Working Paper

ENHANCING DATA SHARING VIA "SAFE DESIGNS"

Kristine M. Witkowski

Inter-university Consortium for Political and Social Research,

University of Michigan

Working Paper

ICPSR Working Paper Series

March 2014

ENHANCING DATA SHARING VIA "SAFE DESIGNS"

Kristine M. Witkowski

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

The social value of data collections are dramatically enhanced by the broad dissemination of research files and the resulting increase in scientific productivity. Currently, most studies are designed with a focus on collecting information that is analytically useful and accurate, with little forethought as to how it will be shared. Both literature and practice also presume that disclosure analysis will take place after data collection. But to produce public-use data of the highest analytical utility for the largest user group, disclosure risk must be considered at the beginning of the research process. Drawing upon economic and statistical decision-theoretic frameworks and survey methodology research, this study seeks to enhance the scientific productivity of shared research data by describing how disclosure risk can be addressed in the earliest stages of research with the formulation of "safe designs". Implications for various research costs are also discussed.

Key words: Confidentiality; access modalities; risk/utility tradeoff; study design; survey design.

Acknowledgements: Research support from the National Institute of Child Health and Human Development (NICHD), Grant 1 R01 HD067184, is gratefully acknowledged. Special thanks are also given to George Alter and Michael Elliott for their thoughtful comments.

Contact Information: Please address all correspondence to Kristine 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: 734-615-9321.

Draft: Please do not cite without author's permission.

Introduction

Between 2000 and 2010, the National Institutes of Health (NIH) and the National Science Foundation (NSF) spent $XX billion on social and behavioral science research, with approximately XX% of these funds devoted to data collection [http://report.nih.gov/budget_and_spending/index.aspx]. The social value of a data collection is reflected by the amount of science produced from the analysis of its research data files. "Increasingly, NIH and the NSF have become interested in data sharing as a means of supporting the scientific process and ensuring the highest return on these investments." [Pienta, Alter, and Lyle 2010; p. 1; Need new phrase.] As Pienta, Alter, and Lyle (2010) have found, research data that are archived typically generate 2.42 times as many publications as collections solely utilized by the original research team (i.e., not shared), controlling for several principal investigator, institutional, and grant award characteristics.

Scientific productivity also depends on the mode under which data are shared (i.e., access modality). As a case in point, let us consider the Adolescent Health Survey (ADDHEALTH), a study which has disseminated both public-use and restricted data files. Tallying the number of journal articles, agency reports, unpublished manuscripts, meeting presentations, and the like (cite ICPSR's database), we find that their one public-use file resulted in 4,060 works. In contrast, all of their 18 restricted files had resulted in only 19 works. These simple statistics illustrate the degree to which access modality limits a dataset's production of science. Barriers to access stem from real and perceived informational, resource, and time constraints associated with acquiring restricted data. Poor publicity, inadequate documentation, difficulty of use, user fees, and administrative work associated with contracting, compliance offices and institutional review boards help overshadow the heightened analytical value of restricted research data (check citation: O'Rourke et. al. 2006). Consequently it is important that datasets are constructed with the long-term goals of unlimited access and wide-ranging utility, so that the social value of data collections are better enhanced.

However most studies are currently designed with a focus on collecting information that is analytically useful and accurate, with little forethought as to how it will be shared. Both literature and practice also presume that disclosure analysis will take place after data collection. But to produce public-use data of the highest analytical utility for the largest user group, disclosure risk must be considered at the beginning of the research process. Statistical methods, as they are applied to existing research data, are hard-pressed when it comes to ensuring safe and unlimited access to complex microdata with sensitive content (e.g., persons nested within spaces, places, and time reporting risky behaviors) comprised of detailed and accurate measures. Yet it is possible that studies can be designed and executed in such a way that subject identities are sufficiently obscured so that competing goals of data confidentiality, utility, access and cost are optimally met. Drawing upon economic and statistical decision-theoretic frameworks and survey methodology research, this study seeks to enhance the scientific productivity of shared research data by describing how disclosure risk can be addressed in the earliest stages of research with the formulation of "safe designs". Implications for various research costs are also discussed.

The Role of "Safe Designs" in a Portfolio Approach to Data Sharing

Reflecting the concentration of U.S. federal agency efforts, there exists a considerable amount of academic and grey literature on the various statistical and technical methods for protecting confidential data of research subjects. But as Lane (2007, p. 300) argues: “[However,] focusing on confidentiality protection alone is likely to lead to piecemeal approaches and result in outcomes that are in the best interest neither of decision-makers nor of society at large. The appropriate approach is to optimize the amount of data access, subject to meeting key confidentiality constraints."

Formulating a decision-theoretic framework encompassing a broad range of statistical, economic, and social determinants, Lane (2007) casts data custodians in the role of ensuring that the societal value to microdata access (U, data utility) meets or exceeds expected societal costs (S), where the method in which confidential data are accessed (access modalities, M_i) is an important determinant of both. Since a single modality is likely not meet all sharing needs, a portfolio of releases holds the key to the optimal dissemination of a study's data, contingent on a mix of data products that maximizes social value and minimizes social cost. Paraphrasing Lane (2007, pp. 310-311), the value to society (U) depends on data quality (Q), researcher quality (R), and the number of times the data are accessed (N). On the other hand, the costs to society (S) stems from the harm incurred to individuals and institutions should a disclosure occur (H), times the probability of a disclosure (D), plus the monetary cost of providing access (C). A host of factors underlie H, D, and C that consist of: the existence and accessibility of other data sources used in reidentification (E), existence of malevolent interlopers (I), population characteristics (X), researcher error (Z), technology (T), legal penalties (L), training/adoptable protocols (P), and the price of providing a certain level of protection (p).

Lane and her colleagues (Lane, Heus, and Mulcahy 2008) go on to illustrate how cyberinfrastructure can be used to implement their portfolio approach by addressing the five principles of "safe projects, safe data, safe people, safe settings, and safe conduct" (citation of presentation or particular page of Lane, et.al. 2008). The foundation of this data sharing system rests on remote access and the integration of various technical, educational, operational, statistical and legal tools. Following established practice and the existing literature, this data sharing approach also presumes that data will be processed for disclosure limitation once it has been collected and that emergent confidentiality issues are the key constraining factor (Lane 2007). With the ever-growing technological threats to privacy and difficulties with implementing statistical tools, Lane et. al. (2008) focus their discussion on how restrictive technologies, legalities, and protocols and accompanying training can help ensure that "safe people", within "safe places", have "safe access" to "safe data". While public-use data still plays an important role in this system, severe constraints are placed on the production of these preeminent and confidentially-formidable data.

In turn, we expand upon Lane's portfolio approach by incorporating the concept of "safe designs". With the explicit goal of maximizing the utility of public-use files, this principle's foundation rests on addressing statistical disclosure risk in the earliest stages of research, where sampling designs are specifically formulated to minimize the probability that subjects are reidentified. As a result, data sharing is constrained by the marginal cost associated with supplementing study designs so that disclosure risk is sufficiently low.

As an expository scenario, let us consider the case where a data producer (embarking on a new study) has formulated a survey instrument to collect information to be disseminated as microdata, where the content of a research data file is of the highest quality (Q) to be used by researchers of a particular quality (R). The producer then formulates a safe study design consisting of two components: (1) a fundamental sample of sufficient size to meet the study's analytical purposes, as estimated by power analyses (Y_0); and (2) a supplemental sample that is an expansion of the baseline design specifically formulated to minimize disclosure risk (Y_S).

The producer must then assess the feasibility of implementing this safe study design (Y_0, Y_S) and releasing a predefined set of research data as a public-use file (M_P), rather than implementing a traditional study design (Y_0) and releasing the research data as a restricted-use file (M_R). The selection of access modality depends on the relative amount of social value (U_P , U_R) and social cost (S_P , S_R).

Justification for releasing a public-use file is found when its estimated return-on-investment is greater than releasing it as a restricted-use file (Equation 1).

$$U_P/S_P \quad > \quad U_R/S_R \qquad (1)$$

Social value is reflected by the number and impact of publications associated with a predefined set of research data accessed through a particular modality (U_P , U_R) (citations from Pienta, Alter, and Lyle 2010). Continuing with our scenario, the producer expects that the social value of public-use data is a consequence of the relatively higher number of times these data are accessed, as compared to disseminating them through a restrictive modality ($N_P > N_R$; holding constant Q and R).

With this done, the producer turns to the work of estimating the social cost associated with each access modality (S_P , S_R). Building on Lane's definition of social and monetary cost (Equations 2 and 3) and an associated constraint (Equation 4), we formulate Equations 5 and 6 for purposes of a feasibility study.

$$S = HD + C \qquad (2)$$

$$C = p_t T + \sum_{Mi} p_{Ai} M_i \qquad (3)$$

$$S - C - HD^* \leq 0 \qquad (4)$$

$$p_A = f\,(p_S\,,\, p_D) \qquad (5)$$

$$[U_P\,/\,(p_{S|P} + p_{D|P})] \quad > \quad [U_R\,/\,p_{D|R}] \qquad (6)$$

Given a predefined set of research data, the producer stipulates the same sensitive content for public-use and restricted data releases. Consequently, the harm associated with disclosing this information ($H_P = H_R$) is held constant across access modality. The producer then determines the conditions under which data can be safely disseminated through each access mode, defined as the target risk of disclosure (D^*). As Lane (2007, p. 310) describes: "The probability of disclosure is typically set at a 'target' level: since most agencies are charged with using reasonable means to protect data, this implicitly means setting reidentification risk to some fixed number." This safety limit reflects concerns about the degree to which researcher error (Z) may provide opportunities for existing malevolent intruders (I) to acquire research data (M) and associate it with other extant data (E) so as to heighten the probability of reidentifying study subjects (D). For purposes of our discussion, we assume that the data producer seeks to lower the probability of reidentification to the same "negligible" level for both public-use and restricted data ($D^*_P = D^*_R \sim 0$). The ability to meet these safety standards is contingent on the monetary cost of providing access (C), as defined by resources devoted to minimizing the probability of disclosure (D) such that $D < D^*$ (Equation 4).

Having set these parameters, the data producer begins his/her feasibility assessment by estimating the amount of disclosure risk (D_0) associated with fundamental and safe study designs (Y_0, Y_S). Next the producer assesses the degree to which different survey and dissemination activities lowers disclosure risk, compiling estimates of disclosure risk (D) and monetary costs (C) associated with alternative modalities. Bringing together these disclosure and cost estimates, producers can better assess the value added for different expenditures (Lane 2007).

In Equation 3, Lane defines two dimensions of monetary cost as they relate to: (1) a preexisting menu of technologies and institutional supports used in data sharing (p_tT) and (2) the provision of "a certain level of protection" for each access modality ($p_{iA}M_i$). But to ascertain the benefits of addressing data confidentiality in the earliest stages of research, it is necessary to fully articulate monetary costs for data collection activities that strictly enhance data sharing (p_S) as well as data dissemination activities

(p_D). For purposes of a feasibility assessment, we formulate Equation 5 where the monetary cost of public and restricted access modalities (p_A) is a function of p_S and p_D, where the expense of data custodian infrastructures (p_tT) and the preemptive disclosure review of a fundamental sampling design and its actual data collection (p_0) are considered sunken costs that are exogenous to feasibility decisions and therefore are set to 0.

We also assume restricted access precludes the need for a safe design ($p_{S|R} = 0$), where producers follow the established practice of conducting disclosure analysis later in the data lifecycle along with implementing a refined set of technical, educational, operational, statistical and legal tools ($p_{D|R}$). In contrast, the public release of research data requires that a responsive survey design be formulated and implemented ($p_{S|P}$); and that an ex post disclosure review also be executed to reveal confidentiality shortcomings that are subsequently addressed ($p_{D|P}$). Having outlined our "safe design" approach to data sharing and bringing all decision-making factors together in Equation 6, we now turn to the data confidentiality and survey research literatures to describe different data collection and dissemination activities associated with implementing safe designs (as compared to those traditionally used), revealing trade-offs in monetary cost.

The Implementation of "Safe Designs"

The literature on statistical disclosure control (SDC) has assessed how different study design factors (implemented at data collection) effect reidentification, among these are: (1) the absolute size of the study population, as defined by known geographic boundaries; and (2) the sampling rates of the study population.[38-40] Holding constant these two study design elements, researchers have also assessed how disclosure risk is shaped by database design factors (considered just prior to dissemination), particularly: (3) the types of identifying personal characteristics of subjects (e.g., age, gender, race/ethnicity) and directly identified geographic locations (e.g., state, zip code); and (4) the application of various disclosure limitation methods (DLMs).[41-42] Finally, researchers have investigated how different DLMs affect the analytical utility of research data and the ability to broadly disseminate these files.[43-44] These studies often state that restricted access modes are the only methods for retaining the analytical utility of anonymized data, in that not much more can be done since data collection has been completed.[45]

In conducting a feasibility assessment in lieu of a data collection, this belief may be refuted by flipping various empirical assumptions on their head. Like SDC research, research informing safe designs assumes: (1) a study population that is sampled at a rate meeting fundamental analytical purposes; and (2) a survey design that gathers a complexity of data where individuals are nested within space, place, and time. But unlike SDC's assessments of how risk varies with database design parameters, we assume a single predefined wish list of attributes, comprised of a data release whose informational content is of the highest analytical utility as defined by numerous detailed measures that are unperturbed. But what is allowed to vary is the supplementing of data collection efforts to meet articulated data sharing goals. Activities performed in the design and implementation of safe design can usurp or negate those required for sharing restricted-use file, enhancing the social value of research data by shifting the monetary costs for data sharing from the end of the data cycle to its beginning.

Before and During Collection: Circumventing Disclosure Risk

A priori knowledge of disclosure factors – specifically the number and composition of respondents who are at-risk of reidentification – allows data producers to responsively modify their study design, such that data collection efforts are extended to meet pre-specified disclosure goals. As

Groves and Heeringa [54] (p. 440) discuss, producers formulate responsive study designs by: (1) preidentifying design features potentially affecting costs and errors of survey estimates; (2) identifying indicators of cost and error properties of these features and monitor indicators in initial phases of data collection; (3) altering survey features in subsequent phases based on cost-error trade-off decision rules; and (4) combining data from separate design phases into single estimator. Data producers capitalize on large amounts of information about how well their survey efforts are meeting data collection goals (i.e., paradata), particularly as it relates to statistical inference and cost efficiencies. We build upon this survey methodology by conceptually integrating disclosure risk into responsive study designs, suggesting empirical data that can be used in their formulation.

Studies with the highest social value tend to be those that gather sensitive information from respondents. Study designs have also increasingly become complex, with geography playing an important role in either sampling designs or informational content or both. Consequently we focus our discussion on a scenario where a producer wishes to disseminate sensitive microdata containing measures of identifying personal characteristics of subjects (e.g., age of respondent) and attributes of geographic locations (e.g., proportion of population in respondent's neighborhood that is poor) to be used in either fixed-effect or hierarchical linear models. Furthermore we consider a disclosure scenario that is empirically conservative, where the largest threat to confidentiality is an "acquaintance" intruder seeking to pinpoint a known subject's record within a research database. A subject is considered easily reidentified when they are a sample unique, as when k-anonymity (e.g., k=3) is not achieved.[32-38] So when producing responsive surveys, producers seek to formulate sampling designs that generate sets of 3 respondents sharing the same identifying personal and geographic attributes, minimizing the number of conspicuous subjects.

Given this definition of disclosure risk and sampling goal, the first step in constructing a responsive safe design is the preemptive disclosure review of a project's fundamental (or baseline) sampling design. Simulated data that mimic a baseline sampling design is required, where the number of synthetic records sharing the same identifying attributes is tallied. Predicted estimates of the number and composition of sample uniques help define the scope of a study's disclosure issues.

These at-risk populations are then targeted in an ex ante modification of the baseline sample design. When addressing k-anonymity, producers rely on well-established surveying techniques for oversampling,[30] where emphasis is placed on locating populations of interest within pre-specified geography. Simulations are again required to estimate the number of additional draws from a population that are needed to meet disclosure goals. The spatial dimensions of a sampling design are only modified when targeted recruitment becomes exceedingly expensive, where geographies resembling those originally surveyed (bounded within primary sampling units) are brought into study in hopes of easing enumeration.[1]

Knowing how many people must be targeted for recruitment and locations where sampling quotas can best be filled, the last step of formulating a responsive design is estimating the cost

[1] The decisions underlying the modification of geography samples take a different form when addressing the disclosure risk associated with "stranger" intruders and population reidentification probabilities. In this case, the intruder is searching for an unfamiliar study subject within the general population, where a respondent is considered easily reidentified when there are a limited number of look-alike persons within a known location. For instance, geographies with populations of 100,000 or more are typically identified in microdata files (such as the Current Population Survey) because of the rare chance that the sample will contain a population unique. Unlike the approach to k-anonymity, producers are not particularly concerned that a respondent stands out among other study subjects (i.e., sample unique). Instead, disclosure risk is reduced by the amassing of populations dispersed across geographic units sharing the same contextual attributes (Witkowski citation), where the scope of study (e.g., known state) and the areal size of geographies (i.e., counties, tracts, blockgroups, pixels) are defining features of the intruder search. Consequently the formulation of safe designs relies on radical changes to a sampling design, particularly as it relates to expanding the number of geographic-based primarily sampling units and/or the spatial scale of contexts. For instance, data collections that characterize a sample of blocks drawn from a single, publicized site are likely to face formidable disclosure issues. An alternative safe design would be to sample tracts drawn from a large number of sites whose identities have been concealed.

associated with its enactment ($p_{S|P}$). Supplemental sampling will be relatively expensive since hard-to-count, precisely-defined populations are the target of safe designs. Therefore producers will be particularly concerned with improving the efficiency of survey efforts, where methods will likely include the use of administrative data, refined screening techniques, and on-the-fly disclosure review during data collection [Need citations].

After Collection: Ensuring Data are Safely Disseminated

Once data have been collected, producers must conduct an ex post disclosure review to reveal any confidentiality shortcomings. If necessary, additional statistical protections are then applied to the research data, with a final review verifying that they can be safely released though a particular access modality.

When enacting safe designs, much of this disclosure work has occurred in the early phases of research. Simulations predict confidentiality problems that are subsequently addressed in supplemental surveying efforts. On-the-fly disclosure review also provides a large part of the information gleaned from ex post reviews. If all goes well, any remaining disclosure risk should be minimal and can be readily addressed through statistical means, where a public-use file can be subsequently released with little additional expense ($p_{D|P} \sim 0$). But when a study is enacted without regard to data sharing outcomes, the monetary cost of this disclosure work is accrued late in the data lifecycle ($p_{D|R} > 0$). This expense is likely lowered by efficiencies gained from performing a preemptive disclosure review of the study's fundamental study design (Y_0), where predicted findings are used to guide ex post processing of restricted data. Hence the value of this early work is retained, justifying a safe design approach regardless of whether it is fully implemented.

With this disclosure work completed, the producer then engages a variety of technical, educational, operational and legal tools so that any remaining disclosure problems can be circumvented and a product can be safely shared via restricted access ($p_{D|R} > 0$, $p_{D|P} = 0$) (Lane, et. al. 2008). [4] [FOOTNOTE: None of these costs exists for the public-use modality, since all previous collection and statistical work have rendered these data safe.] In so doing, the producer will likely rely on established infrastructures to provide non-statistical techniques, whose developmental costs have long been incurred or are being absorbed by the larger scientific enterprise. Consequently a producer is primarily concerned with expenses stemming from (1) maintaining and expanding the capacity of a restrictive mode; (2) meeting its compliance standards; and (3) training data users about how compliance goals can be effectively and efficiently met. When assessing the trade-offs in social cost between access modes, producers must bear in mind that these expenses are a function of the intensity and duration of a modality's use and, therefore, are directly related to its social value.

For restricted-use agreements, mode capacity is contingent on the ability to efficiently negotiate and process material transfer agreements, where compliance is defined by data-usage rules stipulated in contracts and is enforced by monitoring and/or legal sanctions. For a virtual data enclave (VDE), mode capacity is contingent on cybertechnologies that limit disclosure risk while supporting a complexity of analyses, where compliance is defined by the confidentiality of research findings extracted from the VDE and is enforced by their disclosure review. Training reduces the expense of compliance by teaching researchers the importance of protecting confidential data and how they can most effectively meet compliance goals (Lane, et. al 2008). The selection of restricted access modes depends on the sensitivity of a study's content (or harm, H) and the residual probability of disclosure (D_R) after statistical techniques have been applied to render data of a particular analytical quality (Q). With H and Q held constant, data of the highest D_R are typically disseminated through a VDE, while those having a sufficiently low D_R are accessed through licensing agreements.

Now that a file is ready for dissemination and a modality has been chosen, the producer must make a formal announcement of the existence of a data product, its scientific value and how it can be acquired; thereby enhancing the social value of a product by increasing the likelihood of it being accessed and analyzed. Publicizing activities include producing a website and promotional materials for distribution through various venues (i.e., electronic announcements, conferences) as well as assigning staff to answer questions from would-be and actual users.

When estimating the expense of publicizing, it is important to consider the release order of a product as well as its trajectory of data usage. For studies releasing their first data set, a producer is likely to devote significant resources launching their flagship product, with initial expenditures varying little between access modality. Subsequent releases are likely to be free riders, piggybacking on flagship dissemination tools. This is especially true for restricted files since relatively little social value will accrue from a new launch ($p_{D|R} < p_{D|P}$). For purposes of our feasibility assessment, we have assumed that a data product is the first of a study, where this expense is mode invariant and set to zero ($p_{D|R} = p_{D|P} = 0$).

The number of times a data file is accessed typically tapers off over time. When access bottoms out and the benefits of publicizing a dataset are null, it is logical that these expenditures are curtailed. Restricted-use files will likely experience lower absolute levels and faster declines in access activity (as compared to public-use files of the same data utility), resulting in the abbreviation of supports and lowered publicity costs ($P_{D|R} < P_{D|P}$).

Finally, producers need to consider the cost of archival activities (Lane, et. al. 2008) that support data sharing by helping users to: (1) locate a study's research data (e.g., metadata, measurement identification, cross-referencing in bibliographies and data listings); (2) correctly analyze its data (e.g., codebooks and study documentation); and (3) utilize its preserved data (e.g., upgrading of formats). These expenses increase with the amount and complexity of research data. But when conducting a feasibility assessment for a predefined set of data (where file content is held constant), costs of searchability and usability enhancements and preservation are assumed not to vary with access modality; and consequently these expenses are set to zero ($p_{D|R} = p_{D|P} = 0$).

We have just described the monetary costs associated with specific project's safe design approach. However some of this expense can be reduced with expanding the data sharing infrastructure (p_tT).

In Support of "Safe Designs": A Call for Applied Research

As Lane, et. al. (2008, p. XXX) argue: "a major challenge to the data privacy community is developing disclosure limitation techniques that are flexible enough to be used in a wide variety of situations" (GET EXACT QUOTE). The "safe design" approach is such a technique, given its potential to circumvent a wide range of k-anonymity confidentiality problems by addressing disclosure risk in the design of samples.

Safe designs maximize the social value of research data by eradicating the need to use analytically-devaluating statistical techniques and restrictive access modalities. However the heightened productivity of these research data comes at a price as determined by five research activities: (1) preemptive disclosure review of a fundamental design; (2) estimation of monetary costs for a supplemental survey, statistical disclosure processing, restricted access, publicizing, and archival; (3) construction of supplemental survey plan; and (4) implementation of supplemental survey. These research demands can be an inordinate burden to data producers since the planning and implementation of safe designs requires highly-specialized knowledge in the realms of statistical disclosure limitation methods, access modality costs, and survey research. Therefore, if the safe design approach is to be a viable disclosure limitation technique, it is imperative that applied science be conducted to inform all of these activities.

When looking for project financing, producers must be able to efficiently and accurately predict the resource needs of their safe designs, justifying these expenditures to would-be funders. The design and budgeting of projects (under unbillable resource constraints) is best informed by research that generates estimates of disclosure risk, data utility, and monetary costs for a variety of stylized prototypes (characterized by access modality, population of interest, complex design, sampling rate, sensitivity of content, geographic specificity, identifying variables) for use by the broader scientific community. Producers can search among these metadata to identify study/database designs and access modes that best address their project needs, extracting estimates which can then be used for planning and budgetary purposes. These simulated meta data represent a complexity of guidelines that take into account all aspects of the data sharing process, where rules-of-thumb are likely to emerge as knowledge evolves from this "multifaceted approach" (term coined by Lane 2007).

The foundation of this method rests on small area estimates of detailed study populations that allow for the accurate prediction of disclosure risk (D) and utility (Q) outcomes for varying design parameters. This method also builds upon expenditure data reported by survey researchers and archivists, where detailed cost information (C) is gathered on various data collection, dissemination, and archival activities. To round out the picture, this method would benefit from incorporating typological estimates of the number of times accessed and the number of works produced, two measures that capture a study's social value (S), based on observation data compiled from the administrative systems of custodians. Lastly, an assessment of predicted data sharing outcomes is necessary so that modeling inaccuracies can be identified and rectified.

This applied research directly addresses the first two research tasks of safe designs, that of the preemptive disclosure review and monetary cost estimation. Taking this work a step further, models predicting disclosure risk can also be extended to the third research task, the construction of a supplemental survey plan. Derived from sampling designs for stylized or actual studies, detailed simulated data can be formulated to guide the selection of survey sites so that recruitment goals are efficiently met. In sum, this applied research has the potential for generating significant efficiencies, representing a key infrastructural support for data producers wishing to adopt a safe design approach.

Reference List

Abowd, J. and J. Lane. 2003. The Economics of Data Confidentiality. Mimeo, Committee on National Statistics. www7.nationalacademies.org/cnstat/Abowd_Lane.pdf.

Barth, A., A. Datta, J.C. Mitchell, and H. Nissenbaum. 2006. “Privacy and Contextual Integrity: Framework and Applications.” 27th IEEE Symposium on Security and Privacy, IEEE Computer Society. [Cited by Lane, Heus, and Mulcahy 2008]

Weber, T.M. 2005. “Values in National Information Infrastructure: A Case Study of the US Census.” 14th International Conference of the Society of Philosophy and Technology, Delft, The Netherlands. [Cited by Lane, Heus, and Mulcahy 2008]

Burstein, P. 1991. “Policy Domains: Organization, Culture, and Policy Outcomes.” Annual Review of Sociology, 17: 327-350.

Cummings, J., T. Finholt, I. Foster, C. Kesselman, and K. Lawrence. 2008. “Beyond Being There: A Blueprint for Advancing the Design, Development and Evaluation of Virtual Organizations.” [Cited by Lane, Heus, and Mulcahy 2008]

Cummings, J. and S. Kiesler. 2007. “Coordination Costs and Project Outcomes in Multi-University Collaborations.” Research Policy, 36: 1620-1634. [Cited by Lane, Heus, and Mulcahy 2008]

Duncan, G.T., S. Keller-McNulty, and S.L. Stokes. 2003. “Disclosure Risk vs. Data Utility: The R-U Confidentiality Map.” Technical Report 2003-6. Heinz School of Public Policy and Management, Carnegie Mellon University.

Eccles, R. 1985. The Transfer Pricing Problem: A Theory for Practice. Lexington, Mass: Lexington.

Fienberg, S.E., A. Anton, E. Bertino, C. Dwork, E. Viegas, and L. Zayatz. 2007. Workshop on Data Confidentiality, Arlington, VA. [Cited by Lane, Heus, and Mulcahy 2008]

Foster, I., C. Kesselman, and S. Tueck. 2001. "The Anatomy of the Grid: Enabling Scalable Virtual Organizations." International Journal of Supercomputer Applications, 15: 200-222.

Human Subjects Research Subcommittee, Committee on Science. "Achieving Effective Human Subjects Protection and Rigorous Social and Behavioral Research." [Cited by Lane 2007.]

Lane, J. 2007. "Optimizing the Use of Microdata: An Overview of the Issues." Journal of Official Statistics, 23(3): 299-317.

Lane, J., P. Heus, and T. Mulcahy. 2008. "Data Access in a Cyber World: Making Use of Cyberinfrastructure." Transactions on Data Privacy, 1: 2-16.

LeClere, F. 2010. "XXXXXXXXXXXXXXXX." Chronicle of Higher Education, X: XX-XX.

McFarland O'Rourke, J. XXXX. "XXXXXXXXXXXXXXXX." XXXXXXXXXXXX, X: XX-XX.

National Institutes for Health. XXXXa. "XXXXXXXXXXXXXXXX." XXXXXXXXXXXX, X: XX-XX.

National Institutes for Health. XXXXb. "XXXXXXXXXXXXXXXX." XXXXXXXXXXXX, X: XX-XX.

National Science Foundation. XXXXa. "XXXXXXXXXXXXXXXX." XXXXXXXXXXXX, X: XX-XX.

National Science Foundation. XXXXb. "XXXXXXXXXXXXXXXX." XXXXXXXXXXXX, X: XX-XX.

National Science and Technology Council, President's Information Technology Advisory Committee (PITAC). Report on Cyber Security.

Numerous studies undertaken by the National Academy of Sciences, the Committee on National Statistics, and the National Science Foundation (area of cyber trust), and the PORTIA Project (Privacy, Obligation and Rights in Technologies of Information Assessement). [Cited by Lane 2007]

Oberschall, A. and E.M. Leifer. 1986. "Efficiency and Social Institutions: Uses and Misuses of Economic Reasoning in Sociology." Annual Review of Sociology, 12: 233-253.

Pang, L. 2001. "Understanding Virtual Organizations." Information Systems Control Journal, 6.

Pienta, A. XXXX. "XXXXXXXXXXXXXXXX." XXXXXXXXXXXX, X: XX-XX.

Smith, J. 1991. "Data Confidentiality: A Researcher's Perspective." Proceedings of the American Statistical Association, Section on Social Statistics, 117-120. [Cited by Lane 2007. Could not find.]

Trottini, M. 2001. "A Decision-Theoretic Approach to Data Disclosure Problems." Research in Official Statistics, 4: 7-22.

United National Economic Commission for Europe. 2006. "Managing Statistical Confidentiality and Microdata Access." http://www.unece.org/stats/documents/tfcm/l.e.pdf. [Cited by Lane 2007]

UNECE/Eurostat conference. 2005. Monographs of official statistics, Work session on statistical data confidentiality, Geneva 9-11 November 2005, ISBM 92-79-01108-01.

Williamson, O.E. 1975. Markets and Hierarchies: Analysis and Antitrust Implications. New York: Free Press.

Williamson, O.E. 1981. "The Economics of Organization: The Transaction Cost Approach." American Journal of Sociology, 87: 548-577.

Williamson, O.E. 1984. "The Economics of Governance: Framework and Implications." Journal of Institutional Theoretical Economics, 140: 195-223.

1. Goodchild, M.F. and R.P. Haining, GIS and spatial data analysis: Converging perspectives. Papers in Regional Science, 2004. 83(1): p. 363-385.

2. Diez Roux, A.V., Investigating neighborhood and area effects on health. Am J Public Health, 2001. 91(11): p. 1783-9.

3. Donnelly, P.G., An evaluation of the effects of neighborhood mobilization on community problems. J Prev Interv Community, 2006. 32(1-2): p. 61-80.

4. Ewing, R., Can the physical environment determine physical activity levels.? Exercise and Sport Sciences Reviews, 2005. 33(2): p. 69-75.

5. Fisher, J.B., M. Kelly, and J. Romm, Scales of environmental justice: combining GIS and spatial analysis for air toxics in West Oakland, California. Health Place, 2006. 12(4): p. 701-14.

6. Galea, S., J. Ahern, M. Tracy, and D. Vlahov, Neighborhood income and income distribution and the use of cigarettes, alcohol, and marijuana. American Journal of Preventive Medicine, 2007. 32(6): p. S195-S202.

7. Glass, T.A., M.D. Rasmussen, and B.S. Schwartz, Neighborhoods and obesity in older adults the Baltimore memory study. Am J Prev Med, 2006. 31(6): p. 455-63.

8. Gorman, D.M., P.W. Speer, P.J. Gruenewald, and E.W. Labouvie, Spatial dynamics of alcohol availability, neighborhood structure and violent crime. Journal of Studies on Alcohol, 2001. 62(5): p. 628-36.

9. Grady, S.C., Racial disparities in low birthweight and the contribution of residential segregation: a multilevel analysis. Soc Sci Med, 2006. 63(12): p. 3013-29.

10. Greene, N.A., J.D. White, V.R. Morris, S. Roberts, K.L. Jones, and C. Warrick, Evidence for

environmental contamination in residential neighborhoods surrounding the defense depot of Memphis, Tennessee. Int J Environ Res Public Health, 2006. 3(3): p. 244-51.

11. Hannon, L. and M.M. Cuddy, Neighborhood ecology and drug dependence mortality: an analysis of New York City census tracts. Am J Drug Alcohol Abuse, 2006. 32(3): p. 453-63.

12. Harlan, S.L., A.J. Brazel, L. Prashad, W.L. Stefanov, and L. Larsen, Neighborhood microclimates and vulnerability to heat stress. Soc Sci Med, 2006. 63(11): p. 2847-63.

13. Inagami, S., D.A. Cohen, and S.M. Asch, Neighborhood fast food concentration, location of grocery stores and body mass index. American Journal of Epidemiology, 2007. 165(11): p. S129-S129.

14. Ingoldsby, E.M., D.S. Shaw, E. Winslow, M. Schonberg, M. Gilliom, and M.M. Criss, Neighborhood disadvantage, parent-child conflict, neighborhood peer relationships, and early antisocial behavior problem trajectories. J Abnorm Child Psychol, 2006. 34(3): p. 303-19.

15. Israel, B.A., A.J. Schulz, L. Estrada-Martinez, S.N. Zenk, E. Viruell-Fuentes, A.M. Villarruel, and C. Stokes, Engaging urban residents in assessing neighborhood environments and their implications for health. J Urban Health, 2006. 83(3): p. 523-39.

16. Juhn, Y., J. Sauver, S. Katusic, D. Vargas, A. Weaver, and J. Yunginger, The influence of neighborhood environment on the incidence of childhood asthma: a multilevel approach. Social Science & Medicine, 2005. 60(11): p. 2453–2464.

17. Kinney, P.L., M. Aggarwal, M.E. Northridge, N.A.H. Janssen, and P. Shepard, Airborne concentrations of PM2.5 and diesel exhaust particles on Harlem sidewalks: A community-based pilot study. Environmental Health Perspectives, 2000. 108(3): p. 213-218.

18. Kipke, M.D., E. Iverson, D. Moore, C. Booker, V. Ruelas, A.L. Peters, and F. Kaufman, Food and park environments: Neighborhood-level risks for childhood obesity in east Los Angeles. Journal of Adolescent Health, 2007. 40(4): p. 325-333.

19. Kling, J.R., J.B. Liebman, and L.F. Katz, Experimental analysis of neighborhood effects. Econometrica, 2007. 75(1): p. 83-119.

20. Gutmann, Myron P., Kristine M. Witkowski, Corey Colyer, JoAnne McFarland O'Rourke, and James McNally. 2009. Providing Spatial Data for Secondary Analysis: Issues and Current Practices relating to Confidentiality. Population Research and Policy Review 27: 639-665.

21. National Research Council. 2007. Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data. Panel on Confidentiality Issues Arising from the Integration of Remotely Sensed and Self-Identifying Data. M.P. Gutmann and P.C. Stern, Eds. Committee on the Human Dimensions of Global Change. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

22. Rushton, Gerard, Marc P. Armstrong, Josephine Gittler, Barry R. Greene, Claire E. Pavlik, Michele M. West, and Dale L. Zimmerman. 2006. Geocoding in cancer research: A review. American Journal of Preventive Medicine 30: S16-S24.

23. Armstrong, Marc P., Gerard Rushton, and Dale L. Zimmerman. 1999. Geographically masking health data to preserve confidentiality. Statistics in Medicine 18: 497-525.

24. Saalfeld, Alan, Laura Zayatz, and Erik Hoel. 1992. Contextual variables via geographic sorting: A moving averages approach. In Proceedings of the Section on Survey Research Methods, 691-696. Alexandria, VA: American Statistical Association.

25. National Institutes of Health. 2003. Final NIH Statement on Sharing Research Data. Notice NOT-OD-03-032. Bethesda, MD. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

26. Groves, R.M., Fowler Jr., F.J.,Couper, M.P., Lepkowski, J.M., Singer, E. and Tourangeau, R. 2004. Survey Methodology. Hoboken, NJ: Wiley Interscience.

27. Tourangeau, R. 2004. Survey research and societal change. Annual Review of Psychology 55, 775-801.

28. Duncan, George T. and R.W. Pearson. 1991. Enhancing access to microdata while protecting confidentiality: Prospects for the future. Statistical Sciences 6: 219 – 232.

29. O'Rourke, JM, Roehrig, S, Heeringa, SG, Reed, BG, Birdsall, WC, Overcashier, M., Zidar, K. 2006. Solving Problems of Disclosure Risk While Retaining Key Analytic Uses of Publicly Released Microdata. Journal of Empirical Research on Human Research Ethics 1(3): 63-84.

30. Kish, Leslie. 1992. Weighting for Unequal Pi Journal of Official Statistics 8(2): 183-200.

31. National Opinion Research Center. 2009. General Social Survey. University of Chicago. Chicago, IL. http://www.norc.org/GSS+Website/

32. Domingo-Ferrer, J. and V. Torra. 2005. Ordinal, Continuous, and Heterogeneous k-anonymity through Microaggregation. Data Mining and Knowledge Discovery 11: 195-212.

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

34. Lambert, Diane. 1993. "Measures of Disclosure Risk and Harm." Journal of Official Statistics 9: 313-331.

35. Reiter, Jerome P. 2005. Estimating Risks of Identification Disclosure in Microdata. Journal of the American Statistical Association 100: 1103-1112.

36. Skinner, C.J. and D.J. Holmes. 1998. Estimating the re-identification risk per record in microdata. Journal of Official Statistics 14(4): 361-372.

37. 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.

38. Willenborg, Leon and Ton de Waal. 1996. Statistical Disclosure Control in Practice. Lecture Notes in Statistics 111, Springer, New York.

39. Greenberg, Brian and Laura Voshell. 1990. Two Notes on Relating the Risk of Disclosure for Microdata and Geographic Area Size. Survey of Income and Program Participation Working Paper Series (#9029). Washington, DC: U.S. Census Bureau.

40. Voshell Zayatz, Laura. 1991. Estimation of the percent of unique population elements on a microdata file using the sample. U.S. Census Bureau, Statistical Research Division Report Series, SRD Research Report Number: Census/SRD/RR-91/08.

41. Zayatz, Laura Voshell. 1991. Estimation of the Percent of Unique Population Elements on a Microdata File Using the Sample. Research Report Series (CENSUS/SRD/RR-91/08). Washington, DC: Statistical Research Division, U.S. Census Bureau.

42. Bethlehem, J. G., W. J. Keller, and J. Pannekoek. 1990. Disclosure Control of Microdata. Journal of the American Statistical Association 85(409): 38-45.

43. Willenborg, Leon and Ton de Waal. 2001. Elements of Statistical Disclosure Control. New York: Springer-Verlag, Inc.

44. Duncan, George T., Sallie A. Keller-McNulty, and S. Lynne Stokes. 2004. Database Security and Confidentiality: Examining Disclosure Risk vs. Data Utility through the R-U Confidentiality Map. National Institute of Statistical Sciences Technical Report 142, March, 2004.

45. Kennickell, Arthur and Julia Lane. 2006. Measuring the Impact of Data Protection Techniques on Data Utility: Evidence from the Survey of Consumer Finances. In J. Domingo-Ferrer and L. Fanconi, Eds. Privacy in Statistical Databases 291-303.

46. Gomatam, S., A.F. Karr, J.P. Reiter, and A.P. Sanil. 2005. Data Dissemination and Disclosure Limitation in a World without Microdata: A Risk-Utility Framework for Remote Access Analysis Servers. Statistical Science 20(2): 163-177.

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

48. Franconi, Luisa and Julian Stander. 2002. A model-based method for disclosure limitation of business microdata. The Statistician 51 (1): 51-61.

49. Kyle, Susan, Douglas A. Samuelson, Fritz Scheuren, and Nicole Vicinanze. 2007. Explaining discrepancies between official votes and exit polls in the 2004 presidential election. Chance 20: 36-47.

50. Witkowski, K. M. 2008a. Disclosure risk of geography attributes: The role of spatial scale, identified geography, and measurement detail in public-use files. Working paper: http://hdl.handle.net/2027.42/58626

51. Witkowski, K. M. 2008b. Disclosure risk components of contextualized microdata: Identifying unique geographic units and the implications for pinpointing survey respondents. Working paper: http://hdl.handle.net/2027.42/58627

52. Witkowski, K. M. 2008c. Finding a needle in a haystack: The theoretical and empirical foundations of assessing disclosure risk for contextualized microdata. Working paper: http://hdl.handle.net/2027.42/58628

53. Kyle, Susan, Douglas A. Samuelson, Fritz Scheuren, and Nicole Vicinanze. 2007. Explaining discrepancies between official votes and exit polls in the 2004 presidential election. Chance 20: 36-47.

54. Groves, R. M. and S. G. Heeringa. 2006. Responsive design for household surveys: Tools for actively controlling survey errors and costs. Journal of the Royal Statistical Society A 169(3): 439-457.

55. Raghunathan, T.E., James Lepkowski, Peter W. Solenberger, and John Van Hoewyk. 2009. IVEware: Imputation and Variance Estimation Software. Institute for Social Research, University of Michigan. Ann Arbor, MI. http://www.isr.umich.edu/src/smp/ive/

56. Oak Ridge National Laboratory. LandScanTM Global Population Database. 2008 [cited; Available from: http://www.ornl.gov/landscan/.

57. Bhaduri, B. LandScan USA: High-Resolution Population Distribution Model. 2005 [cited 2006 March 2006].

58. U.S. Census Bureau. Geography Division. 2009. TIGER, TIGER/Line and TIGER-Related Products. Suitland, MD. http://www.census.gov/geo/www/tiger/

59. National Aeronautics and Space Administration. 2009. Earth Observing System Data and Information System (EOSDIS). Washington, DC. http://nasascience.nasa.gov/earth-science/earth-science-data

60. United States Environmental Protection Agency. 2009. Toxic Release Inventory (TRI). Washington, DC. http://www.epa.gov/TRI/

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

Dept. of Commerce, Bureau of the Census, and Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producers], 2004. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2004.

62. Cressie, N., Statistics for Spatial Data. 1993, New York: Wiley.

63. Duncan, Greg J. and Stephen Raudenbush. 1999."Assessing the Effects of Context in Studies ofChild and Youth Development." EducationalPsychologist 34:29-41.

64. Duncan, Greg J., James P. Connell, and Pamela K. Klebanov. 1997. "Conceptual and Methodological Issues in Estimating Causal Effects of Neighborhoods and Family Conditions on Individual Development." Pp. 219-50 in Neighborhood Poverty, vol. 1, Context and Consequences for Children, edited by J. Brooks- Gunn, G. J. Duncan, and J. L. Aber. New York: Russell Sage Foundation.

65. Openshaw, S. 1984. The modifiable areal unit problem. Concepts and Techniques in Modern Geography 38: 41.

66. Ott, D., N. Kumar, and P. Thomas, Passive sampling to capture spatial variability in $PM_{10–2.5}$. Atmospheric Environment, 2008. 42: p. 746–756.

67. Bell, N., N. Schuurman, and M.V. Hayes, Using GIS-based methods of multicriteria analysis to construct socio-economic deprivation indices. Int J Health Geogr, 2007. 6: p. 17.

68. Messer, L.C., B.A. Laraia, J.S. Kaufman, J. Eyster, C. Holzman, J. Culhane, I. Elo, J.G. Burke, and P. O'Campo, The development of a standardized neighborhood deprivation index. J Urban Health, 2006. 83(6): p. 1041-62.

69. Rezaeian, M., G. Dunn, S. St Leger, and L. Appleby, Ecological association between suicide rates and indices of deprivation in the north west region of England: the importance of the size of the administrative unit. J Epidemiol Community Health, 2006. 60(11): p. 956-61.

70. Kumar, N., Spatial Sampling for a Demography and Health Survey. Population Research and Policy Review, 2007. 26(5-6): p. 581-99.

71. Kumar, N., An Optimal Spatial Sampling Design for Intra-Urban Population Exposure Assessment. Atmospheric Environment. 2009. 43(5): p. 1153-1155.

72. U.S. Census Bureau. 2009. American Community Survey. Suitland, MD. http://www.census.gov/acs/www/

73. Groves, R. M. and M. P. Couper. 1998. Nonresponse in Household Interview Surveys. Hoboken, NJ: Wiley.

74. Groves, R.M. F.J. Fowler, M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau. 2004. Survey Methodology. Hoboken, NJ: Wiley.

75. Herzog, Thomas N., Fritz J. Scheuren, and William E. Winkler. 2007. Data Quality and Record Linkage Techniques. New York: Springer.

76. 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.

77. Winkler, William. 2006. Overview of record linkage and current research directions. U.S. Bureau of the Census, Statistical Research Division Report. http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf

78. Interagency Confidentiality and Data Access Group, Statistical Policy Office, Office of Information and Regulatory Affairs. 1999. Checklist on disclosure potential of proposed data releases. Washington, DC: Office of Management and Budget.

79. United States General Accounting Office. 2001. Record linkage and privacy: Issues in creating new federal and statistical information, GAO-01-126SP. Washington DC: United States General Accounting Office.

80. Subcommittee on Disclosure Limitation Methodology, Confidentiality and Data Access Committee, Federal Committee on Statistical Methodology. 2005. Statistical policy working paper 22: Report on statistical disclosure limitation methodology, GAO-010126SP. Washington, DC: Office of Management and Budget.

81. 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.

82. 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.

83. DeWaal, A.G., and L.C.R.J. Willenborg. 1996. A view of statistical disclosure control for microdata. Survey Methodology 22: 95-103.

84. Rubin, D. B. 1987. Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

85. Fienberg, S. E., U.E. Makov, and R.J. Steele. 1998. Disclosure limitation using perturbation and related methods for categorical data. Journal of Official Statistics 14(4): 485-502.

86. Raghunathan, T.E., J.P. Reiter, and D.R. Rubin. 2003. Multiple imputation for statistical disclosure limitation. Journal of Official Statistics 19: 1-16.

87. Karr, A.F., C.N. Kohnen, A. Oganian, J.P. Reiter, and A.P. Sanil. 2006. "A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality." National Institute of Statistical Sciences Technical Report 153, June 2006.

88. Domingo-Ferrer, Josep, M. Mateo-Sanz, and Vicenç Torra. 2001. "Comparing SDC Methods for Microdata on the Basis of Information Loss and Disclosure Risk." Proceedings of NTTS and ETK.

89. Domingo-Ferrer, Josep, and Vicenc Torra. 2001. Disclosure control methods and information loss for microdata. In Confidentiality, disclosure, and data access: Theory and practical application for statistical agencies, edited by Pat Doyle, Julia I. Lane, J.M. Theeuwes, and Laura V. Zayatz, 91-110. North-Holland: Amsterdam.

90. Gomatam, Shanti and Alan F. Karr. 2003. “Distortion Measures for Categorical Data Swapping.” National Institute of Statistical Sciences Technical Report Number 131, January 2003.

91. Lane, Julia. 2005. Optimizing the Use of Micro-Data: An Overview of the Issues (August 2005). Available at SSRN: http://ssrn.com/abstract=807624. Accessed February 12, 2009.

92. Hurkens, C.A.J. and S.R. Tiourine. 1998. “Models and Methods for the Microdata Protection Problem.” Journal of Official Statistics 14(4): 437-447.

93. Duncan, George T., Sallie A. Keller-McNulty, and S. Lynne Stokes. 2001. Disclosure risk vs. data utility: The R-U confidentiality map. Technical Report Number 121. National Institute of Statistical Sciences. Research Triangle Park: NC.

SUGGESTIONS

- Publication Activities
  - Have paper reviewed by Michael Elliott or Beth Ellen Pennell, then Myron Guttman (?), and finally Julia Lane
  - Submit to the Journal of Official Statistics (publisher of Lane 2007), Transactions on Data Privacy (publisher of Lane, Heus, and Mulcahy 2008)