified when it has fewer than 20 matches (i.e., at_risk=1).

To ascertain the number of matches, I compare the contextual characteristics associated with a sample of locations – with a master contextual file containing the same measures for the full population of geographic units and their identifying information. Data in the master file are top and bottom coded and collapsed into intervals as defined by the test dataset. In that my test datasets also directly identify population density, division, and state-location, I further refine my matching process by utilizing geographically-specific master contextual files. Because contextual data for a sample of locations are originally drawn from this master file and have not been perturbed, the identification of matches is exact. Consequently, counting the number of matches simply requires that I tabulate the number of geographic units in the population file having a specified set of contextual characteristics, coinciding with those found in my experimental survey (Winkler 2004).

Assessing other disclosure mechanisms that transfer locational risk to individual survey respondents (i.e., $C_{-,j}$ and $E_{-,j|J}$), I apply two additional processing rules that follow established standards for k-anonymity and releasing geographic information on public-use microdata files. A person's geographic location is also considered easily reidentified when less than three respondents share the same contextual characteristics (i.e., at_risk=1, if number of matches in sample is less than 3). This is true even when their location characteristics are commonly found in the population. Furthermore respondents located in

counties with 100,000 or more persons are assigned a null value of risk, for respondents whose contextual characteristics are commonly found in the survey (i.e., at_risk=0, if number of matches in sample is 3 or more).

After assessing whether a sampled geographic unit is considered easily reidentified, the "at-risk" status of this location is then attached to survey respondents who reside in the area (i.e., A_j). A person's "at-risk" status is further refined to reflect their uniqueness within the survey as well as their location's population density (i.e., $C_{-,j}$ and $E_{-,j|J}$). Measuring the amount of aggregate risk associated with a dataset from these three components, I calculate weighted proportions of survey respondents who are considered "at-risk" of being reidentified because of the release of contextual data (i.e., proportion of respondents with at_risk=1). Using this summary measure, I create an indicator that characterizes a dataset as having less than 5% of survey respondents whose geographic location is easily reidentified (i.e., suppress5=1).

For blockgroup- and tract-level contextual data, locational risk directly translates to respondents because these small scale areas always have sparse populations (i.e., less than 100,000 persons). However, risk posed by county-level contextual data may be negated by the protection offered by dense populations. Consequently, I create two sets of estimates for county-level datasets: (1) for all survey respondents and (2) for respondents living in areas with less than 100,000 persons.

Analyzing metadata characterizing my sample of 14,796 datasets (see Appendix Table A-3 for further details), I render estimates that are generalized to all possible datasets. In doing so, I produce estimates of the average aggregate risk and the proportion of datasets with minimal suppression rates for various reidentification components.[3]

5. Presentation of Results

When designing a public-use dataset, a producer seeks to release as much information as possible while ensuring that all locations are kept confidential. In doing so, a producer needs to consider the trade-offs in data utility when defining acceptable levels of (1) local suppression, (2) measurement detail, and (3) contextual content. These sources of information, or information loss, have distinct implications.

Rates of local suppression are reflected by the level of aggregate risk, expressed as the proportion of respondents whose contextual data may not be safely released. Perturbative methods, such as swapping, can then be used to construct confidential information that replaces these missing values. The producer must consider how these ascribed data may distort analyses and whether a group of geographic units is particularly affected by these aberrations.

Establishing an acceptable rate of local suppression, the producer can achieve this aggregate risk-level by coarsening the contextual measures. However, if the necessary amount of coarsening results in analytically useless data, the producer may decide to release fewer contextual variables or less specific geographic identifiers.

Given these decision-making factors, I offer descriptive analyses that inform the design of confidential datasets. Each series in the following figures represents a different facet of the assessment of disclosure risk for contextual data, characterized according to: (1) the configuration of spatial data and (2) the nature of the reidentification process. In a previous study, I conducted disclosure analyses of geographic units, analyzing non-hierarchical spatial data (Witkowski 2007). I now analyze hierarchical datafiles composed of contextualized person-records to assess how locational risk extends from a sample of geographic units to a set of survey respondents nested therein. Segueing between this work and the current study, I provide risk estimates that are at the geographic-level as well as the person-level.

[Figures A, B, C Here]

Identifying uniques within the population of geographic units (i.e., blue bars), half of all counties and just over a third of all tracts and blockgroups are easily reidentified (50%, 37%, and 36%, respectively). Reidentification rates for counties represented in my disperse sample of respondents generally mirror those found in the general population (52%). This is not surprising since 57% of counties are sampled. Given my synthetic methodology, densely populated counties are also more likely to be sampled and these highly populated areas tend to exhibit slightly more unique characteristics, increasing risk by 2 percentage-points. The concentration of persons within rare counties is further indicated by the 11 percentage-point increase in risk when this component of locational risk is translated to survey respondents (63%, up from 52%).

Disclosure risk for sampled tracts (39%) is also approximately equal to its population (37%), with 14% of tracts being sampled. But unlike counties, persons are not concentrated into unique tracts since aggregate risk-levels change very little with its translation to survey respondents (40%). The reason for the lack of a conversion shift is the methodology underlying the construction of these administrative units which places a cap on their population size. With as many as 1,500 to 8,000 people, census tracts designate areas that are relatively uniform in their population characteristics, economic status, and living conditions (U.S. Census Bureau 2000).

With as many as 600 to 3,000 people, blockgroups subdivide tracts into areas bounded by visible and legal features (e.g., streets, property lines) (U.S. Census Bureau 2000); whereby space is further delineated into units that are relatively small in area, large in number, and heterogeneous. As a result, blockgroup-level data exhibit considerably more variation in contextual characteristics, having substantial disclosure ramifications. With a sampling rate of 5%, blockgroups selected for study have relatively common contextual characteristics and, thus, lower levels of risk (27%), compared to the full population of geographic units (36%). This pattern is expected given central tendency theory, whereby the characteristics of a sample reflect those most often found in the population. As with tract-level data, blockgroups also have similar levels of locational risk regardless if estimates are derived from geographic- or person- units (27%).

Taking person-level estimates of the first risk component as a baseline (i.e., solid green bars), the locations of respondents may be further compromised from contexts rarely found among those surveyed. Earlier I searched for matches within the population of geographies, but now I turn my search for matches to those within a sample of persons. Given my dispersed sample of individuals, the likelihood that numerous respondents fall within a similar contextual setting depends on the areal size of locations as well the distribution of contextual characteristics among the geographic population.

As the fourth series indicates (i.e., striped green bars), this type of risk does not emanate from county- and tract- level contexts. However blockgroup-level data experience an 18 percentage-point increase in risk (45%, up from 27%) after assessing for contextual sample uniques. This is expected given the small areal size of blockgroups as well as the relatively large amounts of contextual variation. For reasons provided in Section 3.b., the reidentification of locations should then be minimally affected by the spatial mismatch between primary sampling units and county- and tract- level contexts. On the other hand, blockgroup-level contextual data may be better suited to hierarchical designs that pay particular attention to drawing nested samples. While this is an interesting finding, further research using personal identifiers is needed to fully assess its implications.

Building upon these modified estimates of “person sample uniques”, I then assess the “needle-in-haystack” protection offered by large populations falling within county borders, where these results are presented in the last two series of Figure A (i.e., solid and striped red bars). Approximately XX% of my survey respondents live in highly populated counties that have rare contextual characteristics. Assuming that the reidentification of these unique geographies poses no real harm to respondents, locational risk drops by 45 percentage-point (18%, down from 63%). But for those living in low-density areas who are not offered such protection, 51% of respondents remain in easily reidentified counties. Consequently, the

potential for bias is considerable when contextual information for large fractions of rural populations is further altered for confidentiality.

Turning to the bottom panel of the figures, we see how these components of risk ultimately affect dataset design. In contextualizing microdata, producers must choose the scale, number, and coarseness of contextual measures and decide whether they will also release identified geography. Underlying these decisions is the potential distortion of contextual information stemming from local suppression or perturbation. If our goal is to construct a dataset that has a minimal level of distortion, what is the likelihood of identifying a design that meets this benchmark of utility? And how does the translation of locational risk to individuals – captured by my modified components of risk – ultimately influence the chances of meeting this design goal.

To answer these questions, I take estimates of locational risk for my test datasets and identify those having less than 5% of respondents with easily-reidentified locations. I select this benchmark strictly for expository purposes since no research has assessed how the analytical utility of contextual data varies with aggregate risk. These test datasets systematically vary in four experimental traits (i.e., scale, identified geography, number of contextual keys, and measurement coarseness), representing pertinent design elements. I then calculate the proportion of "promising" datasets who meet this utility criterion, estimating the likelihood of identifying designs that require minimal perturbation.

For datafiles containing county-level contexts, the likelihood of identifying a promising design nearly quadruples when high-density areas are considered protective (26%, up from 7%). However the design benefits of this assumption do not translate in the production of analytically useful data for rural populations, where only 12% of datasets would require a minimal amount of perturbation to ensure confidentiality for respondents living in counties with less than 100,000 persons.

For datafiles containing tract-level contexts, approximately one-quarter (23%) are promising in their design. Furthermore, the number of favorable tract-level designs is not reduced by the risk posed by sample uniques. Although this component is negligible for tracts, blockgroup-level data experience a jump in aggregate risk from sample uniques. I expected that such an increase in risk would result in more datasets requiring higher rates of perturbation. But surprisingly, the chance of identifying a promising blockgroup-level design is not dampened by this component (16% versus 17%). Assessing the first component of risk, an extreme level of homogeneity within the population of blockgroups was required to meet the stringent 5% benchmark. Consequently, the design elements that produced high match rates in the geographic population are the same as those producing large numbers of matches within a sample.

6. Conclusions

In writing this paper, I have two complementary objectives. My first aim is to develop an analytical framework for conducting disclosure analysis of contextualized microdata. My second aim is to describe the intricacies involved in implementing such an analysis. In doing so, I illustrate the mechanisms that translate disclosure risk of geographic units to individual survey respondents (i.e., "locational" risk), providing empirical estimates as a way to demonstrate the processes involved and their implications for dataset design.

In meeting my first objective, I describe the empirical steps involved in reidentifying geographic units and individuals that are represented within a survey dataset. From this description, I develop a matching algorithm that lays the foundation for assessing of disclosure risk of contextualized microdata. Hierarchical in nature, this algorithm considers both the individual and contextual characteristics of survey respondents as well as their spatial dispersion, whereby five components of risk are formulated and brought together in an analytical framework that is recursive.

As for my second objective, I implement a modified version of my hierarchical matching algorithm to assess three of the five components of risk, that of identifying the location of survey respondents whose contextual characteristics: (1) are rarely found among the total population of geographic units; (2) are rarely found within a survey; and (3) pose no disclosure risk given the protection offered by the area's dense population. Expounding upon these components, I discuss the relationships between these determinants, the mechanisms that shape each components' contribution to aggregate risk, and factors that should be considered when setting parameters that define locational risk. Utilizing a synthetic set of survey respondents that are thinly dispersed across geographic units, I conduct reidentification experiments for a series of test datasets that vary in the scale, variable composition, and measurement detail of contextual information. Gathering metadata from these simulations, I then assess patterns of aggregate risk for these three components of locational risk.

Analyses of these metadata provide insight into how the construction of administrative units plays a role in disclosure risk. Counties, tracts, and blockgroups vary dramatically in their areal size, number of units, and the heterogeneity in the size and composition of their populations. These scale factors of geographic units influence the translation of location risk to survey respondents through the identification of "personal sample uniques" and "needle-in-haystack" protection. A survey's sampling methodology also influences the reidentification process, particularly as it relates to compilation of respondents within similar contexts.

Most interesting is how different components of risk help intruders pinpoint the location of survey respondents and how these sources vary with the spatial scale of contexts. For instance, my results indicate that the risk posed by inadequate numbers of respondents sharing common contexts may be considerable when blockgroup-level information is associated with a geographically dispersed sample of persons. Yet the construction of contextualized datafiles may ultimately be unaffected by this component of risk since the selection of designs with minimal suppression rates seems to be unconstrained by sample uniques.

The current study utilizes a synthetic set of survey respondents derived from a dispersed sampling of blocks reflecting the areal distribution of the population. While this sampling approach is necessary for the initial stages of this work, further research needs to be conducted to assess how locational risk varies with clustered and hierarchical sampling designs, thereby better reflecting current methodological practice.

Furthermore I plan to conduct a full assessment of disclosure risk that brings together a limited set of personal identifying variables and my experimental and approximated sets of contextual variables. Wholly implementing my hierarchical matching algorithm under alternative sampling frameworks, a series of re-identification experiments would consider all five components both individually and additively, assessing their relative contributions to disclosure risk.

7. References

Armstrong, Marc P., Gerard Rushton, and Dale L. Zimmerman. 1999. "Geographically Masking Health Data to Preserve Confidentiality." *Statistics in Medicine* 18: 497-525.

DeWaal, A.G. and L.C.R.J. Willenborg. 1995. "Global Recodings and Local Suppressions in Microdata Sets." *Proceedings of Statistics Canada*, 95: 121-132.

DeWaal, A.G. and L.C.R.J. Willenborg. 1996. "A View of Statistical Disclosure Control for Microdata." *Survey Methodology*, 22: 95-103.

Domingo-Ferrer, Josep and Vicenc Torra. 2001a. "A Quantitative Comparison of Disclosure Control Methods for Microdata." Pp. 111-133 in *Confidentiality, Disclosure, and Data Access: Theory and*

Practical Application for Statistical Agencies, edited by P. Doyle, J.J. Lane, J.J.M. Theeuwes, and L.M. Zayatz. North-Holland: Amsterdam.

Domingo-Ferrer, Josep and Vicenc Torra. 2001b. “Disclosure Control Methods and Information Loss for Microdata.” Pp. 91-110 in *Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies*, edited by P. Doyle, J.J. Lane, J.J.M. Theeuwes, and L.M. Zayatz. North-Holland: Amsterdam.

Duke-Williams, Oliver and Philip Rees. 1998. “Can Census Offices Publish Statistics for More than One Small Area Geography? An Analysis of the Differencing Problem in Statistical Disclosure.” *International Journal of Geographical Information Science*, 12(6): 579-605.

Duncan, George and Diane Lambert. 1989. “The Risk of Disclosure for Microdata.” *Journal of Business and Economic Statistics*, 7(2): 207-217.

ESRI. 2006. *GIS Dictionary*. Last modified April 26, 2006. <http://support.esri.com/index.cfm?fa=knowledgebase.gisDictionary.gateway>

Interagency Confidentiality and Data Access Group. 1999. Checklist on Disclosure Potential of Proposed Data Releases. Statistical Policy Office. Office of Information and Regulatory Affairs. Office of Management and Budget. Washington, D.C.

Kyle, S., D. A. Samuelson, F. Scheuren, and N. Vicinanze. 2007. “Explaining Discrepancies between Official Votes and Exit Polls in the 2004 Presidential Election.” *Chance*, 20(2): 36-47.

Lambert, Diane. 1993. “Measures of Disclosure Risk and Harm.” *Journal of Official Statistics*, 9(2): 313-331.

Raghunathan, T.E., J.P. Reiter, and D.R. Rubin. 2003. “Multiple Imputation for Statistical Disclosure Limitation.” *Journal of Official Statistics*, 19: 1-16.

Saalfeld, A., Laura Zayatz, and E. Hoel. 1992. “Contextual Variables via Geographic Sorting: A Moving Averages Approach.” Proceedings of the Section on Survey Research Methods. American Statistical Association. Alexandria, VA. Pp. 691-696.

Subcommittee on Disclosure Limitation Methodology, Confidentiality and Data Access Committee, Federal Committee on Statistical Methodology. 2005. Statistical Policy Working Paper 22 (Second version, 2005): Report on Statistical Disclosure Limitation Methodology. Revised December 2005: Report GAO-010126SP. Washington, DC: Statistical and Science Policy. Office of Information and Regulatory Affairs. Office of Management and Budget.

Sweeney, Latanya. 2002. “Achieving K-Anonymity Privacy Protection Using Generalization and Suppression.” *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5): 571-588.

United States General Accounting Office. 2001. Record Linkage and Privacy: Issues in Creating New Federal and Statistical Information, GAO-01-126SP. United States General Accounting Office. Washington DC.

U.S. Census Bureau. 2003. *Census 2000, Public Use Microdata Sample (PUMS), United States, Technical Documentation*. http://www.census.gov/prod/cen2000/doc/pums.pdf

U.S. Census Bureau. 2001. U.S. Department of Commerce, Economics and Statistics Administration. *Survey of Income and Program Participation Users' Guide, Third Edition*. Washington, DC. http://www.sipp.census.gov/sipp/usrguide/sipp2001.pdf

U.S. Census Bureau. 2000. *Census 2000 Geographic Terms and Concepts*. Reference Resources for Understanding Census Bureau Geography. http://www.census.gov/geo/www/tiger/glossry2.pdf

VanWey, Leah K., Ronald R. Rindfuss, Myron P. Gutmann, Barbara Entwisle, and Deborah L. Balk. 2005. "Confidentiality and Spatially Explicit Data: Concerns and Challenges". *Proceedings of the National Academy of Sciences of the United States of America*,102 (43): 15337-15342.

Winkler, William. 2004. Masking and Reidentification Methods for Public-use Microdata: Overview and Research Problems. Issued: October 21, 2004: Research Report Series (Statistics #2004-06). Washington, DC: Statistical Research Division, U.S. Census Bureau.

Witkowski, Kristine M. 2007. "Disclosure Risk of Contextual Data: The Role of Spatial Scale, Identified Geography, and Measurement Detail in Public-Use Files." Submitted to *Journal of Official Statistics*.

Zayatz, Laura. 2005. Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update. Revised August 31, 2005: Research Report Series (Statistics #2005-06). Washington, DC: Statistical Research Division, U.S. Census Bureau.

8. Data Sources

U.S. Census Bureau, Population Division. 2002. Census 2000 PHC-T-3. Ranking Tables for Metropolitan Areas: 1990 and 2000 (Table 3: Metropolitan Areas Ranked by Population). Last Revised: July 31, 2002
< Web Page: http://www.census.gov/population/www/cen2000/phc-t3.html>
< Direct Link: http://www.census.gov/population/cen2000/phc-t3/tab03.xls>

U.S. Census Bureau, Geography Division, Cartographic Products Management Branch. 2005. *Cartographic Boundary Files*. Last Revised: August 24, 2005.
<http://www.census.gov/geo/www/cob/index.html>

U.S. Census Bureau, Population Division. 2006a. *Geographic Relationship Files: 1999 MA to 2003 CBSA* (Excel file). Last Modified: August 18, 2006.
<http://www.census.gov/population/www/estimates/metroarea.html>
<Direct Link: http://www.census.gov/population/www/estimates/CBSA03_MSA99.xls>

U.S. Census Bureau. 2006b. 2000 *Census of Population and Housing, Summary File 1 (Matrices P1)* generated by Kristine Witkowski; using American FactFinder; <http://factfinder.census.gov>; (6 November 2006).

U.S. Department of Commerce, Bureau of the Census. CENSUS OF POPULATION AND HOUSING, 2000a [UNITED STATES]: SUMMARY FILE 1 SUPPLEMENT, STATES [Computer file]. ICPSR release. Washington, DC: U.S. Dept. of Commerce, Bureau of the Census [producer], 2003. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, [distributor], 2003.

U.S. Department of Commerce, Bureau of the Census, and Inter-university Consortium for Political and Social Research. CENSUS OF POPULATION AND HOUSING, 2000b [UNITED STATES]: BLOCK GROUP SUBSET FROM SUMMARY FILE 3 [Computer file]. ICPSR ed. Washington, DC: U.S. Dept. of Commerce, Bureau of the Census, and Ann Arbor, MI: Inter-university Consortium for Political and Social

Research [producers], 2004. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2004.

U.S. Department of Commerce, Bureau of the Census, and Inter-university Consortium for Political and Social Research. CENSUS OF POPULATION AND HOUSING, 2000c [UNITED STATES]: SELECTED SUBSETS FROM SUMMARY FILE 3 [Computer file]. 2nd ICPSR ed. Washington, DC: U.S. Dept. of Commerce, Bureau of the Census, and Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producers], 2004. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2004.

9. Endnotes

[1] To clarifying terms used throughout this paper, contextual measures are referred to as “key variables” being how combinations of their values are used to locate uniques within its target population. On the other hand, geographic identifiers are considered “block variables” being how they are used to subset locations within the matching process.

[2] I conduct another set of simulations analyzing contextual measures that are coarsened based on their percentile distribution (i.e., percentile recoding). Twenty percent of all counties have at most 66% of their population being non-Hispanic White (i.e., 20th percentile at 66.15%); while thirty percent of all counties have at most 76% of their population being non-Hispanic White (i.e., 30th percentile at 76.14%). Coarsening the measure into deciles categories, my exemplar county – having 72% of its population being non-Hispanic White – would be characterized as falling between 20th and 30th percentiles (i.e., the third decile).

As illustrated in Appendix Table A-3, disclosure risk is heightened considerably by this global recoding approach. Counties having a rare characteristic – those with an outlying value at the tails of a contextual variable’s continuous probability distribution – are less likely to be reidentified with percentile coarsening. However, counties sharing a relatively common characteristic – those within the middle of the distribution – are actually more likely to be reidentified with percentile coarsening.

Building upon my previous example, let us consider my exemplar county which is one among a population of 3,140. With absolute recoding, this county has approximately 346 matches with values between 70% and 80%. With percentile recoding, there are 315 matches with values between 66.15% and 76.14%. In turn, percentile recoding automatically sets an upper bound to the number of matches, resulting in relatively higher risk for more typical counties.

[3] The standard errors for my paper’s point estimates are available from the author upon request. But suffice it to say, all standard errors fall below .03.

10. List of Tables and Figures in Paper

11. List of Tables and Figures in Appendix

Figure A. Disclosure Risk of Geographic Units and Survey Respondents and Suppression Rates of Contextual Data, By Risk Components of County-Level Measures

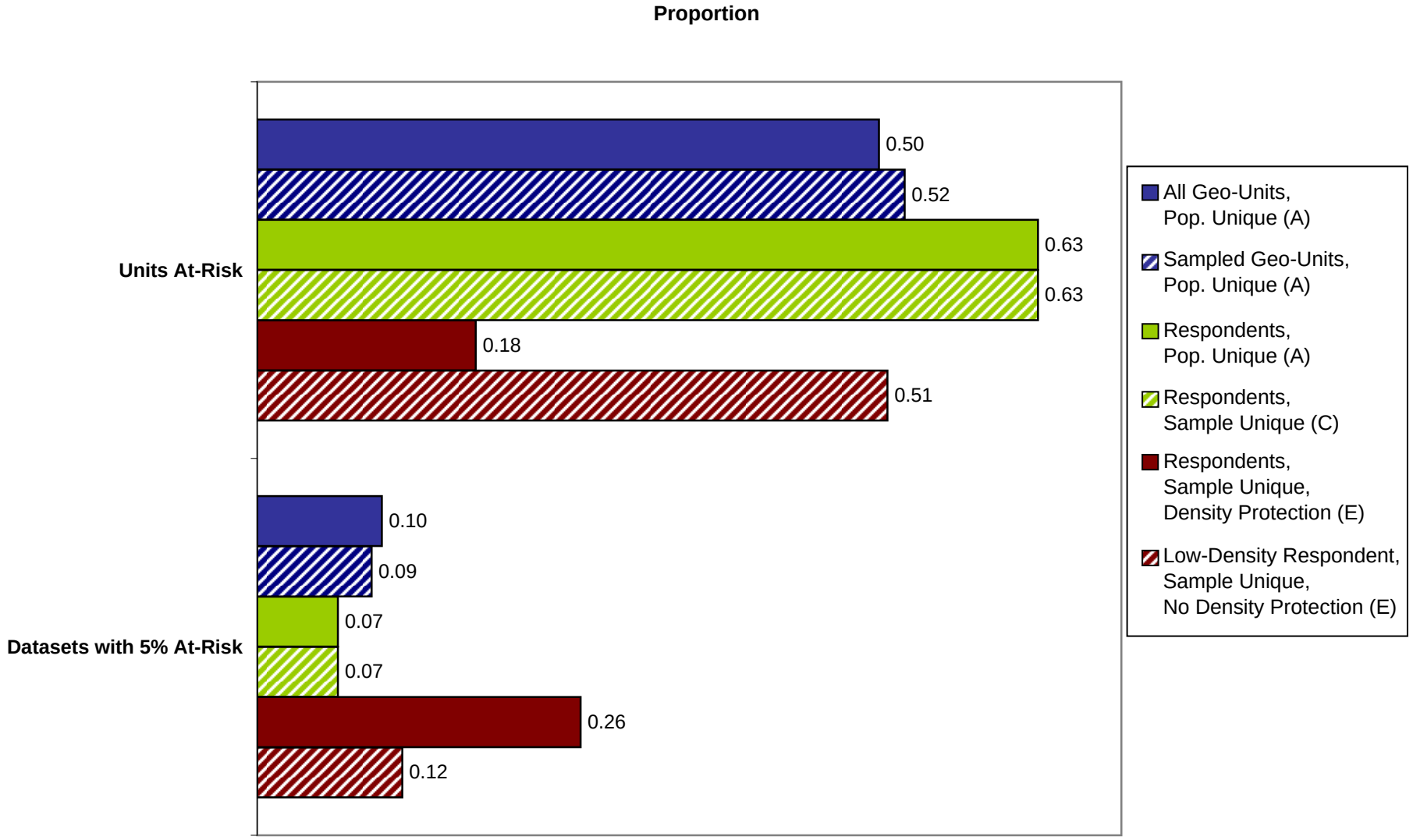

Figure B. Disclosure Risk of Geographic Units and Survey Respondents and Suppression Rates of Contextual Data, By Risk Components of Tract-Level Measures

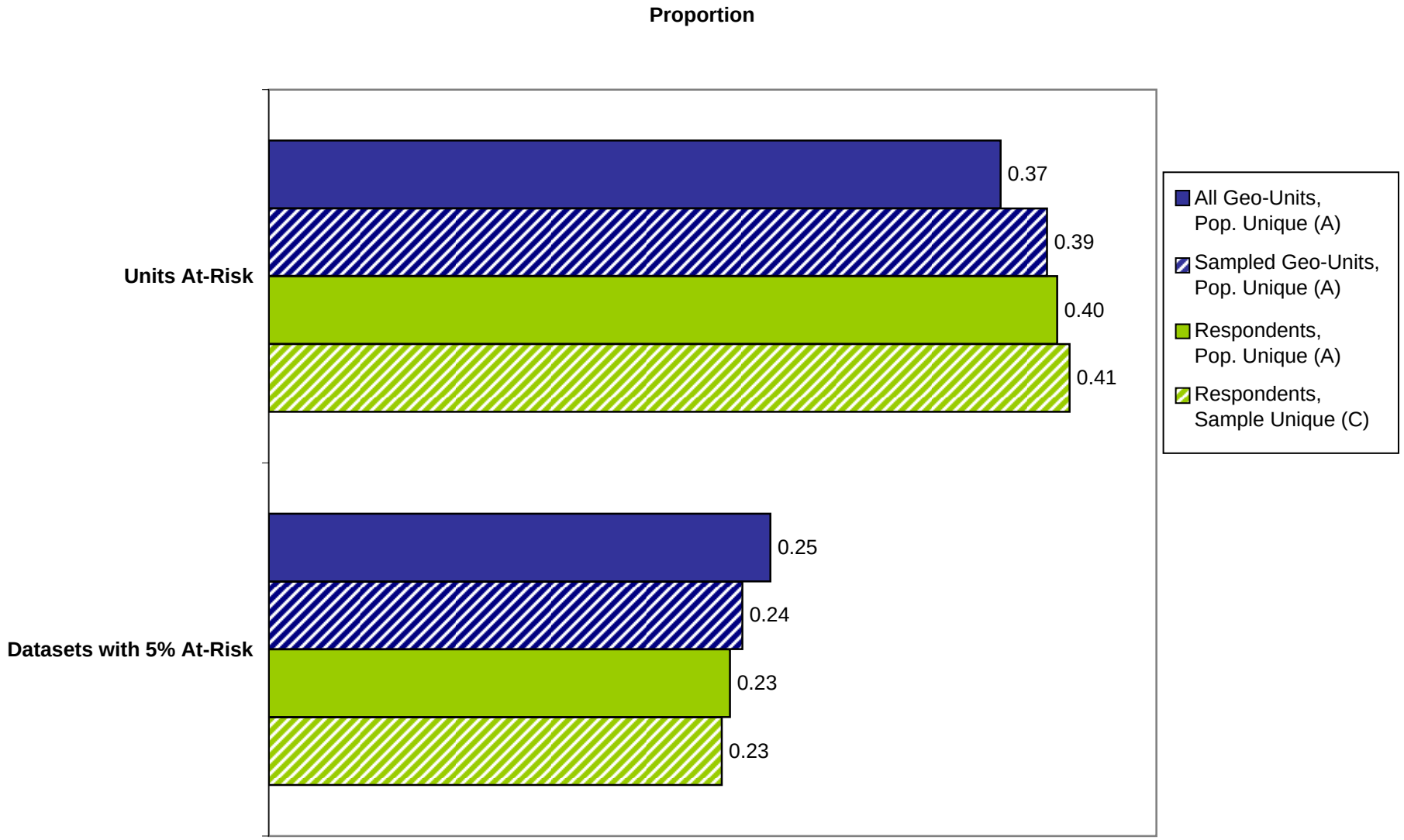

Figure C. Disclosure Risk of Geographic Units and Survey Respondents and Suppression Rates of Contextual Data, By Risk Components of Blockgroup-Level Measures

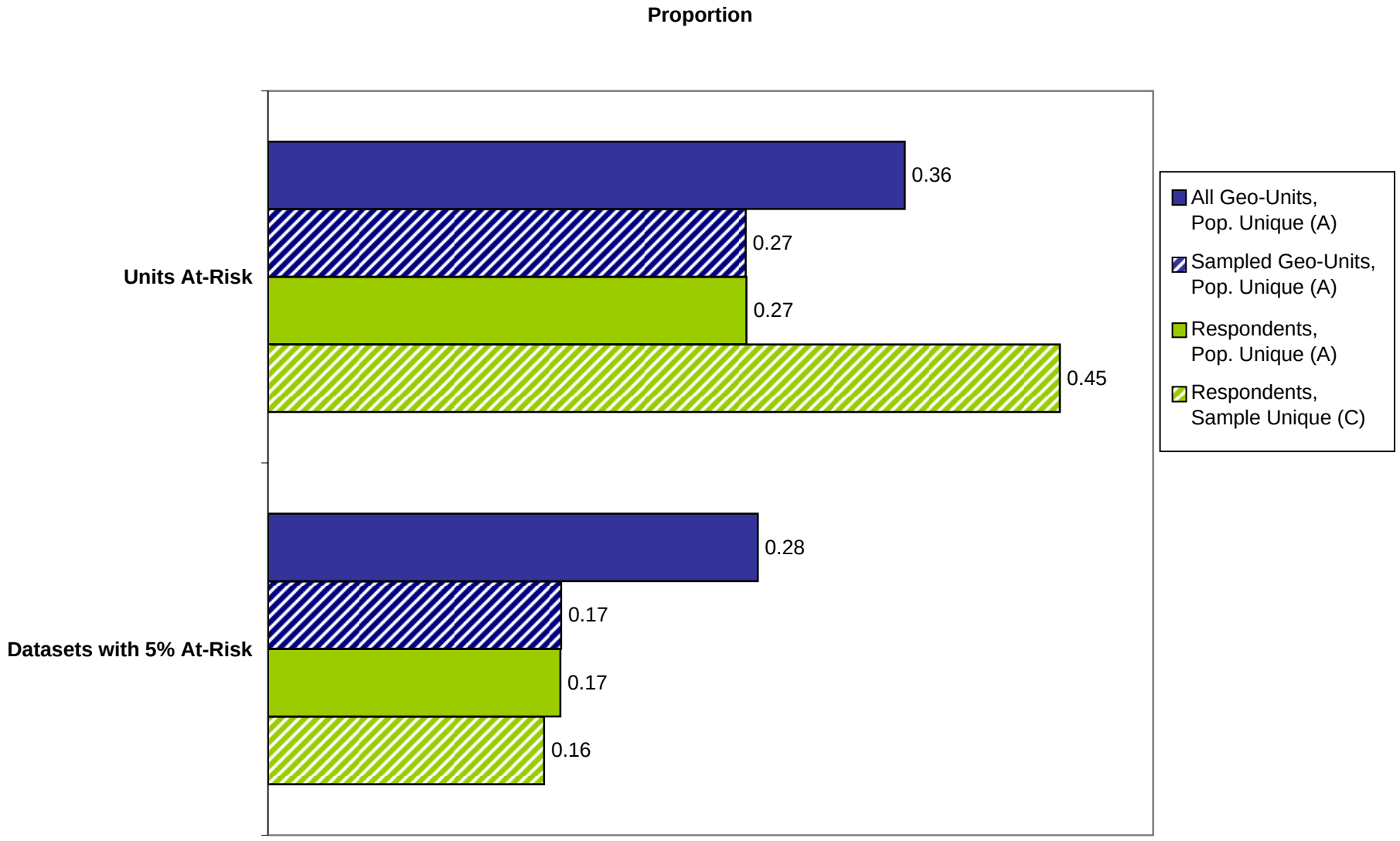

Table A-1: Characteristics of Test Datasets (N=14,796)

Unweighted Averages Across Datasets	Proportion of Datasets
Spatial Scale	
Counties	0.33
Tracts	0.33
Blockgroups	0.33
Geographic Identifiers	
None	0.17
Population Density	0.17
Division	0.17
Division & Population Density	0.17
State	0.17
State & Population Density	0.17
Coarseness	
1%-Categories	0.17
5%-Categories	0.17
10%-Categories	0.17
15%-Categories	0.17
20%-Categories	0.17
Top-25%, Bottom-25%, Other	0.17
Number of Contextual Variables	
1-Key	0.12
2-Keys	0.22
3-Keys	0.22
4-Keys	0.22
5-Keys	0.22
Conceptual Composition	
% Persons, Non-Hispanic White	0.31
% Persons, Non-Hispanic African-American	0.17
% Persons, Non-Hispanic Asian or Pacific Islander	0.07
% Persons, Non-Hispanic Other Race	0.24
% Persons, Hispanic	0.18
% Persons, Foreign-Born	0.17
% Foreign-Born, Naturalized Citizens	0.18
% Households, Linguistically Isolated	0.23
% Persons, In-Poverty	0.19
% Households, With Wage Income	0.18
% Households, Receiving Public Assistance	0.15
% Persons Age 25+, College Degree	0.18
% Families, Female-Headed	0.20
% Persons Age 16-19, Neither Enrolled nor Graduated from High School	0.15
% Housing Units, Owner-Occupied	0.15
% Persons Age 16+, Civilian Labor Force	0.23
% Civilian Labor Force, Unemployed	0.22

Table A-2: Sampling of Synthetic Persons, Resulting Geographic Contexts, and Size of Geographic-Unit Populations

Sampling	Total	Sample	Percent
Synthetic Persons	280,849,847	11,475	0.0041
Geographic-Units			
Blocks	8,199,368	11,475	0.14
Blockgroups	208,235	10,401	4.99
Tracts	65,133	8,877	13.63
Counties	3,140	1,784	56.82
States	50	50	100.00
Population Size	**Average**	**Mininum**	**Maximum**
Counties			
Geographic Identifiers			
None	3,140	3,140	3,140
Population Density	1,047	389	2,294
Division	349	67	618
Division & Population Density	116	15	546
State	63	3	254
State & Population Density	21	0	196
Tracts			
Geographic Identifiers			
None	65,133	65,133	65,133
Population Density	21,711	13,860	36,070
Division	7,237	3,204	11,346
Division & Population Density	2,412	576	7,456
State	1,303	127	7,049
State & Population Density	434	0	5,838
Blockgroups			
Geographic Identifiers			
None	208,235	208,235	208,235
Population Density	69,412	47,713	112,251
Division	23,137	11,025	36,768
Division & Population Density	7,712	1,860	23,494
State	4,165	398	22,133
State & Population Density	1,388	0	17,987

Note: Excludes Washington, DC as well as tracts and blockgroups located in water (i.e., by definition, no populuation possible).

Table A-3: Aggregate Disclosure Risk of Geographic Units in Test Datasets, By Experimental Traits (N=4,932 Datasets At Each Spatial Scale)

Coarseness Based on Absolute Values

Weighted Averages Across Datasets	Proportion At-Risk					
	Counties		Tracts		Blockgroups	
	Mean	(SE)	Mean	(SE)	Mean	(SE)
Total	0.52	(0.01)	0.39	(0.01)	0.27	(0.00)
Geographic Identifiers						
None	0.31	(0.02)	0.25	(0.02)	0.16	(0.01)
Population Density	0.37	(0.02)	0.29	(0.02)	0.20	(0.01)
Division	0.46	(0.02)	0.36	(0.02)	0.24	(0.01)
Division & Population Density	0.57	(0.02)	0.42	(0.02)	0.30	(0.01)
State	0.65	(0.02)	0.49	(0.02)	0.33	(0.01)
State & Population Density	0.78	(0.01)	0.56	(0.02)	0.39	(0.00)
Number of Contextual Variables						
1-Key	0.19	(0.00)	0.04	(0.00)	0.07	(0.00)
2-Keys	0.33	(0.01)	0.16	(0.01)	0.17	(0.01)
3-Keys	0.41	(0.01)	0.26	(0.01)	0.22	(0.01)
4-Keys	0.49	(0.01)	0.34	(0.01)	0.25	(0.01)
5-Keys	0.55	(0.01)	0.42	(0.01)	0.28	(0.01)
Coarseness						
1%-Categories	0.99	(0.00)	0.98	(0.00)	0.36	(0.00)
5%-Categories	0.77	(0.01)	0.66	(0.01)	0.44	(0.01)
10%-Categories	0.53	(0.02)	0.36	(0.01)	0.32	(0.01)
15%-Categories	0.39	(0.02)	0.21	(0.01)	0.23	(0.01)
20%-Categories	0.29	(0.01)	0.12	(0.01)	0.17	(0.01)
Top-25%, Bottom-25%, Other	0.17	(0.01)	0.04	(0.00)	0.10	(0.01)

Coarseness Based on Percentiles

Weighted Averages Across Datasets	Proportion At-Risk					
	Counties		Tracts		Blockgroups	
	Mean	(SE)	Mean	(SE)	Mean	(SE)
Total	0.94	(0.00)	0.75	(0.01)	0.35	(0.00)
Geographic Identifiers						
None	0.84	(0.02)	0.56	(0.02)	0.27	(0.02)
Population Density	0.88	(0.01)	0.65	(0.02)	0.30	(0.01)
Division	0.94	(0.01)	0.75	(0.02)	0.34	(0.01)
Division & Population Density	0.97	(0.00)	0.81	(0.02)	0.38	(0.01)
State	0.99	(0.00)	0.86	(0.01)	0.41	(0.01)
State & Population Density	1.00	(0.00)	0.89	(0.01)	0.41	(0.00)
Number of Contextual Variables						
1-Key	0.39	(0.00)	0.07	(0.00)	0.10	(0.00)
2-Keys	0.69	(0.01)	0.33	(0.01)	0.25	(0.01)
3-Keys	0.85	(0.01)	0.54	(0.01)	0.32	(0.01)
4-Keys	0.92	(0.01)	0.69	(0.01)	0.35	(0.01)
5-Keys	0.95	(0.00)	0.79	(0.01)	0.36	(0.01)
Coarseness						
1%-Categories	1.00	(0.00)	1.00	(0.00)	0.33	(0.00)
5%-Categories	1.00	(0.00)	0.99	(0.00)	0.41	(0.00)
10%-Categories	1.00	(0.00)	0.93	(0.00)	0.53	(0.01)
15%-Categories	0.99	(0.00)	0.80	(0.01)	0.17	(0.01)
20%-Categories	0.96	(0.00)	0.60	(0.02)	0.45	(0.01)
Top-25%, Bottom-25%, Other	0.68	(0.02)	0.19	(0.01)	0.21	(0.01)

Note: Assume there is risk of disclosure when there are fewer than 20 matches (i.e., based on reidentification confidence-level of p=.05).

Note: Standard errors (in parentheses) are adjusted to account for the complex survey design of sampled variables sets.