ICPSR
A PARTNER IN SOCIAL SCIENCE RESEARCH
INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH
UNIVERSITY OF MICHIGAN

Working Paper

Finding a Needle in a Haystack: The Theoretical and Empirical Foundations of Assessing Disclosure Risk for Contextualized Microdata

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

ICPSR Working Paper Series
Working Paper No. 4

June 2008

Finding a Needle in a Haystack: The Theoretical and Empirical Foundations of Assessing Disclosure Risk for Contextualized Microdata

Kristine M. Witkowski

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

Contextualized microdata are one way to safely release geographic data without identifying the location of survey respondents. This study informs the design of such datafiles with its needle-in-haystack approach to disclosure and its discussion of associated methodological concerns. Drawing a sample of counties, tracts, and blockgroups, I illustrate how the reidentification of individuals is shaped by aggregating geographies into look-alike sets. I detail the complexity of reidentification patterns by assessing the likelihood that young adult white and black males would be pinpointed within reconstituted haystacks given: (1) the size of the total population of aggregated contexts; (2) the amount of error in population counts; and (3) differential search costs stemming from spatially-dispersed contexts.

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 and Felicia LeClere for their 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.

Finding a Needle in a Haystack: The Theoretical and Empirical Foundations of Assessing Disclosure Risk for Contextualized Microdata

Kristine M. Witkowski

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

Contextualized microdata are one way to safely release geographic data without identifying the location of survey respondents. This study informs the design of such datafiles with its needle-in-haystack approach to disclosure and its discussion of associated methodological concerns. Drawing a sample of counties, tracts, and blockgroups, I illustrate how the reidentification of individuals is shaped by aggregating geographies into look-alike sets. I detail the complexity of reidentification patterns by assessing the likelihood that young adult white and black males would be pinpointed within reconstituted haystacks given: (1) the size of the total population of aggregated contexts; (2) the amount of error in population counts; and (3) differential search costs stemming from spatially-dispersed contexts.

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 and Felicia LeClere for their 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.

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, Zayatz, and Hoel 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; Energy Information Administration 2001). 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 2008), I have conducted reidentification experiments to assess how easy it is to pinpoint locations within the total population of counties, tracts, and blockgroups. 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. 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 with its needle-in-haystack approach to disclosure and discussion of associated methodological concerns.

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 of counties, tracts, and blockgroups. Utilizing a synthetic test datafile, I illustrate how the reidentification of individuals is shaped by aggregating geographies into look-alike sets. Furthermore, I assess one dimension of risk associated with contextualized microdata, that of identifying survey respondents whose personal characteristics (i.e., sex, age, and race) are rarely found among populations sharing the same contexts. Using geographic-units and their "aggregated look-alike contexts" as my units of analysis, the number of persons in the population with a distinct set of personal characteristics as the outcome of interest, and indicators of different factors that underlie the definition of risk, I detail the complexity of reidentification patterns by assessing the likelihood that young adult white and black males would be pinpointed within reconstituted haystacks given: (1) the size of the total population of aggregated contexts; (2) the amount of error in population counts; and (3) differential search costs stemming from spatially-dispersed contexts.

2. Empirical Approach to Disclosure Analyses of Contextualized Microdata

In this section, I outline the analytical steps involved in evaluating disclosure risk for contextualized microdata. In doing so, I describe my approach to defining and assessing risk as well as some of the methodological concerns involved.

2. a. The Role of the Intruder

As a first step of a risk assessment, one must consider the roles, motivations, and resources of would-be intruders and how these characteristics influence search behavior. An intruder may be an absolute stranger to a respondent, or they may be a family member or an acquaintance. While these are examples of individuals, institutional actors, such as insurance companies, employers, or government officials, may also play the role of intruder. An intruder's target of disclosure varies with their underlying motives,

consisting of either a person known to them, any person within a database reporting compromising information, or a group of individuals sharing conspicuous characteristics. Given their diversity of roles, intruders have access to a variety of resources used in their quest for confidential information, influencing their method of reidentification.

Formulating a disclosure scenario that is of general concern to data producers, I consider an intruder with high-levels of resources and empirical expertise, but who lacks knowledge of any particular survey respondent. This type of intruder begins her search by obtaining a public-use datafile containing sensitive information collected by a national survey as well as contextual data describing each respondent's county of residence. Assuming she is looking for a target of extortion, an intruder searches the database for persons reporting information that is potentially harmful if widely known (e.g., crime, health problem). Reviewing the datafile, she finds several such respondents who she is considering for reidentification. But to provide a concrete example, I limit my discussion to one target respondent who is a 20-year-old, non-Hispanic, white male living in a metropolitan county with fewer than 1 million persons, having contextual characteristics described as: (1) 70 to 79.9% Persons, Non-Hispanic White; (2) 0 to 9.9% Persons, Foreign-Born; (3) 10 to 19.9% Persons, In-Poverty; (4) 60 to 69.9% Housing Units, Owner-Occupied; and (5) 0 to 9.9% Civilian Labor Force, Unemployed. Unknown to her, this man lives in Peoria County, Illinois, along with approximately 1,100 of his "twins" (i.e., subpopulation members).

Since she does not know her target's identity, the intruder needs to pinpoint the name and location of this respondent. To do so, she must locate a list of names and addresses of people residing within counties sharing the above contextual characteristics. Besides encompassing all counties of interest, this identifying extant file must also contain personal attribute information, such as age, sex, and race, so that specific names and addresses may be linked to the survey respondent. If attribute data is inaccurate or if a file lacks information for a significant proportion of the geographic units and/or their populations, she may mistakenly assign identifiers to the survey respondent who may or may not exist within the extant file. Grappling with measurement and coverage error, the intruder finds several records in the extant file with the same demographic characteristics as her target. Having to select among numerous records, *n*, that resemble her target, the intruder has a one-in-*n* chance of assigning the wrong identifiers.

The intruder may pay a hefty price – ranging from embarrassment to costly lawsuits to imprisonment – if she acts on an erroneous assumption that she has correctly identified the survey respondent. Hence she spends a considerable amount of effort gathering information to supplement and verify her list of possible matches, increasing computation, transaction, and/or ground costs. If the cost or risk assumed by the intruder is sufficiently high, she is likely to give up her search and to dismiss her extortion plans; and as a result, the risk of disclosure can then be considered negligible to the respondent. Given this scenario, I now detail my definition of risk as it relates to contextualized microdata.

2. b. The Definition of Risk

To help explain the conceptual underpinnings of anonymized data, I rely on the English idiom of the "needle in a haystack", which refers to an item that is difficult to find because it is hidden within a larger set of objects (Cambridge University Press 2003). A survey respondent, or the "needle", is a specific person within a public-use datafile who has a particular set of individual traits that are easily ascertained by an intruder, such as their sex, age, race, ethnicity, marital status, and education-level. This survey respondent is also a member of a group, or "haystack", of persons in the population sharing the same identifying individual and contextual characteristics. As with finding a needle in a haystack, the chance that a respondent is correctly reidentified is considered faint when an intruder must search among a sufficiently large number of persons sharing the same characteristics within an identifying extant file (Citation).

Underlying my needle-in-haystack approach to disclosure is an intruder behavioral model that considers both uncertainty and cost. Building upon a statistical inference argument, we can refute an intruder's

assertion that she has reidentified a survey respondent, by certifying that the chances of her being correct is no better than if she chose a person at random from the population (VanWey, et al. 2005). Evidence is gathered by estimating the cost associated with pinpointing a survey respondent among members of a subpopulation within a specified geographic area. The producer must then decide upon a cost-threshold that defines the upper bound of risk. If estimated search costs fall below this threshold, then the risk of reidentification is considered intolerable and therefore a respondent's contextualized microdata may not be safely released.

$$\text{Search Cost for Respondent}_{i,j} = \sum \left(\text{Number of Straws}_{I,J} \times \text{Search Cost for Straw}_{I,J} \right) \quad (1)$$

$$\text{Number of Straws}_{I,J} = f\left(\text{Number of Shared Keys, Coarseness of Measures, Response Error of Survey and Extant Data} \right) \quad (2)$$

$$\text{Search Cost for Straw}_{I,J} = f\left(\text{Access to Extant File, Coverage in Extant File, Spatial Dispersion of Haystack} \right) \quad (3)$$

The fundamental component of search costs is the number of members in a subpopulation, or the size of a haystack. Each member of the subpopulation, or "straw", imposes a certain amount of effort from an intruder, requiring the compilation of identifying information and the determination that a chosen name and address most likely belongs to a target respondent. Providing the direct identifiers for large proportions of the population, several commercial databases (e.g., Experian, ChoicePoint, KnowX) exist that may significantly lower search costs. Having gathered information from a variety of public sources, such as departments of motor vehicles, real estate purchase records, and voter registrations, these extant files offer an easy way to link names and addresses to survey respondents. However, their usefulness in reidentification is determined by whether an intruder can use the provided information to accurately limit the number of matches while minimizing supplemental search costs.

If an intruder has the name and address of a survey respondent, the cost of accessing a commercial database is minimal since a single record can be readily extracted using the file's search engine and typing in the known information. But access costs increase considerably when the intruder is on a fishing expedition. Not knowing a name or an address, the intruder searches the database for persons having a specified set of characteristics that can be associated with a record in a survey datafile. Given this reidentification approach, the extraction of records from the commercial database may be quite cumbersome and costly. But access costs may be reduced if an extant file has accurate and detailed data that overlaps with information provided by a survey. When both a survey and a commercial database have detailed age, race, and geographic information, the intruder can better refine the list of potential matches for her target respondent, thus reducing uncertainty and search costs (cite Raghu).

Even when sufficient access to extant files containing useful linkage information is secured, the intruder will likely have to address coverage error (cite Raghu). An error in coverage occurs when there is an omission, duplication or wrongful inclusion of persons represented in the population (Robinson, et al 1993). Assessing the amount of undercoverage, the intruder tallies population counts from the decennial census to estimate the number of haystack members. She then compares the number of nonduplicate records within an extant file, estimating the likelihood that a respondent is not represented in the identifying file. However calculations are not straightforward. Census population counts are also subject to undercounting and overcounting (Robinson, et al 1993). Consequently an intruder must refine her assessment of uncertainty and cost by adjusting census counts to address the spatial distribution of hard-to-count and double-counted populations.

With this said, if there is a significant chance that a respondent is incorrectly reidentified, then the intruder may choose to pursue alternative datasets and search methods. Additional computational costs may be minimal when incorporating electronic data using existing (albeit modified) processing systems. But if an intruder must review several datafiles produced and distributed by different state and local agencies, associated search costs can quickly increase. Furthermore information that is only available in hard copy

or from direct observation also represent a significant increase in labor and travel costs since these data must be collected on-site. In turn, I suspect that the cost of supplemental search activities may be positively associated with the spatial dispersion of haystacks, following a rudimentary function of the number of geographies where a target respondent may possibly be located and the areal size and distance between each of these locations.

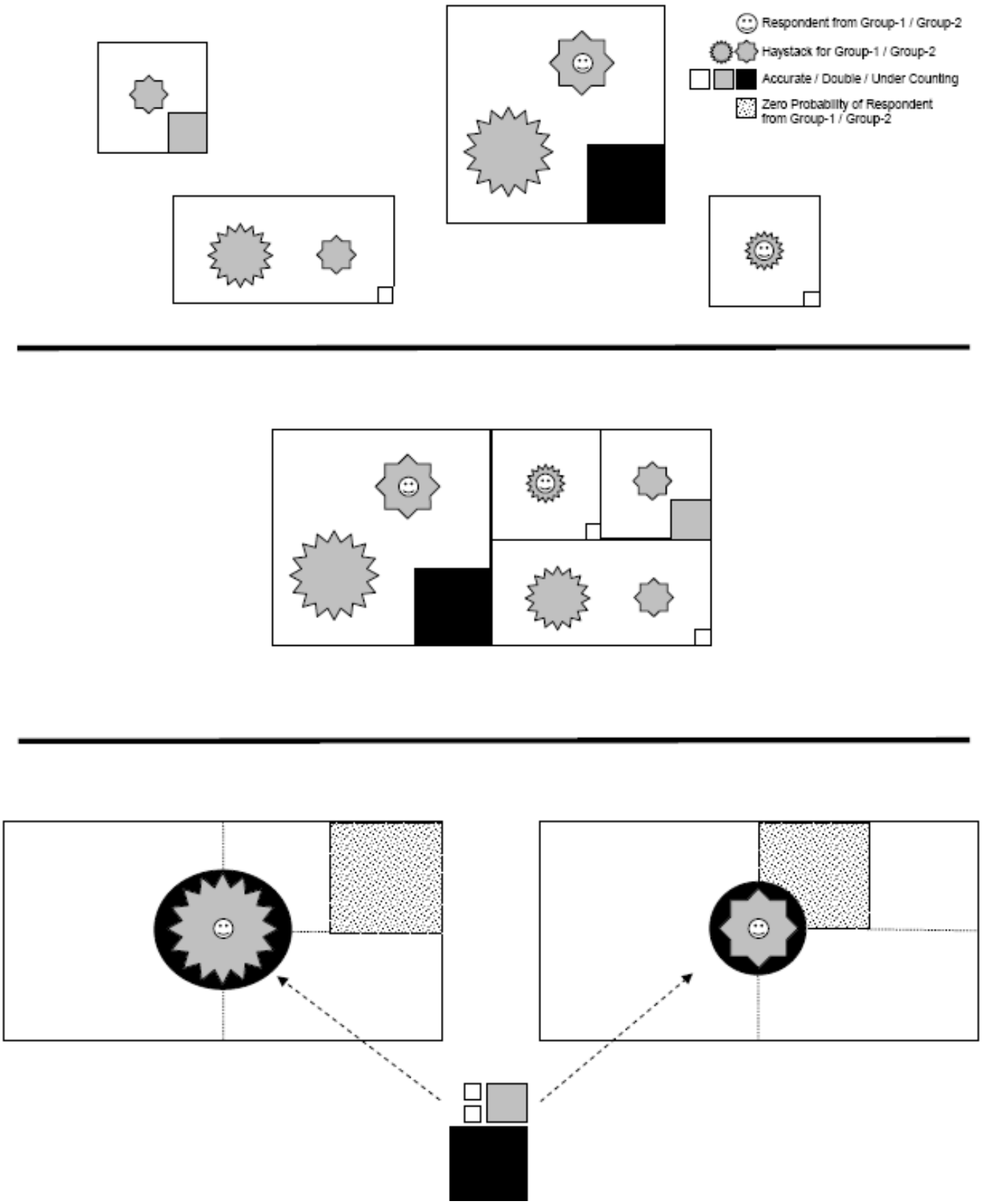
Given the above factors that define the cost and ultimately determine the risk associated with contextualized microdata, I now describe the empirical steps involved in conducting a reidentification experiment that utilizes this type of data.

2. c. Laying the Groundwork for the Search

Trying to meet user demands for geographically-rich microdata, a producer can either directly identify the location of respondents or attach contextual data to these records. When a dataset identifies geography, an intruder's search is limited to a single known location, heightening the chances that a target is correctly reidentified. Releasing only contextual data reduces this risk because search costs are likely to grow with the size of haystacks which are aggregated across a spatially dispersed set of "look-alike" locations (i.e., geographic units sharing the same contextual characteristics). As illustrated in Figure A and further described in this section, an intruder must consider a complexity of factors and information when reidentifying respondents in contextualized microdata files.

Knowing the contextual characteristics of a target respondent, the intruder's first task is to compile a list of counties having the same characteristics. The survey database may also provide geographic information, such as the region, state, or population density of respondent locations. If so, the compilation of geographic units is limited to these geographic areas. As the original source of contextual information attached to a survey datafile, widely-available summary files of census data identify all counties, tracts, and blockgroups (as well as other types of spatial units) in the United States and provide measures describing the characteristics of the population located within these geographies. Searching within the survey file, the intruder identifies her target respondent's contextual characteristics (listed above), which describe a particular county in a metropolitan area with fewer than 1 million persons. Using the same contextual indicators, she then identifies counties in the summary file that have similar measurement values (located in MSAs with less than 1 million persons). I purposely use the term "similar" since the intruder should be aware that the contextual information provided in the survey's public-use file may have been perturbed to remove any one-to-one correspondence with the original tabulation files for the full population of geographies. Adding to this confusion is the Census Bureau's policy of directly introducing noise into the original tabulations; whereby they swap household records between neighboring geographies when a location has an extremely small number of persons with a specific set of characteristics. Compiling the names of these look-alike counties, the intruder now has a list of 22 locations where a respondent may live.

Figure A. Constructing Aggregated Contexts and Measuring their Disclosure Characteristics



For her second task, the intruder must then assess the probable amounts of uncertainty and cost that she will face in her effort to reidentify a respondent within this set of geographies' populations. Utilizing the same source of her contextual measures, decennial census tabulations – particularly those derived from the 100 percent count of the population – indicate the size of haystacks as they vary across space. By virtually amassing populations from the 22 counties resembling her target's location, the intruder must, in effect, search within an aggregated haystack consisting of 20,814 young-adult white males emerging from a compiled population of 4,422,343 persons.

However this assessment of haystack size may be inaccurate. The U.S. Census Bureau has estimated that significant proportions of males age 18-49, renters (i.e., non-owners), and non-Hispanic blacks were uncounted in the decennial census, with national net rates ranging from 1.12% to 1.84%. Furthermore the Bureau found that some demographic groups were actually over-counted by .60% to 2.53%, namely non-Hispanic whites, children age 10-17, females age 18 and over, and males age 50 and over. Besides these demographic patterns, enumeration error also varies with the size of a geography's population. Estimates of census coverage for small geographic areas indicate that less populated places (i.e., less than 100,000 persons) have predominately overcounted populations, with error rates as high as 6%. However large places (i.e., 100,000 or more persons) are more likely to have their populations undercounted, reflecting the concentration of hard-to-count populations in large urban areas (U.S. Census Bureau 2003a).

The amount of coverage error associated with a particular subpopulation within a particular geographic area influences the amount of intruder uncertainty and cost in two ways. First of all, error in estimates of haystack size increases uncertainty for an intruder since she cannot determine the precise number of needles that she will need to identify and verify. Dropping persons who are double counted and contemplating those who are uncounted within an area, reconstituted haystacks may offer significantly more or less anonymity, depending on the personal characteristics and geographic location of the survey respondent. Hence, an intruder should consider the spatial and demographic patterns in coverage error when designing her reidentification strategy. But whether or not the size of a haystack has been over- or under- estimated, large amounts of coverage error should generally increase skepticism about the accuracy of reidentifications; thereby lowering risk.

The difficulty of the decennial census to enumerate the U.S. population brings forth a second issue affecting disclosure risk. The 5.8 billions of dollars and the long-established infrastructure devoted to the 2000 census implies that other external databases (derived from relatively limited resources) are likely to be subject to even higher levels of coverage error. As Raghu (2008) has illustrated, error in extant data may be most effective in thwarting intruders. Consequently patterns of coverage error identified in the decennial census may also point to groups of persons and locations that are most likely to be missing from or double-counted within identifying extant files.

In planning for the 2010 decennial census, the U.S. Census Bureau has conducted research characterizing all U.S. tracts in terms of how difficult it may be to enumerate their populations. Each tract is assigned a composite score reflecting population and housing attributes associated with mail nonresponse (Bruce and Robinson 2003). Twelve variables are used in the creation of hard-to-count scores, consisting of: (1) % Housing Units, Vacant; (2) % Housing Units, Not Single Detached or Attached Units; (3) % Housing Units, Renter-Occupied; (4) % Housing Units, More than 1.5 Persons per Room; (5) % Households, Not Husband/Wife Families; (6) % Occupied Housing Units, No Telephone Service; (7) % Persons Age 25+, Not High School Graduate; (8) % Persons, In-Poverty; (9) % Households, Public Assistance Income; (10) % Persons, Unemployed; (11) % Households, Linguistically Isolated; and (12) % Occupied Housing Units, Householder Moved into Unit 1999-2000. Readily available from the Census website, the intruder analyzes these data to get a sense of whether a particular county is susceptible to undercounting or overcounting; thereby providing a rough assessment of the accuracy of its population counts. Taking the first and third quartiles of the full distribution of hard-to-count scores (ranging from 0 to 132), she assigns tracts into two categories whose populations are: (1) most likely to be double-counted (i.e., scores below Q_1=9); and (2) most likely to be undercounted (i.e., scores above

Q_3=52). For all counties of interest, she then calculates the proportion of the total population living in tracts that are most likely undercounted and double-counted.

Analyzing these data, she finds that 27.1% of the population within this set of look-alike counties lives in extremely "hard-to-count" tracts; while 24.7% lives in tracts that are most likely double-counted. As indicated by national estimates, approximately 1.12% of males age 18-49 are undercounted, while 1.13% of non-Hispanic whites are overcounted (Tech.Assess. ACE Revision II, 2003). Given the relatively balanced sources of coverage error, there may be a good chance that the estimate of this haystack's size is fairly accurate, consisting of an aggregated 20,814 young-adult white males.

However there is one important caveat for the above measures of coverage error: their inability to capture race and ethnic differences in enumeration difficulty within locations. Disproportionate numbers of minorities exhibit characteristics that are associated with being undercounted; while members of the majority population are more likely to have characteristics associated with being double-counted. For instance, native non-Hispanic whites are most likely of all racial groups to report owning a home in the 2002 Current Population Survey. While 75.0% of this majority population are homeowners, only 22.2 to 70.3% of minorities held this asset (http://www.census.gov/prod/2003pubs/h121-03-1.pdf). If an intruder does not account for these compositional differences among subpopulations residing in the same area, she may erroneously assume that the amount of coverage error estimated from the total population equally applies to every haystack therein.

The proportion of occupied housing units that are rented is one of the key correlates of hard-to-count scores characterizing each tract's full population. Subpopulations having significantly higher rates of homeownership, compared to the tract's population as a whole, indicate groups who are relatively easier to reidentify given their ability to acquire stable housing. One such example is found in newly gentrifying areas, where a small number of affluent persons have purchased homes in a high-poverty, and therefore, highly undercounted area. Conversely subpopulations with dramatically higher rental rates may be more difficult to enumerate than what is indicated by the general population. Some examples of relatively conspicuous but hard-to-count housing situations are rental units catering to low income, student, and other transient populations that are positioned within wealthy areas.

Using census tabulations of subpopulations in occupied housing units with different tenure status, an intruder can assess whether a particular racial group is more or less likely to be living in rental housing units, as compared to the rest of the population in the area. Analyzing tract-level rental rates of persons residing in the 22 look-alike counties of interest, the intruder finds that, on average, 17.3% fewer whites are renters (compared to 45.2% of the non-white population who are renting). This relatively low rate of rentership, juxtaposed to the general population's low hard-to-count scores, indicates a heightened chance of whites being double-counted. Hence the intruder may conclude that the actual number of young-adult white males is considerably lower than the estimated 20,814. She may also infer that identifying extant data will likely cover a large proportion of this subpopulation.

As illustrated above, these summary indicators draw attention to haystacks whose sizes are likely in need of adjustment. They also provide a rough indication of the sources and relative amounts of coverage error that an intruder will face in her efforts to link identifying information from extant files. A large proportion of an aggregated context's population that is likely overcounted indicates that an intruder may have to locate and delete a sizeable number of duplicate records within extant files. Furthermore a large proportion of a population that is likely undercounted points to the heightened possibility that significant amounts of identifying information will be missing. More research is needed to assess the degree to which coverage error influences the process of reidentification by translating this information into accurate estimates of hidden and overblown haystacks and missing extant information. Paramount to this work is the refinement of undercount and overcount measures.

Finishing my description of how one collects and uses different types of data in this reidentification process, I now discuss the factors involved in setting priorities in the quest for confidential survey information.

2. d. Setting Priorities in the Search

If aggregated haystacks or the amounts of missing data are expected to be exceptionally large, the intruder may decide to gather supplemental information, decreasing levels of uncertainty while driving up her search costs. More research is needed to accurately predict intruder behavior and to estimate search costs. But if the observed size and spatial dispersion of this scenario's haystack is any indication, potential costs may be considerable since the intruder needs to search for her target respondent, the proverbial needle, within a haystack composed of 21,000 straws which is dispersed across 22 counties and 12 states, covering 16,500 square miles of land area.

Given these search parameters, the intruder may decide that the most thorough approach to reidentification – that of hunting within all look-alike counties – is too daunting. However the intruder may significantly reduce her efforts by ignoring counties that are unlikely to have respondents drawn from the area. If the probability of a target respondent being drawn from a county is nearly zero, the intruder may take a calculated risk and assume that a particular geographic unit was not included in the survey's sampling frame. Hence the chances of a geography being searched rises with the likelihood that a respondent is drawn from the area. This prioritization scheme has the tendency to increase the size of haystacks that are most likely examined. Ironically, a respondent concealed by a thin haystack may actually be less likely to be reidentified if we assume that an intruder has taken this search approach. As for the exemplar scenario, the intruder cannot safely ignore any of the 22 look-alike counties since at least 192 young-adult male exists within each location.

But in one last effort to limit the set of look-alike geographies, an intruder can return to the public-use datafile to see if more reidentifying information may be gleaned from weights or other variables used in the administration of a survey (e.g., codes for field interview regions). Identifying records sharing the same primary sampling units (although these areas are not directly named), the intruder can ascertain whether a set of respondents, sharing the same contexts, are constrained to a single sampling unit or are distributed across more than one location. In addition, respondent characteristics may also be used to target look-alike geographies. A new set of approximated contextual measures can be derived from the personal characteristics of survey respondents within primary sampling units or shared contexts. Research by Ragunathan, et al. (2007; need citation) has shown that an intruder may be quite successful in correctly reidentify geographies when using approximated contextual data.

Laying the groundwork by gathering and analyzing information as described above, the intruder gains a sense of the magnitude of her search effort. Even though I discuss a scenario where an intruder attempts to reidentify a single respondent, the intruder will likely review materials for a large number of targets. So before fully launching her search, she must bring to light the different issues she will likely face. Hence, in the next section, I discuss underlying factors that shape levels of uncertainty and cost for the intruder.

2. e. Factors Shaping Dimensions of the Search Effort

The Foundation of Haystacks: Composition and Size of Populations and the Scale of Geography

Defined by a set of linkage variables found in a survey and identifying extant files, the number of people resembling a target respondent, or the absolute size of haystacks, fundamentally depends on the composition and size of an area's population that are a function of the scale of geography. When a national survey suppresses all geographic information in its microdata file, intruders must search the whole country for target respondents, limiting their efforts to small subpopulations that are accurately reidentified within extant files. However data producers, in their efforts to provide geographically-rich microdata, may choose to directly identify the locations of respondents. Carving up finite space into units

of various sizes influences disclosure risk by setting the composition and number of persons within geographic boundaries that are sub-national in scale.

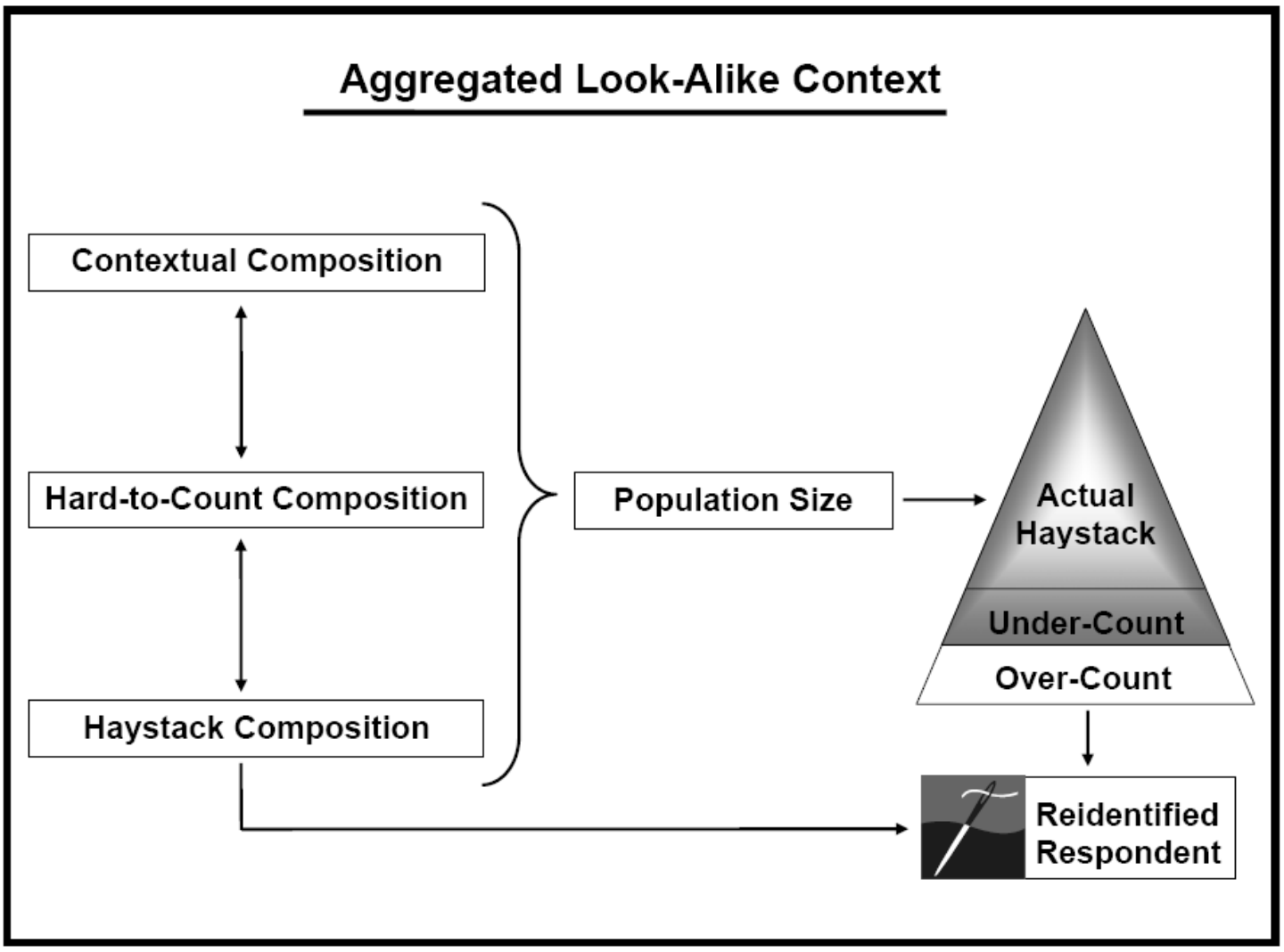
The composition of a population reflects the ways in which its members are distributed according to age, ethnic or racial category, and sex. The spatial distribution of individuals with particular sets of characteristics emerges from a complexity of factors that underlie residential segregation patterns. The sociological and demographic literature has found, for instance, that the commingling and concentration of racial groups is related to an area's labor and housing markets, social networks, and discriminatory barriers which differentially attract and retain majority and minority populations (citations, Annual Review of Sociology, Demography, others?). Geographic units characterized by high levels of racial segregation produce relatively small haystacks for uncommon populations. While U.S. minority and majority groups are prevalently defined by race and ethnicity, the reidentification of individuals, and therefore the construction of haystacks, is not limited to this single compositional element. Compounding the disclosure effects of racial segregation, intruders are likely to consider other identifying personal characteristics, such as the age, sex, and education of respondents. There is considerable socioeconomic variation within race and ethnic groups that influences residential patterns and, subsequently, the absolute size of haystacks. Holding constant the size of the total population, the absolute number of members declines when subgroups are explicitly defined. Consequently disclosure risk is heightened when haystacks are delineated by a variety of detailed personal characteristics. This relationship between the specificity and size of haystacks exists regardless of compositional shifts resulting from the scale of geography.

Given the underlying composition of the U.S. population as it is spatially dispersed, let us consider how the size of an area's population and the scale of its geography influence the construction of haystacks. Holding constant the composition of a population, the absolute number of subgroup members, by definition, declines with the total size of the population. Consequently disclosure risk should rise when the population is constrained into smaller and smaller geographic units.

There are dramatic differences in the absolute numbers of counties, tracts, and blockgroups as well as the size of populations within their borders. Compared to the number of counties, there are 20.7 times and 66.3 times as many tracts and blockgroups (3,141 counties or county-equivalents, 65,174 tracts, and 208,125 blockgroups). Variation in the number of geographic units results from the methodology underlying the construction of these administrative units which places a cap on their population size. Having population sizes ranging between 67 to 9,519,338 people, counties are entities that have been defined legally; that is, they are created by State law or some other administrative action. However census tracts and blockgroups have been defined specifically for data collection purposes. Census tracts designate areas that are relatively uniform in their population characteristics, economic status, and living conditions, with as many as 1,500 to 8,000 people. Composed of 600 to 3,000 people, blockgroups further subdivide tracts into areas bounded by visible and legal features (e.g., streets, property lines) (U.S. Census Bureau 2000).

The limited number of people within rural counties, tracts, and blockgroups precludes the safe release of their identifiers, representing an important barrier to providing rich information about relatively unpopulated areas and small-scale environments. One viable solution is to release the contextual characteristics of respondent locations instead of directly identifying geographies. In doing so, small populations within look-alike geographic units are virtually aggregated into a larger pool of people, who are parceled out into haystacks of increased size and diversity. If an aggregated context reaches a sufficient population size, it is likely that emergent haystacks will also be large enough to obscure the identities of respondents. Controlling for population distributions within individual geographies, aggregated contexts are likely to achieve a large base population (along with anonymizing haystacks therein) when they are derived from a high number of look-alike units. As with haystacks, disclosure risk is heightened when aggregated contexts are derived from an explicit set of contextual characteristics. As described by Witkowski (2008), the relationship between the specificity and size of aggregated contexts varies with the scale of geography.

Figure B. Population Composition of Aggregated Context and the Absolute Size of Emergent Haystacks



The U.S. Census Bureau has set disclosure standards allowing for the release of geographic information onto microdata files for areas with 100,000 or more persons (5% sample, Public Use Microdata Area, PUMA) and 400,000 or more persons (1% sample, super-PUMA) (U.S. Census Bureau 2003b). Similar standards have been set for surveys collecting more detailed data from even smaller samples. For instance, the Current Population Survey (CPS) identifies metropolitan areas and counties with 100,000 or more persons, while the Survey of Income and Program Participation (SIPP) identifies states and metropolitan areas with populations over 250,000 (U.S. Census Bureau 2001). The SIPP's panel design collects relatively more detailed information than the cross-sectional CPS. In turn, established guidelines indicate the important role of survey design and sampling methodology in determining population-size thresholds, where higher thresholds are necessary to offset risk that is heightened by complex surveys and elevated sampling rates.

Starting in 1970 with their construction of county groups (http://usa.ipums.org/usa/), the Census Bureau has gathered a formidable amount of knowledge and experience in carefully demarcating areas for identification in public-use microdata files. However much of the justification of population-threshold values has not been published. The lack of publicly-available documentation is expected since governmental policy prohibits the release of detailed information about disclosure practices. While this policy is understandable, its application has buried valuable scientific insights. However my work is not confined by a real set of survey respondents; and so I am able to elaborate on the empirical

underpinnings of these standards and assess whether they may be extended to contextual information, answering such questions as: What determines whether the population within an aggregated context will produce sufficiently large haystacks for both majority and minority populations? If a haystack is too small, what additional factors operate to reduce disclosure risk?

The Premise Underlying Population-Size Thresholds and the Convergence of Three Population Compositions in Aggregated Contexts

An important premise of population-size thresholds is the belief that (1) a large base population will contain sufficient numbers of members of any subpopulation that: (2) will likely be drawn into a survey as a respondent and (3) will likely be represented in identifying extant files. Breaking down this premise into three discrete parts, we can better understand the complexity of protection offered by aggregated contexts resulting from the convergence of a triad of population compositions.

In the initial stage of a reidentification experiment, I assume that an intruder will rely on three sets of compositional variables to assess the difficulty of pinpointing survey respondents. The gender, age, and race/ethnicity of survey respondents define the composition of haystacks where these "needles" are hidden. Enriching microdata files with geographic information, variables describing the contexts of unknown surveyed locations (i.e., % Persons, Foreign-Born; metropolitan status) further constrain where respondents may be found. A third set of composition variables, in the form of hard-to-count scores and racially-specific rentership rates, are used to assess the accuracy of estimated haystack sizes and the likely amounts of coverage error in extant files. The correlation between these three sets of contextual measures is considerable. It is this measurement overlap that determines the shift in factors contributing to the correct reidentification of respondents, reflecting the degree that an aggregated context is residentially segregated, characteristically unique, and difficult to enumerate.

A unique geographic unit, sharing contextual characteristics with only a few locations, typically has extreme measurement values. This location's outlying contextual characteristics indicate that residential segregation may be considerable, resulting in small haystacks for uncommon subpopulations. For instance, an area with an extremely large foreign-born population is likely to contain a small haystack of non-Hispanic whites. The ability to offset the risk of this unique location's small haystack is limited since there is only one other geographic unit with the same characteristics. Hence it is unlikely that we can sufficiently increase the size of this non-Hispanic white haystack because of the small number of additional straws gathered from the single look-alike unit.

In turn, we see that a selection process operates in the creation of small haystacks from unique contexts. Compensating for reduced levels of uncertainty, this same process brings rise to contexts with inordinately high levels of coverage error. Contextually-unique geographic units also exhibit extreme hard-to-count characteristics. Continuing with my example, foreign-born populations are typically difficult to enumerate since they have relatively high rates of rentership. Non-Hispanic whites within the same location may exhibit similar hard-to-count characteristics as the majority foreign-born population. If this is true, the actual size of white haystacks may be considerably larger; and it is likely that this subpopulation is also underrepresented in identifying extant files. Support for this conclusion is found when (1) white and foreign-born rentership rates mirror each other and (2) the base population is extensive, suggesting that a substantial white population may still be hidden.

But let's consider the alternative case when an aggregated context has a small base population, derived from a few unpopulated geographic units. Regardless of the population's enumeration difficulty, chances are slim that the Census Bureau has missed counting a substantial number of small-haystack members. However it is also unlikely that a respondent will be sampled from this particular haystack. Unless a survey's sampling framework is designed to capture persons with rare characteristics (i.e., oversampling of minority populations), is constrained to rural populations, or is otherwise subnational in its geographic scope (i.e., survey of a particular place), a non-Hispanic white will probably not even be represented in a survey with the aforementioned contextual characteristics. But if a survey respondent (with unique

personal characteristics) is drawn from the area, there is indication that its subpopulation has been severely undercounted; and the danger posed by this partially-hidden haystack may be mitigated by a mix of population-count error, extant coverage error, and increased search costs.

3. Assessing Disclosure Risk of Masked Contextual Data

As the empirical basis of my study, I conduct an experiment to assess the contributions of various determinants of disclosure risk for contextualized survey data. In Section 2, I have already described the compilation and application of different information used in my needle-in-haystack approach to reidentification. Hence, in the remainder of this section, I only round out the details of my analysis that follows four methodological steps:

- construct a microdata file composed of a single synthetic sample of survey respondents, attaching contextual information to person-level records;
- identify geographic units that resemble a respondent's location, using available contextual data for counties, tracts, and blockgroups;
- assemble base population, haystack, coverage error, and spatial dispersion information for geographic units and aggregated contexts, attaching these search estimates to person-level records; and
- calculate summary statistics for test microdata file, reflecting the distribution of survey respondents.

3. a. Synthetic Sample of Survey Respondents and their Geographic Locations

I derive a set of synthetic person-records that are tied to a specific location 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 a synthetic dataset of person-records containing population-counts of various haystacks, estimates of coverage error, and spatial dispersion and contextual characteristics of locations.

Fifty-one state-specific block samples (including 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. By sampling one person per block, I disperse my survey respondents as thinly as possible over space. Consequently 11,562 blocks are sampled, representing 11,562 synthetic persons dispersed across approximately 5.0% of all blockgroups, 13.7% of all tracts, and 56.8% 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.

3. b. Sources and Measures of Contextual Data, Haystack Size, and Coverage Error

My sources of contextual data and 100% population-counts are also from the 2000 decennial census and its tabulated files (U.S. Department of Commerce 2000a, 2000b). These summary files are prominent public-use databases within the social sciences, providing a diversity of measures and a range of geographic detail. Contextual and haystack data are compiled from published tabulations for all counties, tracts, and blockgroups in the United States.

Covering a broad spectrum of social scientific inquiry, I have selected five contextual variables to be represented in my test dataset: (1) % Persons, Non-Hispanic White; (2) % Persons, Foreign-Born; (3) %

Persons, In-Poverty; (4) % Housing Units, Owner-Occupied; and (5) % Civilian Labor Force, Unemployed. Following my earlier work (Witkowski 2008), I have applied a nonperturbative masking technique in hopes of broadly informing the design of datasets containing contextual information at various spatial scales. After top-coding and bottom-coding these continuous variables to conceal outliers, I recode contextual measures into ten metric spaces of 10% categories (i.e., 0 - 9%, 10 - 19%, 20 - 29%, 30 - 39%, 40 - 49%, 50 - 59%, 60 - 69%, 70 - 79%, 80 - 89%, 90 -100%). Outliers were identified as those within the top and bottom 0.5% of each variables distribution (Zayatz 2005), given geographically-specific distributions defined by the metropolitan status of the geographies. Contextual variables are recoded into aggregated categories based on their absolute values (i.e., absolute recoding).[1]

To ascertain sets of look-alike geographic units (i.e., matches), I compare the contextual characteristics associated with my sample of counties, tracts, and blockgroups – 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 described earlier. Since my test dataset directly identifies population density, my matching process is constrained to three sets of geographies defined by their 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). Because contextual data for a sample of locations are originally drawn from this master file and have not been perturbed, the identification of look-alike units is exact (Winkler 2004).

For each sampled geographic unit (1,785 counties; 8,947 tracts; 10,478 blockgroups) and look-alike units within an aggregated context (315; 2,280; and 3,090 of aggregated contexts based on counties, tracts, and blockgroups), I compile three sets of information regarding their surveyed population as well as the size and composition of their total population. First I count the total number of respondents who fall within each location and context. I then enumerate the total population size of individual and aggregated areas. Finally, I tally the total number of persons who have a selected set of personal characteristics, providing estimates of haystack size for each location and context. Since analyses for a broad array of haystacks are beyond the scope of this paper, I must simplify my study by assessing only one majority and one minority haystack consisting of 20-year-old males who are either (1) non-Hispanic white alone or (2) African-American or black alone. With this approach, I am able to investigate how minority-status influences the reidentification of respondents, holding constant their gender and age.

Reflecting upon these estimates of haystack size, I have developed a simple way of identifying locations whose population-counts are susceptible to two sources of coverage error. The hard-to-count score is a composite measure reflecting the relative difficulty in enumerating populations across tracts. I assume that an extremely high score indicates that a tract's population is likely to be under-counted. I have built upon this score's measurement variation by also assuming that the lower-end of this score's distribution represents an opposing component of coverage error, the likelihood that a population is over-counted. The Census has stated that factors related to over-counting are likely to differ from those related to under-counting (p. 20 of ACETechAssess.pdf). However, a census study of duplicate counting (Mule 2002) finds that homeowners and non-Hispanic whites are most often double-counted. In direct contrast, renters and minority populations are most often under-counted and are likely to be found in tracts that are difficult to enumerate (ACETechAssess.pdf). Hence there is some evidence that low hard-to-count scores may provide a fairly accurate indicator of over-counting.

Census hard-to-count measures are only estimated for tracts. Consequently I assign a blockgroup the same values as the tract it is nested within. Higher-level estimates of coverage error are calculated as the proportion of tracts within counties and look-alike contexts that are likely under- and over- counted. To maintain consistency across measures of coverage error, I also calculate the difference in rentership rates between race/ethnic groups at the tract-level (e.g., rentership rate of "non-Hispanic white households" minus rentership rate of "other households"). I then create blockgroup-level estimates of rentership

[1] As discussed by Witkowski (2008), disclosure risk is heightened considerably by global recoding contextual measures based on their percentile distributions. Consequently, I present analyses for contextual information that is recoded according to their absolute values.

following the same methodology as the hard-to-count measures; while higher-level rentership estimates are derived from the average of tract-level values for each county and aggregated context. Supplemental analyses indicate that tract-level estimates of rentership rates conceal significant racial differences in enumeration difficulty occurring at the blockgroup-level. Given the limited scope of my study, I must set aside this scale issue and utilize a consistent methodological approach so that I may clearly assess the broad implications of coverage error to disclosure risk.

The size of haystacks indicates the amount of uncertainty associated with correctly reidentifying a particular respondent. However additional search costs may accrue from the spatial dispersion of these haystacks. Assessing the distribution of populations within and between search locations, I tally the numbers of look-alike geographic units, square miles of land area, and the states represented within individual and aggregated locations. Taking a calculated risk, the intruder may decide to ignore geographic units that are unlikely to a respondent drawn from an exceptionally small haystack. The degree to which this factor operates in the reidentification process is indicated by the percent of look-alike geographic units in an aggregated context that have at least one member of a subpopulation.

The analytical unit of ultimate concern is the sampled person. Consequently, I attach the above higher-order information to my set of synthetic person-records. Producing summary statistics reflecting the distribution of survey respondents, this approach bridges a complexity of estimates derived from either individual geographies or aggregated contexts. While synthetic survey respondents are provided a set of contextual characteristics describing their locations, they are not assigned any personal characteristics. Hence, I do not know the sex, age, and race of any particular respondent. But I do know the size of haystacks and the total population within each geographic unit and aggregated context as well as the proportion of respondents who have been drawn from these areas. Bringing this information together, we can ascertain the likelihood that a respondent will have a set of personal characteristics as defined by the haystack of interest.

4. Presentation of Results

Given my premise for population-size thresholds and the above measures, I conduct two sets of analyses that evaluate the potential role of haystack size, coverage error, and dispersion in determining disclosure risk of contextualized microdata. For the first set of analyses presented in Table 1, I assess how the reidentification process is changed when I attach contextual information instead of directly identifying geographic units. Described in broad strokes, I illustrate how the above haystack characteristics are functions of sparsely and densely populated geographic units; and how the aggregation of look-alike geographies moderates disclosure risk by increasing uncertainty and search costs. In doing so, I produce separate analyses for survey respondents nested within individual geographic units and aggregated contexts that are highly populated (“dense”) or less populated (“sparse”), defined as those whose populations are above and below 100,000 persons.

[Table 1 Here]

Looking at Table 1, analyses reveal that the release of contextual information, instead of the direct identification of geographies, has important ramifications for disclosure risk. The chances of a survey respondent being located in a densely populated area dramatically increases when contextual data, at any spatial scale, are attached to their records. Most survey respondents (96%) are found in densely populated contexts derived from counties; while over a majority (59 to 67%) live in highly-populated contexts aggregated from small-scale geographies. For all scales of geography, the total population size (on average) rises to over 2.4 million persons for high-density aggregated contexts; and for low-density contexts the population is at least 29,000 persons.

For all scales of geography, the accumulation of base populations within aggregated contexts gives rise to significantly larger haystacks. High-density contexts have at least 11,000 non-Hispanic white males age 20, while sparse contexts have at least 125 members of this majority subpopulation (on average). As expected the minority subpopulation of young black males typically has much smaller haystacks than its majority counterpart. But aggregating look-alike geographies increases the size of haystacks emanating from small-scale units to the point that at least 59 twins are found in sparse contexts and as many as 467 are found in dense contexts. Minority haystacks are even larger for dense contexts derived from counties, with an average size of 2,641 members.

Looking at the shifting patterns of coverage error resulting from aggregation, I find evidence of the selection of unique geographies into contexts with less than 100,000 persons. With twice the difficulty of being enumerated as individual units and dense contexts, 47 to 71% of look-alike geographies in sparse contexts are likely undercounted. The concentration of hard-to-count populations within relatively unpopulated and unique contexts is further indicated by the exceptionally low likelihood of double-counting (i.e., 1 to 5%). With national rates of undercounting ranging between 1 to 2%, the disproportionate number of difficult-to-enumerate tracts may prove to be a significant barrier to reidentification.

However the protection offered by coverage error, and therefore the translation of hard-to-count scores to actual rates of undercounting, is intricate. First of all, there is evidence that the level of enumeration difficulty captured by hard-to-count scores does not sufficiently reflect the complexity of residential segregations patterns surrounding underrepresented race and ethnic groups. Regardless of the spatial scale, population density, and geographic detail of contexts, non-Hispanic white households are actually easier to enumerate than other populations within their tract, while African-American or black households are actually more difficult to count. Compared to others in their tract, approximately 2 to 11% more black households are renters while 13 to 17% more non-Hispanic white households are homeowners. Given this offsetting pattern of homeownership, coverage error tends to close the minority-majority gap in disclosure risk. However the chance that a small minority haystack has been largely hidden may be capped by contexts with small base populations. For instance, a sparse context (derived from tracts) typically has 34,668 persons in its total population. Even if 71% of these contexts' tracts are probably undercounted, it is unlikely that this small population would contain an exorbitant number of hidden members.

While the potential size of hidden haystacks within sparse contexts may not be sufficiently large, the same high error rates could indicate the lack of coverage within identifying extant files. Consequently the intruder will likely have to perform supplemental search activities for respondents from these areas. While cost estimates for these activities are not available, we do know that the search – for respondents in aggregated contexts with less than 100,000 persons – would cover an average of: 2 counties, 7,503 square miles, and 2 states; 9 tracts, 581 square miles, and 5 states; or 23 blockgroups, under 1 square mile, and 10 states. If an intruder is willing to take a calculated risk, she could lower these costs by ignoring geographic units that are unlikely to have a survey respondent drawn from the area. The savings could be pronounced, especially for minority populations residing in sparse aggregated contexts, where 58 to 78% of look-alike geographies have not been assessed a single black male aged 20. But as stated earlier, disregarding these unlikely locations is likely a misstep for intruders since the Census Bureau has introduced noise into the original tabulations for locations with extremely small haystacks.

5. Conclusions

In writing this paper, I have two complementary objectives. My first aim is to describe an empirical approach for conducting disclosure analysis of contextualized microdata. In doing so, I describe a complexity of disclosure factors that underlie populations bounded by small area geographies and extend these determinants to aggregated contexts. My second aim is to develop an analytical framework that accesses these factors that influence the correct reidentification respondents, illustrating the intricacies involved in interpreting such an analysis.

To meet these objectives, I layout the theoretical underpinnings of the intruder search by describing the role of the intruder, taking a "needle-in-haystack" approach to reidentification, and incorporating economic concepts of uncertainty and cost. I then detail the empirical steps involved in reidentifying geographic units and individuals that are represented within a survey dataset. Expounding upon the reidentification process, I discuss the relationships between factors and the mechanisms that shape each components' contribution to risk. Utilizing a synthetic set of survey respondents that are thinly dispersed across geographic units, I conduct a reidentification experiment for a dataset that contains the metropolitan status of respondents as well as five county, tract, and blockgroup-level contextual variables, collapsed into 10% categories. Incorporate measures of coverage error and haystack dispersion, I assess the relative contribution of each factor.

My study indicates that contextualized microdata may prove to be a viable method of safely distributing geographically rich information. This finding is particularly pertinent for county-level contextual information, where only 4% of survey respondents are typically located in aggregated contexts with fewer than 100,000 persons. However, more work needs to done to fully understand the implications of premises underlying population-size thresholds. While my results show the potentially important role of coverage error in ensuring the anonymity of respondents, further research is needed to create and analyze data that better captures spatial variation in undercounting for different subpopulations. Furthermore this study is also limited to a single set of contextual information and two emergent haystacks. A more complete assessment is needed for a comprehensive set of haystacks as well as an expanded set of contextual information that varies in measurement composition and detail.

6. References

Robinson, J., Ahmed, B., das Gupta, P., and Woodrow, K., "Estimation of Population Coverage in the 1990 United States Census Based on Demographic Analysis," Journal of the American Statistical Association, 88: 423, pp. 1061-1079. 1993

Mule, Thomas. 2002. "A.C.E. Revision II Results: Further Study of Person Duplication." Decennial Statistical Studies Division, U.S. Census Bureau.

Mule, Thomas. 2002. "Further Study of Person Duplication Statistical Matching and Modeling Methodology," A.C.E. Revision II Memorandum Series PP-51, December 31, 2002.

Mule, T. (2001). ESCAP II Person Duplication in Census 2000. Executive Steering Committee for A.C.E. Policy II, Report 20, Oct. 11, Decennial Statistical Studies Division, U.S. Census Bureau, Washington, DC.

Mule, T. (2003). *A.C.E. Revision II Results: Change in Estimated Net Undercount.* Decennial Statistical Studies Division, A.C.E. Revision II Memorandum Series PP-58. Washington, DC: U.S. Census Bureau.

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

Cambridge University Press. 2003. *Cambridge Dictionary of American Idioms*. Cambridge University Press: Cambridge, UK.

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.

Energy Information Administration. 2001. *Residential Energy Consumption Survey.* http://www.eia.doe.gov/emeu/recs/recs2001/codebook82001.txt (accessed December 27, 2007).

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, 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. Census Bureau. 2003a. *Technical Assessment of A.C.E. Revision II*. http://www.census.gov/dmd/www/pdf/ACETechAssess.pdf

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

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. 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

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. 2008. "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.

7. 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]: 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.

Table 1. Aggregation of Sampled Geographic Units into Look-Alike Contexts, Weighted to Reflect Spatial Distribution of Survey Respondents (N=1,785; 8,947; and 10,478 of Sampled Counties, Tracts, and Blockgroups; N=315; 2,280; and 3,090 of Sampled Aggregated Contexts Based on Counties, Tracts, and Blockgroups; N=11,562 of Synthetic Survey Respondents)

	Survey Respondents Residing in											
	County as Contextual-Base				Tract as Contextual-Base				Blockgroup as Contextual-Base			
	Geographic Unit		Aggregated Context		Geographic Unit		Aggregated Context		Geographic Unit		Aggregated Context	
	<100K	100K+	<100K	100K+	<100K	100K+	<100K	100K+	<100K	100K+	<100K	100K+
Proportion of Respondents in Context	0.34	0.66	0.04	0.96	1.00	0.00	0.33	0.67	1.00	0.00	0.41	0.59
Average Size of Total Population within Sampled Context	41,756	955,163	51,233	3,469,881	4,898	---	34,668	2,484,499	1,613	---	29,243	2,489,206
Average Size of Haystack Subpopulation of												
White alone, non-Hispanic, Males Age 20	275	3,128	264	17,715	25	---	155	11,808	8	---	126	11,155
African-American or Black alone, Males Age 20	33	1,222	130	2,641	5	---	76	467	2	---	59	318
Proportion of Tracts in Sampled Geographic Units and Aggregated Contexts that are Likely Under-Counted	0.17	0.28	0.47	0.23	0.29	---	0.71	0.08	0.29	---	0.59	0.07
Proportion of Tracts in Sampled Geographic Units and Aggregated Contexts that are Likely Over-Counted	0.18	0.28	0.05	0.25	0.19	---	0.01	0.28	0.19	---	0.03	0.32
Difference in Proportion Renters												
White alone, non-Hispanic (minus Others)	-0.16	-0.15	-0.13	-0.16	-0.16	---	-0.15	-0.17	-0.16	---	-0.16	-0.16
African-American or Black alone (minus Others)	0.02	0.11	0.04	0.09	0.10	---	0.09	0.10	0.10	---	0.11	0.11
Average Number of Geographic Units Resembling Sampled County	117	14	---	---	381	---	---	---	1,098	---	---	---
Average Number of Geographic Units in Aggregated Context [1]	---	---	2	51	---	---	9	566	---	---	23	1,838
Average Sq.Miles of Land Area of Sampled Geographic Units and Aggregated Contexts	1,848	1,272	7,503	44,691	113	---	581	39,759	0.03	---	0.44	32.52
Average Number of States in Aggregated Context	---	---	2	12	---	---	5	31	---	---	10	37
Proportion of Geographic Units with Subpopulation (in Context)												
White alone, non-Hispanic, Males Age 20	1.00	1.00	0.99	1.00	0.95	---	0.98	1.00	0.88	---	0.96	1.00
African-American or Black alone, Males Age 20	0.79	1.00	0.79	1.00	0.59	---	0.94	1.00	0.38	---	0.92	1.00

Table 1 (cont). Aggregation of Sampled Geographic Units into Look-Alike Contexts, Weighted to Reflect Spatial Distribution of Survey Respondents (N=1,785; 8,947; and 10,478 of Sampled Counties, Tracts, and Blockgroups; N=315; 2,280; and 3,090 of Sampled Aggregated Contexts Based on Counties, Tracts, and Blockgroups; N=11,562 of Synthetic Survey Respondents)

Note: Excluded from analyses are those geographic units with no population, resulting in 3,141 counties; 65,174 tracts; and 208,125 blockgroups considered from the geographic-unit population.

Note: Dataset contains five county, tract, and blockgroup-level contextual measures of (1) % Persons, Non-Hispanic White; (2) % Persons, Foreign-Born; (3) % Persons, In-Poverty; (4) % Housing Units, Owner-Occupied; and (5) % Civilian Labor Force, Unemployed, recoded into 10% categories (i.e., 0 - 9%, 10 - 19%, 20 - 29%, 30 - 39%, 40 - 49%, 50 - 59%, 60 - 69%, 70 - 79%, 80 - 89%, 90 -100%). This dataset also directly identifies MSA-status of geographic units: (1) MSA 1-million or more, (2) MSA less than 1-million, and (3) Non-MSA.

Note: Weighted to reflect the spatial distribution of survey respondents, averaged values are derived from sets of geographic units and aggregated contexts having a total population of size that is either less than 100,000 persons or 100,000 persons or more, indicating: (1) number of geographic units resembling sampled geography; (2) number of geographic units in aggregated context; (3) size of total population in a geographic unit or an aggregated context (i.e., distributed across all look-alike geographic units); (4) proportion of tracts in an individual geographic units or geographic units in an aggregated context having a "high" or "low" Hard-To-Count score; (5) number of states in an aggregated context; and (6) number of square miles of land area in a geographic unit or an aggregated context (i.e., distributed across all look-alike geographic units).

Note: Details of the construction of Hard-to-Count scores are provided by Bruce and Robinson (2003). Tract-level scores are assigned to nested blockgroups and are aggregated into county- and context-level estimates. A "high" and "low" levels of error in extant data are reflected by the top and bottom quartiles of Hard-to-Count scores, as derived from tract distributions.

Note: "0.00" indicates a proportion of respondents in contexts greater than zero but less than 0.05, while "---" indicates an absolute value of zero. "---" also indicates that a particular population-size category did not apply to a set of geographies and, therefore, associated statistics were not calculated.

[1] The low number of geographic units in aggregated contexts with less than 100,000 persons reflects the selection of a limited set of relatively unpopulated "look-alike" units into these contexts.