Confidentiality and spatially explicit data: Concerns and challenges

Leah K. VanWey*†, Ronald R. Rindfuss‡, Myron P. Gutmann§, Barbara Entwisle‡, and Deborah L. Balk¶

*Department of Sociology, Indiana University, Bloomington, IN 47405; ‡Department of Sociology and Carolina Population Center, University of North Carolina, Chapel Hill, NC 27516-2524; §Inter-University Consortium for Political and Social Research, University of Michigan, Ann Arbor, MI 48106-1248; and ¶Center for International Earth Science Information Network, Columbia University, Palisades, NY 10964

Edited by Susan Hanson, Clark University, Worcester, MA, and approved September 14, 2005 (received for review July 10, 2005)

Recent theoretical, methodological, and technological advances in the spatial sciences create an opportunity for social scientists to address questions about the reciprocal relationship between context (spatial organization, environment, etc.) and individual behavior. This emerging research community has yet to adequately address the new threats to the confidentiality of respondent data in spatially explicit social survey or census data files, however. This paper presents four sometimes conflicting principles for the conduct of ethical and high-quality science using such data: protection of confidentiality, the social–spatial linkage, data sharing, and data preservation. The conflict among these four principles is particularly evident in the display of spatially explicit data through maps combined with the sharing of tabular data files. This paper reviews these two research activities and shows how current practices favor one of the principles over the others and do not satisfactorily resolve the conflict among them. Maps are indispensable for the display of results but also reveal information on the location of respondents and sampling clusters that can then be used in combination with shared data files to identify respondents. The current practice of sharing modified or incomplete data sets or using data enclaves is not ideal for either the advancement of science or the protection of confidentiality. Further basic research and open debate are needed to advance both understanding of and solutions to this dilemma.

data preservation | data sharing | disclosure risk | social surveys | spatial social science

Considering individuals in their spatial contexts opens a rich array of analytic possibilities. Geographers have traditionally considered the spatial organization of populations and their characteristics. Other social scientists have focused on the importance of social location (defined by age, race, gender, education, position in social networks, etc.) for actions of individuals and families, with little attention to the importance of spatial location and spatial relationships. The increasing integration of these two lines of inquiry is made possible by methodological advances in the spatial sciences and data collection advances in the social sciences. As a result, spatially explicit data sets that contain information on the attitudes and behaviors of individuals and households are now being created that permit researchers to address important scientific questions that have heretofore resisted methodologically defensible empirical analysis.

Although the creation of these new data sets is good news for the spatial and social scientific communities, their exploitation to further our understanding of the causes and consequences of human behavior is currently being hampered by uncertainty about the effects of the availability of such spatially explicit data on the risk of causing harm to respondents through confidentiality breaches. This uncertainty is leading to the underutilization of data and is increasing the possibility that well intentioned scientists could inadvertently disclose information that could harm respondents. Because of this uncertainty, we begin this paper with four principles that must guide the collection, analysis, publication, distribution, and archiving of spatially explicit, micro-, social science data, showing how, in combination, they can stand in opposition to one another. We do so in the belief that before a problem can be solved, it needs to be understood. In our view, the combination of spatially explicit displays of data and the sharing of data across research teams creates a new problem that is insoluble with accepted practices in geography or other social sciences. We thus provide more detail on these two issues after discussing the four principles.

The Principles and the Problem

Protection of Confidentiality. Ensuring the confidentiality of information collected about individual human research subjects is fundamental to the ethical conduct of research. Information that, if known, might lead to physical, emotional, financial, or other harm must not be able to be linked to individuals or households [see, for example, the codes of conduct for the Association of American Geographers (1), the American Psychological Association (2), the American Political Science Association (3), and the American Sociological Association (4)]. In following this principle, researchers are representatives of the larger scientific community. The promise of confidentiality of responses is not a bargain between the individual researcher and respondent, nor is it only for the duration of the research project. Ensuring actual and perceived confidentiality of scientific data is necessary to guarantee the continued participation of the public in censuses and social surveys. Individual researchers must be concerned with the protection of confidentiality at all stages and in all types of research, including when they collect, disseminate, use, and read about data.

The Social–Spatial Linkage. The social–spatial linkage is the key to the advancement of science in a variety of fields. By linking the characteristics and actions of individuals, households, or communities to a point in geographical space, researchers can conduct a wealth of analyses about spatial pattern, process, and the importance of relative location for respondent behaviors. For the most rapid advancement of science, this linkage must be available to researchers outside of the original data collection team. In contrast to names or Social Security numbers, the traditional personal identifiers in social survey data that would allow linkage across a variety of data sources, location and the information that it provides about relative position are important inputs into statistical models. Although one would never argue that knowing names was essential for analyzing the relationship between alphabetical position and behavior, one can easily argue that geographic location is essential for understanding the diffusion of behaviors.

Conflict of interest statement: No conflicts declared.

This paper was submitted directly (Track II) to the PNAS office.

Abbreviation: Add Health, the National Longitudinal Study of Adolescent Health.

†To whom correspondence should be addressed. E-mail: lvanwey@indiana.edu.

Data Sharing. Data sharing is essential on both scientific and financial grounds. Replication is essential for the advancement of science. For other research teams to be able to replicate, falsify, or modify key research findings, data must be shared with the larger research community. Spatial location is a key attribute of social survey or census respondents that, when made available, allows other researchers to conduct their own analyses and to link new data to existing data. This necessity is recognized by funding agencies, which now mandate the sharing of data collected using their funding. The highest-quality data embody a great deal of money and time that should not benefit only the research team who collected them.

Data Preservation. Similarly, the preservation of data in its most usable format, including the spatial location and other attributes of respondents, for future generations of scientists is key to the advancement of science. Although sometimes carried out by the same organizations, preservation and data sharing remain fundamentally different activities. It is possible to think of a long-term preservation strategy that postpones the sharing of information for several generations (as is the case under the U.S. law that guards original census records for 72 years after their creation), and it is possible to think of a data-sharing strategy that exists for the duration of a project's active life (while it has funding, for example) yet makes no long-term plans for preservation.

The Problem. Whether through the publication of data (including maps) or the direct transfer of data files, the dissemination of data to other researchers poses confidentiality protection problems for spatially explicit social survey or census data. Consider a simple example: Imagine social survey data collected on health-related work absences, which a research team used in the absence of any environmental data. If the data set contained the spatial location of work settings (e.g., respondent no. 72118 works at 35.9123°N, 79.0569°W), other researchers could easily use such information to add data on air pollution, average number of cloudy days, or elevation for their own analysis. If the original data were disseminated with work location to facilitate such analyses, respondent no. 72118 could be identified by locating her workplace with a handheld global positioning system (GPS) and then comparing the characteristics of employees at that location with other data on respondent no. 72118 in the public data file. Although this example is intuitively clear, it is important to remember that there are relatively few simple cases or simple solutions. Disclosure risk associated with display or dissemination of spatially explicit social survey or census data depends on the geographic coverage of the data, whether and how geographic units are sampled, whether and how individuals are sampled within those units, and the heterogeneity of individuals and sample clusters.

The following sections of this paper deal in more detail with issues related to the graphic display of locations of respondents or sample clusters and to the public release of data. We argue that this combination of data display and data sharing increases the possibility that the sharing of data will conflict with confidentiality promises. It is obvious that researchers' promises of confidentiality require that the exact location of respondents must be excluded from publicly released data. However, it also is important that researchers realize that even indirect identification of respondents or sampling clusters must be avoided. For example, it is tempting to publicly thank institutions that constitute sampling clusters for their cooperation in a web site or press release, but such release of information is potentially dangerous and should be avoided. It is similarly tempting to provide a map showing (even in general terms) the distribution of respondents or sample clusters, potentially with some interesting but nonsensitive information. The difficulty arises when that nonsensitive information or distribution is combined with information contained in the public-use data file (see discussion of Fig. 1 below).

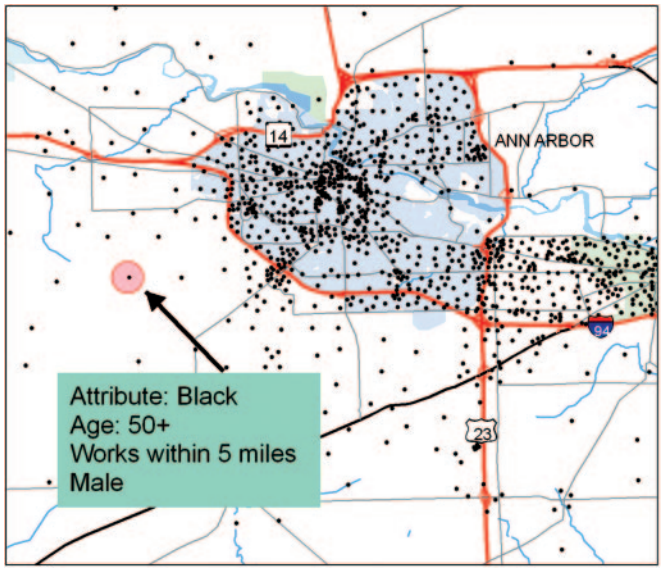

Fig. 1. Spatial distribution of residences of hypothetical survey respondents in Washtenaw County, Michigan, with attributes of one respondent.

In recent years, a small number of data producers and data archives, including the U.S. Census Bureau, have attempted simultaneously to share data and limit the risk of disclosure by restricting the data to use within a secure workplace, sometimes called an "enclave" or a "cold room." These facilities give carefully selected researchers who are able to travel to their location limited access to data; these facilities also control and delay the removal of research results from the premises and require that researchers pay a substantial daily or monthly fee for access. Most observers of these facilities report that, although they permit research to be completed, they are costly for both the operator and the data user, and because of their cost and the temptation for data holders to enclose more and more data, they reduce the extent to which data are used for the most sophisticated forms of analysis. Some of these facilities are hardly used at all as a result, and even the most frequently used are not used to capacity. Although they are an important solution to the problem and one that we support, restricted data enclaves need to be improved substantially before they can be considered successful.

The amount of information about location that may ever be made public in any form (e.g., in maps that are part of papers and presentations) is still something that the research community must conclusively work out. Although standards exist for the presentation of purely geographical data (5, 6) and separately for the presentation of tabular data (7), no standards exist for the presentation of maps based on extensive tabular data. Similarly, we have not begun to estimate disclosure risk from combinations of map data and tabular data. We begin this conversation in this paper by considering the identification of sample clusters in a map and by considering in more depth the options for dissemination of data with social–spatial linkages.

Display and Visualization

Maps and displays based on spatially referenced social survey data have great potential to inform the research process and communicate complex results. At the same time, display and visualization can threaten confidentiality, and it is incumbent on researchers to minimize disclosure risks. The risk of disclosure varies from project to project, and there are currently no

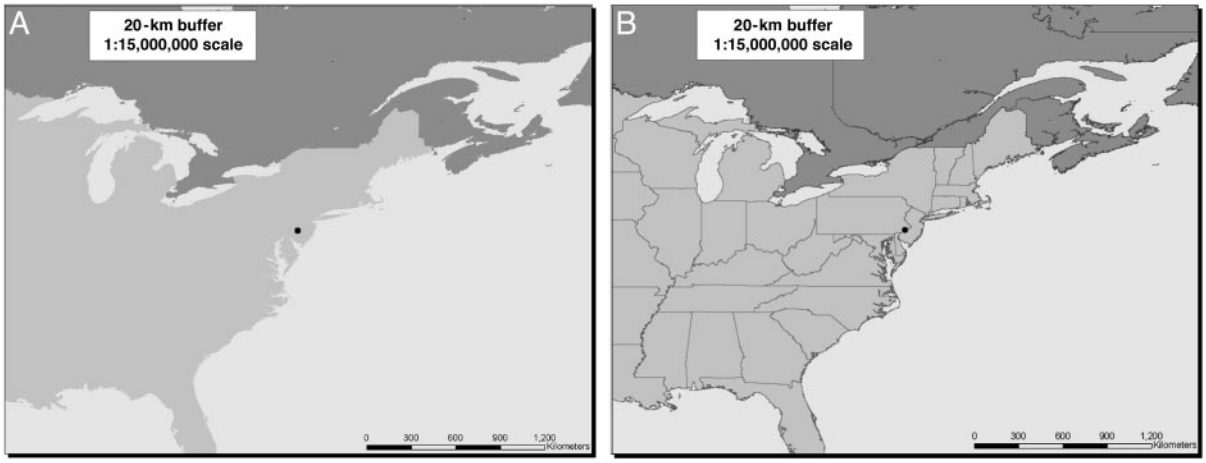

Fig. 2. Point location of school in hypothetical public release data file. (*A*) Without state boundaries displayed. (*B*) With state boundaries displayed.

accepted methods for estimating disclosure risk or standards for minimizing disclosure risk in the display of maps. This section provides an example for assessing such risk for one project.

Presenting information cartographically is a very useful tool, one that more and more researchers use. At the same time, disclosure risks are associated with cartographic data presentation. Among these risks is the long-term harm that may come from the publication of maps that indicate, even in general terms, the location of an individual who is included in a linked (or potentially linked) social science research database. No responsible researcher would publish the name or other precise identifier, such as an exact address, of a research subject in a journal, web-based publication, or handouts at a public presentation. They might, however, publish other information that they thought was innocuous, e.g., the mapped location of individuals with certain kinds of responses, or publicly available natural information about a place where the subject lived or worked. Fig. 1 shows a simple example of the sort of risks accompanying maps.

Fig. 1 is a map of Washtenaw County, Michigan, overlaid with a hypothetical set of locational points indicating respondent residences. The figure shows some attributes of a respondent. Although seemingly harmless, these data could be linked with public-release survey data that have no locational information to find out other, potentially more compromising, information about the respondent. The published map (Fig. 1) only shows age, sex, race, and distance to work. But suppose the article also provides a link to a publicly available data set. If the data set is a random sample of the adult population of the United States, knowing that the person is male narrows the search for the respondent to about half the cases in the data set. Further, knowing that the person is black, over 50 years old, and lives in the Midwest would narrow the search for the respondent to one or at most a handful of cases in the data set.

Further, consider the disclosure risks associated with putting a visual display of a clustered sample design in a publication or presentation. We present examples drawn from the National Longitudinal Study of Adolescent Health (Add Health), which involved surveys administered to ≈20,000 adolescents at three time points, to consider possible rules for the display of spatial data that limit disclosure risk. One of the motivations for Add Health was to examine how social contexts such as families, friends, peers, schools, neighborhoods, and communities influence adolescents' health and risk behaviors. Thus, the survey implemented a school-based sample design in which schools served as clusters in the first stage of sample selection, and then students were sampled from these schools in the second stage. The clustered sample design increases disclosure risks, and visual display adds to these risks because it may reveal clues about the identity of the clusters as well as about the respondents themselves.

In considering rules for the display of spatial data, we first must specify a level of disclosure risk that can be tolerated. Disclosure risks are twofold: (*i*) those that are associated with identification of the cluster and (*ii*) those that are associated with identification of respondents given a known cluster. To keep things simple, we restrict our attention to the first, the risk of identifying a cluster, and specify a median probability of identification of 0.05 as the risk level that we can tolerate in a visual display. A 0.05 risk level is for illustrative purposes only. In practice, we expect that most researchers and their institutional review boards would set this tolerance level much lower.

Add Health is but one of a large number of studies that use a sampling design that begins with selecting schools. To illustrate our point about cluster designs, we simulate a sampling frame of public schools containing an eleventh grade by using information from the National Center for Education Statistics (http://nces.ed.gov/ccd/address.asp). These data have the geographic location of each public school in the continental United States in 2000 containing an eleventh grade. We can assume that a potential intruder also has access to these publicly available data.

The question then is how to design a display to achieve an average of a 1-in-20 chance of identifying a given school. As mentioned, this risk is only one part of that associated with the risks of disclosing respondent identities. The chances of identifying a respondent will be considerably less than 1 in 20, with the actual likelihood depending on the characteristics of the respondent and cluster (relative to the characteristics of the populations of respondents and clusters) as well as sampling (if any) within the cluster.

Given a disclosure risk level of 0.05 or lower and given information about the locations and spatial distributions of schools, map symbols can be designed so that a point indicating the location of a specific school could be associated with any of 20 schools. There are 13,126 schools in all. For each school, the size of the buffer that would contain 19 other schools was determined. The median size of the buffer is 20.5 kilometers (km). This median indicates that, at least half the time, a circle having a radius of 20.5 km drawn around a specific school will contain ≥20 schools. The points or dots indicating school location could be set to this size. Fig. 2*A* shows the location of an arbitrarily selected school in the Central Atlantic region with a point or dot corresponding to a 20-km buffer. If a 0.05 level of disclosure risk could be tolerated for the median school, then location of a sample school could be shown at this scale of

presentation. Clearly, if the level of disclosure risk were set more stringently, say at 0.01, then the size of the buffer would be considerably larger.

Threats to data security at a given scale also will depend on what else can be inferred about the school. Fig. 2*A* does not show state boundaries. If state boundaries or other additional themes were included in the map, disclosure risks would increase. Fig. 2*B* adds state boundaries. Looking at this map, given the buffer's location toward the southern end of the boundary between Pennsylvania and New Jersey, one would likely guess (correctly) that the selected school is located in Philadelphia. The general point is that layers or themes potentially displayable on a map add to the security threat.[‖]

Threats to data security also depend on heterogeneity in the spatial distribution of the sample clusters. Schools are not evenly spread over the map of the United States. The density of schools is much higher in large cities than rural areas, for example. In the case of our example, knowing or guessing that the school selected for illustrative purposes is in Philadelphia actually tells us less than it might appear. Again using a 0.05 level of disclosure risk, for schools in large cities such as Philadelphia, we need only a 5.8-km buffer on average, much less than the 20-km buffer incorporated in Fig. 2. Putting it another way, given that the school of interest is in Philadelphia, the 20-km buffer in Fig. 2 includes many more than 19 other schools. In fact, for schools in large cities such as Philadelphia, the use of a 20-km buffer in these maps is equivalent to a disclosure risk level of <0.01, not <0.05.

However, the problem is the reverse in rural areas, where schools and respondents are widely dispersed. A buffer of 20 km is insufficient. We would need a buffer of 51.2 km to achieve no more than a 0.05 level of disclosure risk in the median case. Further, it might be possible to infer from a map that the selected school is likely in a rural area. Even if state boundaries were not shown, it might be possible to guess from position on a continental map that a school was located in, for example, South Dakota, a heavily rural region of the country. If clusters are randomly distributed, then other than their general location, no additional information will be revealed in a map. If there are regional or rural–urban patterns or other systematic patterns in the spatial distribution of the clusters, however, additional care must be taken.

This example has been based on Add Health, a data set that contains very sensitive data on sexual and other behaviors. It is unlikely that the identities and locations for their sample schools will ever be released. Nevertheless, our examination of disclosure risk in visual displays related to this data set illustrates important points associated with visual display of any spatially referenced survey data set based on a multistage clustered design. The risk associated with any given display depends on the scale of presentation and the various layers displayed. It also depends on the available public-use data, both whether the data are available and what variables are included. It also is important to note that, before we did the actual analysis, we did not know what could be reasonably displayed in a publication, presentation, or on a web site. The general conclusion of this exercise is that, as our capabilities for linking geographic data with survey data improve, we also need to better understand the risks that are being created for disclosure of the identities of respondents, and this type of research is just beginning. We are still learning what we do not know.

Protecting Location Information in Data That Are Archived and Shared

Many social scientists have made a practice of both long-term data preservation and sharing their data with others for secondary analysis. In recent years, these practices have been encouraged by the policies of governmental and nongovernmental research funders.** It is because of this practice of sharing and preserving data that we see risk that goes beyond a map showing confidential information at the exact location of respondents. Introducing the notion of data preservation and sharing into the confidentiality protection equation adds the role of the data archivist (and the funding agencies who often mandate data sharing) to the list of actors who are involved and highlights the rewards and risks that flow from collecting, sharing, and making use of these data. We count as data archivists those whose primary organizational responsibility falls in that area as well as those who lead and manage research centers where data collection and data sharing are undertaken, even if they do not normally think of themselves as archivists.

The data archivist shares with the original producer of the data the responsibility for making the data available to secondary users without compromising the promise of privacy and confidentiality that was made when the data were collected. The archivist also accepts the responsibility to promote the advancement of science by sharing data that are as useful as possible, thereby ensuring that secondary data users, and the larger scientific community, have the greatest possibility to use the data. By definition, the archivist lives with the choices that the data producer has made about what data to collect, how they were collected, what information was revealed to the public during the time of data collection, and what information the data producer published before or after the transfer of data to the archive.

The archivist also has the responsibility of thinking about the long-term preservation of the data in the context of potential future opportunities and risks that might emerge. Some disclosure risks may diminish over time, e.g., population mobility reduces the certainty that a spatial link to a residence reflects where someone currently lives. Preserving some data that will be released only after a period of embargo (not necessarily as long as the U.S. Census Bureau's 72 years) may ameliorate some confidentiality problems but is no panacea. The timely secondary analysis of data is required for continued scientific advancement. But the archivist also needs to consider disclosure threats that might arise in the future as increasing amounts of data are linked and linkages become easier to make.

Responsible data producers and data archives have long included an examination and reduction of potential disclosure risk in the procedures that they follow when making data available to others and when accepting data into their collections (8, 9). A substantial and growing literature exists about ways to limit disclosure in tabular data with area identifiers but no precise spatial locations (10–15). Data suppression, such that data from areas with small numbers of observations are not released, is one possible technique that preserves original data for larger areas. Alternatively, data producers or archivists may transform data while maintaining important characteristics of the original data (16, 17) by changing records (e.g., swapping a record from one spatial area with one from another area), by

[‖]As part of a restricted-use contract, Add Health releases relative household location within primary sampling units, with no links to any other data. With these data, it is possible to map the characteristics of respondents in Euclidean space (20). Adding layers to such a map would increase not only its scientific value but also the security risk. Even without spatial links, the relative household location can be compared with census-based maps of the school-age population to narrow the list of possible sampling units. Extreme care must be taken in the presentation of any visual display based on these data.

**The National Institutes of Health, the National Science Foundation, and many other research funders require some or all of their grantees to share data. The National Institutes of Health and National Science Foundation policies on data sharing may be viewed at: http://grants2.nih.gov/grants/policy/data_sharing and www.nsf.gov/pubs/2001/gc101/gc101rev1.pdf (section 36).

changing attributes (e.g., by recoding values so that extreme values are combined with less extreme values), or more recently, by creating data that are partly or completely synthetic.

The above-mentioned methods work for tabular data without spatial locations of individual respondents because it is possible to limit the possibility that someone enumerated in a survey or administrative database can be uniquely identified as a single individual in the population at large. More recently, geographers have developed limited methods for adding noise or transforming those locations to limit disclosure risk while preserving the ability to analyze those data (18). These methods go beyond the basic technique of aggregating over space to reduce the likelihood of identifiable individuals and include adjusting geographic coordinates by various means and attaching contextual variables to the microdata so that secondary data users do not know the exact location. However, these methods may not be as useful as they could be and themselves carry dangers, e.g., they may allow for sufficient inference to identify the location of a respondent or sampling cluster. Geographers also have used a variety of methods to convert point data or aggregated census data to continuous surfaces to represent the distribution of population characteristics without identifying individual respondents (19, 20). These surfaces provide an adequate representation of distributions on single variables and can be combined for understanding correlations between variables. However, the dissemination of only such surfaces does not provide the microdata on the individual characteristics and relative positions of individuals that are necessary for cutting-edge analyses of spatially explicit census and survey data.

Until recently, virtually all shared social-science data were made available for secondary analysis as public-use data files, in which all known potential identifiers were removed or obscured. These data could then be distributed widely to the research community, who were able to share them without concern that individual respondents could be identified. When spatial locations are linked to social science data, such public-use distribution systems become almost impossible to maintain. As mentioned, one solution to this dilemma is for data producers and data archivers to limit access to their data, either by entering into contractual agreements with potential data users (who agree to restricted terms of data use) or by insisting that data users only use the data in a restricted-use enclave facility. These solutions still require improvements, perhaps by developing something that approaches a true virtual enclave, in which restricted access to data can take place, without requiring travel, access fees, or delays before the results are available to the researcher. One goal we advocate is a clearly delineated partnership of data producers, data archivists, and data users that is designed to ensure that everyone acts together to share data that have been designed and produced with an eye toward maximizing both primary and secondary use.

Discussion and Conclusion

To summarize, many questions require linking microlevel survey or census data with spatially explicit data that characterize the social, economic, and biophysical context in which survey or census respondents live, work, and/or engage in leisure activities. Once the precise spatial locations of a person's activities are known, these locations serve as identifiers that can be used as links to a vast array of spatial and social data. This linkage poses challenges to issues of confidentiality, data sharing among scientists, and archiving data for future scientific generations. We have a moral responsibility to protect the confidentiality of respondents, and if we shirk that responsibility, respondents will not provide data on future censuses and surveys. Similarly, we have a moral responsibility to advance the scientific agenda on issues that could benefit from the sharing of data that link respondents to their social and biophysical environments. Further, the history of science provides many examples of new and important uses of old data, which argues for the preservation of data; but, as data are archived, how can investigators be assured of the safeguarding of respondents' confidentiality by future researchers?

In this paper, we have indicated how the principles of confidentiality protection, useful social–spatial linkages, data sharing, and data preservation are currently in conflict. This situation is not permanent but instead requires careful thought, research, and debate. We do not have solutions but conclude by noting why ignoring the problem is unacceptable, how consensus does not yet exist on what level of risk of confidentiality breach is acceptable or even how to estimate that risk, and how new institutional arrangements for making linked data available need to consider the burdens imposed on legitimate researchers.

Ignoring Is Unacceptable. At the front line of those who protect the confidentiality of research subjects are the investigators who are responsible for the collection of such information. In the past, their traditional practice was either to strip away all personal identifiers, such as names and addresses, before making data available to the broader research community or to release tabular data that had been aggregated to the point where no information was provided about any individual. The aggregation and tabular-release approach could be used with spatially linked data, but doing so precludes the microlevel analyses necessary for a large fraction of the most pressing research problems. Stripping away the personal identifiers means stripping away the spatial links, and, hence, removing locational identifiers. Either approach, tabular or stripping, means that only a small number of investigators will have access to the full power of the data and is, therefore, unacceptable if the scientific community is going to make progress in understanding a variety of microlevel social and environmental processes.

Acceptable Risk Level. There is no consensus now about what level of risk of confidentiality breach is acceptable given the research benefits that might accrue if linked data were made available outside the institution that collected the data. The example with Add Health data used a 1-in-20 median risk level for disclosing the primary sampling unit. Although this level was for illustration and disclosure of the primary sampling unit and not the respondent, it is clear that a 1-in-20 disclosure risk of a respondent's identity is totally unacceptable. What would be acceptable: 1 in 100? 1 in 1 million? Should the median risk level be our guide, or should we discuss acceptable levels for the maximum risk? If, instead of a 1-in-20 risk level for the median school, we chose a 1-in-20 risk level for each and every school, the buffer around schools to achieve this risk level would be 255.6 km, not 20.5 km. Does the level of acceptable risk depend on the type or magnitude of societal benefit that would be gained by answers to the research question? These questions need to be aired by all actors who have a stake in the outcome (e.g., respondents, data-collecting investigators, funders, secondary data researchers, institutional review boards, archivists, and journal editors).

Nor is there consensus on the disclosure risks associated with various approaches that have already been tried. When spatial explicitness is added to the tabular approach, disclosure risk clearly increases, but basic research is just beginning to reveal the size of this increase. A common current practice is to share the linked data with other researchers only after these researchers have agreed to a data security plan. Although there is not yet published research on the safety of this approach, anecdotal reports suggest that compliance with such security pledges can be flawed. Even the use of enclaves, where the researcher's activities are carefully monitored, carries unknown risks. Despite these unknown risks, most who have worried about these issues believe that, among current options, enclaves carry the lowest

risk of confidentiality breaches. We turn now to our last point, namely the relationship between the current manner in which enclaves have been structured and the lives of researchers.

Enclaves and Researchers. A number of enclaves have now been set up around the United States. Researchers conduct their research at the enclave, and all of the output that they take from the enclave is carefully monitored from the perspective of disclosure risk. With few exceptions, enclaves are operated by federal data-collection agencies or those responsible for large data collections, i.e., the data producers set up these enclaves. Because they want to maintain control over the security of the enclave, they tend to set them up at their own institution. The U.S. Census Bureau is the main exception, and even here there are a very small number of census enclaves.

There are equity concerns with the use of enclaves, because some charge fees and users must travel to the facilities. Moreover, enclaves do not always work well for interdisciplinary research teams. The team member with the technical skills necessary to work in the enclave may not be able single-handedly to incorporate the ideas of the full team. Numerous decisions are made in the course of analysis that might be handled differently if all team members could see all of the intermediate output. In sum, the enclave is not an ideal data-dissemination approach.

We believe that the scientific community needs a broader understanding of the potential confidentiality breach problems that are part of linking census or survey data to spatially explicit data. We need research on the risks associated with various protection schemes and with the display of linked data in publications and the like. It will be necessary to quantify the known and, to the extent possible, unknown risks. Simulated microdata provide a promising avenue for evaluating risk without risking disclosure of survey respondents. Although that exercise is beyond the scope of this paper, it is an obvious next step in which the research community should collectively invest. Once we further understand the problem, we need to forge solutions that both protect respondents and make critical data available to relevant members of the research community. In the current research milieu, explicit future funding opportunities for evaluating these risks and developing new methods to overcome these risks could be an important catalyst for finding acceptable solutions. These opportunities should be designed to incorporate the many stakeholders and perspectives that we have laid out here.

This paper grew out of a session at the 2005 meeting of the Population Association of America, organized by Rebecca L. Clark of the National Institute of Child Health and Human Development (NICHD). We thank John Spencer for assistance with analysis and maps and Tom Swasey and Lisa Isgett for preparation of figures. This work was supported by NICHD Grants 5 U24 HD048404 (to B.E. and M.P.G.), P01 HD045753 (to M.P.G.), R01 HD25482 (to B.E. and R.R.R.), and R01 HD35811 (to L.K.V.); and National Aeronautics and Space Administration Grant NAS5-03117 (to D.L.B.).

1. Association of American Geographers (1998) *Statement on Professional Ethics*. Available at www.aag.org/publications/other%20pubs/ethicsstatement.html.
2. American Psychological Association (2002) *Am. Psychol.* **57,** 1060–1073.
3. American Political Science Association (1998) *A Guide to Professional Ethics in Political Science* (Am. Political. Sci. Assoc., Washington, DC), 2nd Ed.
4. American Sociological Association (1999) *Code of Ethics and Policies and Procedures of the ASA Committee on Professional Ethics* (Am. Sociol. Assoc., Washington, DC).
5. Onsrud, H., Johnson, J. & Lopez, X. (1994) *Photogramm. Eng. Rem. S.* **60,** 1083–1095.
6. Croner, C. M. (2003) *Annu. Rev. Pub. Health* **24,** 57–82.
7. Zayatz, L. (2002) in *Inference Control in Statistical Databases*, ed. Domingo-Ferrer, J. (Springer, Berlin), pp. 181–202.
8. Dunn, C. S. & Austin, E. W. (1998) *ICPSR Bull.* **19,** 1–8.
9. O'Rourke, J. M. (2003) *ICPSR Bull.* **24,** 3–9.
10. Duncan, G. & Lambert, D. (1989) *J. Bus. Econ. Stat.* **7,** 207–217.
11. Lambert, D. (1993) *J. Off. Stat.* **9,** 313–331.
12. Duncan, G. T. (2001) in *International Encyclopedia of the Social and Behavioral Sciences*, eds. Smelser, N. J. & Bates, P. B. (Elsevier, London), pp. 2521–2525.
13. Willenborg, L. & de Waal, T. (1996) *Statistical Disclosure Control in Practice* (Springer, New York).
14. Willenborg, L & de Waal, T. (2001) *Elements of Statistical Disclosure Control* (Springer, New York).
15. Doyle, P., Lane, J. I., Theeuwes, J. J. M. & Zayatz, L. V., eds. (2001) *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies* (Elsevier, New York).
16. Abowd, J. M. & Lane, J. (2004) in *Privacy in Statistical Databases*, eds. Domingo-Ferrer, J. & Torra, V. (Springer, New York), pp. 282–289.
17. Rubin, D. B. (1993) *J. Off. Stat.* **9,** 461–468.
18. Armstrong, M., Rushton, G. & Zimmerman, D. (1999) *Stat. Med.* **18,** 497–525.
19. Goodchild, M. F., Anselin, L. & Deichmann, U. (1993) *Environ. Plan. A* **25,** 383–397.
20. Cressie, N. A. C. (1993) *Statistics for Spatial Data* (Wiley, New York).