The Economics of Data Confidentiality

by

John M. Abowd and Julia I. Lane[1]
Cornell University and The Urban Institute
(this version: October 16, 2003)

1. Introduction

The mission of national statistical institutes is to collect and disseminate data. Decades ago, this meant producing books and reports primarily consisting of tabular data – designed to answer pre-defined questions. The increasing complexity of 21st century society, however, has put increasing pressure on such institutes to produce micro-data – designed to allow policy analysts and researchers to pose and answer questions of their own choosing. This pressure creates both opportunity and challenge. On the one hand, the relevance and stature of statistical agencies can be enhanced by their dissemination of data that policy makers can use to answer complex questions quickly. On the other hand, the well-known confidentiality challenges to the creation of public use files and other access modalities have been exacerbated by the development of new types of micro-data, as well as substantial computing and technological advances.

The creation of public use products by the agencies that collect the data reflects the public good nature of this information. The costs associated with the re-use of the micro data to address questions that were not the original focus of the collection effort are directly related to the agency's pledge of confidentiality to the respondents. Absent these costs the optimal dissemination strategy would be to release the complete micro data. The decision to gather the data as collective public activity acknowledges that the private (for-profit) production of the data, with its attendant proprietary use, would result in socially inefficient under-use of the data. Once the data have been collected, however, their re-use via public use products, licensing arrangements, secure access, and other modalities is a standard investment/production problem. Each access modality provides benefits (satisfies certain demands) and each requires additional resources from the data provider. Safeguarding the confidentiality of the micro data, then, requires an investment by the data provider in protection methods for each type of dissemination. In general, the use of a portfolio of protection methods, each coupled with an appropriate dissemination technology, provides more social benefit from the underlying data for any given level of confidentiality protection than could be accomplished by relying on a single protection method. We make this argument formally and illustrate some of its consequences in this paper.

Finding creative ways to address the fundamental tension between data dissemination and the protection of respondent confidentiality goes to the core of each statistical institute's mission. Failure to do so has tremendous costs to society. We first illustrate with an example from one of the authors' research. Lane worked with the World Bank on and off for over a decade, in a number of less developed countries. One common characteristic of the statistical institutes of the

---

[1] This paper is based upon a speech delivered by Lane at the Conference of European Statisticians Geneva, Switzerland, June 12, 2003 and a presentation made by Abowd and Lane to the NSF Workshop on Data Confidentiality, May 11-12, 2003.

countries in which she worked was a reluctance to provide access to micro-data – and in every case, this led to incomplete analysis and wasted resources in countries that could afford them least. In one case, the country in question was concerned about the low labor force participation rate of women, which had hampered development for over a decade. Several policy options were on the table – including providing free child care, flexible work-weeks, and subsidized education. However, no micro-data analysis had been undertaken. Although labor force surveys were regularly fielded, they were not even released to the Ministry of Human Resources or the Ministry of Education. Analysis of the micro-data revealed that, even after controlling for education, industry and occupation, women were paid 60% less than men – and had been for the ten years in question. The conclusion, which would have been apparent to any analyst working with these data, was that the country in question would have been better served by investigating the sources of these earnings differentials, rather than investing in the expensive set of options initially identified. Had the country in question permitted broader access to the micro-data a decade earlier, the appropriate policies could have been in place much earlier.

Eurostat has recently issued a new regulation (831/2002) to codify access to confidential data[2]. How can statistical agencies determine the "optimal" amount of micro-data to release – and find creative ways to seek this optimum? Our answer is that an accurate assessment depends on the benefits derived from the use of such data, the costs, and the tradeoff between the two. This paper is to attempt to explicitly delineate these benefits and costs, identify changes and summarize the consequences and opportunities for statistical agencies.

2. The benefits of micro-data use and why they are increasing.

The benefits associated with micro-data access are myriad. The most obvious is that micro-data permit policy-makers to pose and answer complex questions, but others are also apparent. Access to micro-data permits analysts to calculate marginal, rather than average effects. Access also acts as an important scientific safeguard because it permits others to replicate findings. Discovery and replication create a virtuous cycle of knowledge for the statistical institute because data use inevitably reveals data quality and processing anomalies as well as new data needs. Finally, re-use of micro-data creates a core constituency for the statistical agency itself.

   a) Micro-data permit analysis of complex questions

One of the most important findings in economics over the past decade has been that the analysis of aggregate statistics does not give policy makers an accurate view of the functioning of the economy. Indeed, the creative turbulence which is the hallmark of the United States economy, and a major contributor to its success, is not apparent from macro level indicators. Analysis of micro-data suggests that the widespread reallocation of factors of production from one firm to

---

[2] See Jean-Louis Mercy and John King's paper "Developments At Eurostat For Research Access to Confidential Data" Joint ECE/Eurostat work session on statistical data confidentiality (Luxembourg, 7-9 April 2003) Working Paper 12.

another firm even within narrowly defined industries is a major contributor to U.S. productivity growth – more important than investment in equipment and structures[3].

We illustrate this phenomenon using data provided to policy makers in Illinois that examine employment changes in a detailed industry – industrial machinery – in a detailed geographic area - Peoria, IL. Aggregate statistics indicated that this industry had lost a total of 20 jobs in the previous year. An analysis of the micro-data, summarized in Figure 1, revealed a very different picture. The net employment loss of 20 jobs was the sum of positive employment gains for workers 44 and under, and employment losses for workers 45 and older. About 160 jobs were reallocated from older to younger workers. The micro-data revealed even more reallocation than this. If we simply tabulate up the job gains from expanding and new firms, there were over 250 jobs gained for workers of all ages (including older workers). The gross job reallocation, achieved by summing up 250 jobs gained and 270 jobs lost, exceeds 520 jobs. The worker flows are greater yet. Over the same period, over 710 workers were hired and 730 separated – for a total of 1400 workers reallocated.

Workforce Dynamics: Industrial Machinery, Peoria, IL
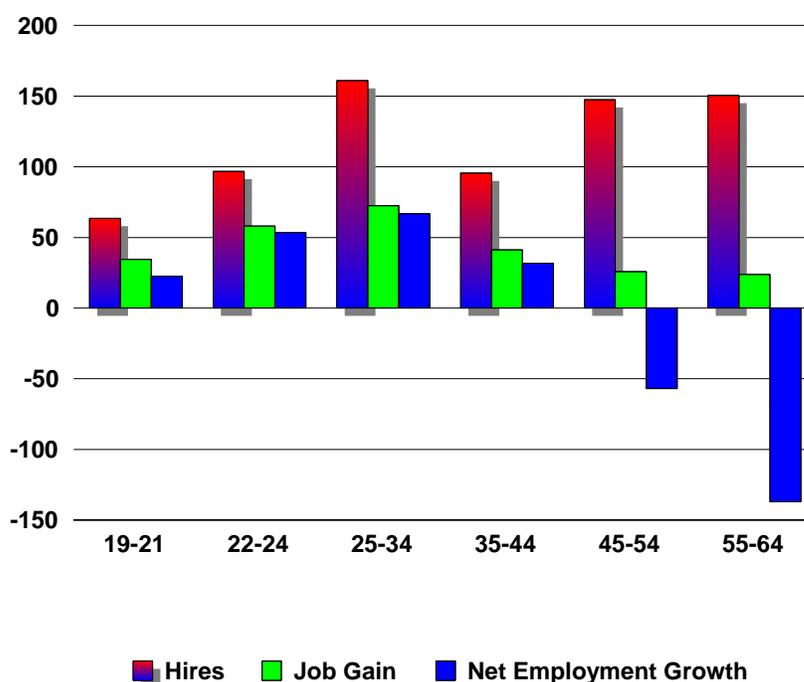
Source: LEHD Program, US Census Bureau and Illinois Department of Employment Security

Figure 1

The importance of knowing that even quite small net job changes can represent enormous job and worker reallocation is non-trivial information for policy-makers so that the productive potential of this reallocation process can be realized to its fullest. In this case, for example, the

---

[3] Foster, Lucia, John Haltiwanger, and C.J. Krizan (2001). "Aggregate Productivity Growth: Lessons from Microeconomic Evidence." *New Directions in Productivity Analysis*, (eds. Edward Dean, Michael Harper, and Charles Hulten), University of Chicago Press, (forthcoming).

analysis showed Illinois policy makers that the aging of the industrial machinery workforce would lead to a demand for trained workers to replace oncoming retirements.

The new challenge that this increasing value of micro-data poses to statistical agencies is that the micro-data sets that permit such in-depth understandings of the economy – which involve the longitudinal linkage of firm and worker data over time – are also very large and complex, and often involve the integration of administrative and survey records. External researcher access is often the only way to create such data because many of the decisions require subject matter knowledge as well as statistical expertise.

b) Calculating marginal rather than average effects

The ability to estimate marginal effects goes to the heart of the use of micro-data. Micro-data enable analysts to do multivariate analyses, whereby the marginal impact of key variables, controlling for other factors, can be isolated. An excellent example is provided by a recent study[4] which investigates the distributional impact of Medicare. The importance of this healthcare program for the elderly population is difficult to overstate – it cost $220 billion (in 1998) and its costs are growing faster than Social Security. Understanding program use, and the correlation of this with income and health, is critical to understanding the effects of the program.

The micro-data reveal important facts about program use that, again, would not be available from an analysis of aggregate data. Program use is heavily skewed – a very small proportion of the elderly population account for a very large proportion of expenditures. Program use is very persistent: those who account for a high proportion of expenditures in one year are highly likely to be heavy users in subsequent and preceding years. Even more interesting, however, is the effect of examining the relationship between income and expenditures, which is described in Figure 2.

---

[4] Lee, McClellan and Skinner "The Distributional Effects of Medicare", NBER working paper 6910, January 1999.
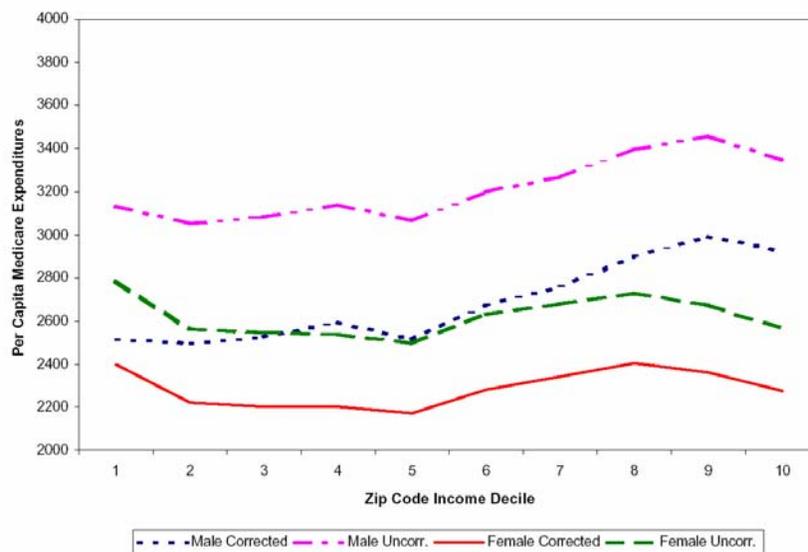
Source: Lee, McClellan and Skinner, 1999
Figure 2

Briefly, it is clear from an examination of Figure 2 that the marginal effect of income on expenditures is broadly positive for men, but that the relationship is not only much flatter for women but women spend less. The marginal effect of correcting for health status (whether or not the individual died during the analysis year) at all income levels is also evident. Thus this analysis of the micro-data provides a quantification of the marginal effects of the key contributing factors to health expenditures: sex, income and health.

This example controls only for demographic effects – yet the increasing complexity of economic activity requires the production of data that can be used to separate out not just complex demographic interactions, but also economic and spatial effects. The expansion of research on the human dimensions of environmental change has increasingly meant that researchers want to include the contextual variables surrounding an individual-the schools they go to, the neighborhoods they live in, the firms they work for, and the people with whom they interact. As Rindfuss points out, "Linking data on people and their environments is at the very core of IHDP"[5] The imperative to identify marginal effects in such an environment will put tremendous pressure on statistical agencies.

c) *Scientific safeguard*

Access to micro-data is critical to ensure that other scientists can replicate important research. This acts as an important discipline device for both government statisticians and academic researchers. That there is overwhelming temptation for scientists to misrepresent results is, sadly, evident from the all-too-frequent news stories of data fabrication. That there is similar pressure on statistical institutes should be taken as self-evident. Constant vigilance in this area is important. When the gains to monopoly power over information are great, in terms of either

---

[5] Ronald Rindfuss "Confidentiality Promises And Data Availability" in IHDP Update, 02/2002, Newsletter of the International Human Dimensions Programme on Global Environmental Change.

political or professional prestige, it would be naïve to think that there were no malfeasance in even the most pristine of agencies. The consequences to the statistical system of such malfeasance can be devastating if unchecked.

d) Data quality

Although statistical institutes expend enormous resources in quality assurance to ensure that they produce the best feasible product, there is no substitute for actual research use of micro-data to identify data anomalies. Indeed, there is general recognition of the direct correlation between the quality of a national statistical institute and that institute's openness to external research in international agencies, such as the World Bank. The United States Internal Revenue Service (IRS) and the United States Census Bureau have actually formalized the role of researcher use of selected tax micro-data to improve national statistics. Because this inter-agency agreement only permits the IRS to release selected micro-data[6] to the Census Bureau in order to improve the economic and demographic censuses, surveys and inter-censal population estimates researchers who use Census Bureau's tax-derived micro-data must document the benefits according to the following criteria:

- Understanding and/or improving the quality of data produced through a Title 13, Chapter 5 survey, census or estimate;
- Leading to new or improved methodology to collect, measure, or tabulate a Title 13, Chapter 5 survey, census or estimate;
- Enhancing the data collected in a Title 13, Chapter 5 survey or census. For example:
    Improving imputations for non-response;
    Developing links across time or entities for data gathered in censuses and surveys authorized by Title 13, Chapter 5.
- Identifying the limitations of, or improving, the underlying business register, household Master Address File, and industrial and geographical classification schemes used to collect the data;
- Identifying shortcomings of current data collection programs and/or documenting new data collection needs;
- Constructing, verifying, or improving the sampling frame for a census or survey authorized under Title 13, Chapter 5;
- Preparing estimates of population and characteristics of population as authorized under Title 13, Chapter 5;
- Developing a methodology for estimating non-response to a census or survey authorized under Title 13, Chapter 5;
- Developing statistical weights for a survey authorized under Title 13, Chapter 5.

A sterling example of how this can work is a new project between the Census Bureau and researchers at the Sloan Industry Centers. The Sloan Foundation has invested heavily in case study research of a number of industries, five of which (semi-conductors, software, retail trade, finance and trucking) are involved in this project. The Sloan researchers work directly with

---

[6] As in many countries, selected tax data form the heart of the Census Business Register – the business sample frame – and play a critical role in developing intercensal population estimates.

Census staff – and their rich industry specific knowledge should lead to contributions ranging from help with industry classification to identifying new survey questions that could hone in on the driving forces of change in their industry.

Statistical institutes operating in an environment where the blurring of firm and industry boundaries is accelerating, where the differentiation between place of work and place of residence is increasingly unclear, and where the engine of economic growth has changed from measurable machines and equipment to the much less measurable workforce quality will increasingly need to turn to external researchers for guidance.

e) Development of core constituency

The funding of a statistical agency depends on the development of a constituency and greater use of data – which includes the creation of new products from existing data - creates a constituency beyond that of those who access the data. More analysis, more publicity and more insights lead to a greater understanding of the value associated with products produced by the statistical agency with associated funding benefits.

The value of a core constituency goes beyond the funding aspects. The quality of staff that can be hired is directly correlated with the prestige and visibility of the institute, and the perceived quality of work that can be done within its walls. External researchers, who are often academics, also advise and counsel students about career opportunities. Cultivating this network is an important first step to developing a high quality staff – maintaining the dynamic interaction between staff and their mentors can create an ongoing virtuous cycle of information exchange and education.

3. The costs of micro-data use – and how they are changing

The most obvious costs of micro-data use include the cost of providing access, potential reputational costs and the costs associated with identification of the sampled entities and the concomitant potential disclosure of confidential attributes. These are the costs that must be weighed against the benefits of providing access.

a) The cost of providing access

Clearly the cost of providing access depends on the modality, and several have been developed by statistical institutes across the world – public use micro-data, licensing, remote access sites, research data centers. The agencies' explicit costs for each of these methods are substantial in terms of staffing, support and documentation. The costs to users vary dramatically – public use data are clearly the lowest cost option, while the explicit and opportunity costs of accessing research data centers are substantial.

The most important of these modalities – and the one subject to most change - is public use micro-data. Statistical institutes have worked very hard to make these available, with dramatic success. It is not an overstatement to say that since such data were first created over 40 years ago, they have had a major impact on decision making. Indeed, decisions are often made in

developing countries based on results from European and North American public use data sets. Funding decisions for some entire data collection activities are predicated on the existence of public use micro data. However, the cost and feasibility of producing high quality public use datasets is unfortunately increasing. A combination of technological advances in computing capacity, computer linking software and increased online availability of administrative data threaten their very existence[7].

Dealing with the threats to public use files is an area in which much needs to be done – and one in which statistical agencies can join forces. One under-investigated area is the effect of the choice of different disclosure protection techniques on data quality. The lack of agency focus on this is evident: agencies that pour resources into producing top quality data - for example, survey design to improve response quality, and response follow up to reduce attrition bias – will spend much less on the decision to top-code, data-swap or suppress information. While this lack of focus was rational in a less technologically savvy era, it is unlikely that statistical institutes will continue to be able to be so sanguine.[8]

An attractive technical development that has received much attention in the last three years is the creation of inference-valid synthetic datasets[9]. These datasets often use multiple imputation and other Bayesian techniques to create data with the same analytical structure as the underlying protected data. They can be used by researchers at a remote site to develop an understanding of the structure of the confidential data, develop analysis code, and even estimate basic relationships before sending the code to the secure site to estimate the underlying relationships on the original confidential data. The quality of this approach is evident in Figure 3 – using French data, Abowd and Woodcock show that there is almost no difference between results estimated using some forms of synthetic data and the actual confidential data. Other forms of synthetic data suffer some analytic difficulties but they appear to be manageable.[10]

---

[7] See, for example, Chapter 1 in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, edited by Pat Doyle, Julia Lane, Jules Theeuwes, and Laura Zayatz North Holland, 2001.

[8] For a rigorous formal treatment of this problem see Dobra, Fienberg and Trottini "Assessing the Risk of Disclosure of Confidential Categorical Data," in *Bayesian Statistics* 7 (2003): forthcoming.

[9] See "Disclosure limitation in longitudinal linked data" Abowd and Woodcock (2001) in Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies edited by P. Doyle, J. Lane, J Theeuwes and L. Zayatz, North Holland, Amsterdam, 2001 and T. Ragunathan, J. Reiter, and D. Rubin "Multiple Imputation for Statistical Disclosure Limitation," *Journal of Official Statistics* (January 2003).

[10] See also J. Reiter "Satisfying disclosure restrictions with synthetic data sets," Journal of Official Statistics (2003).
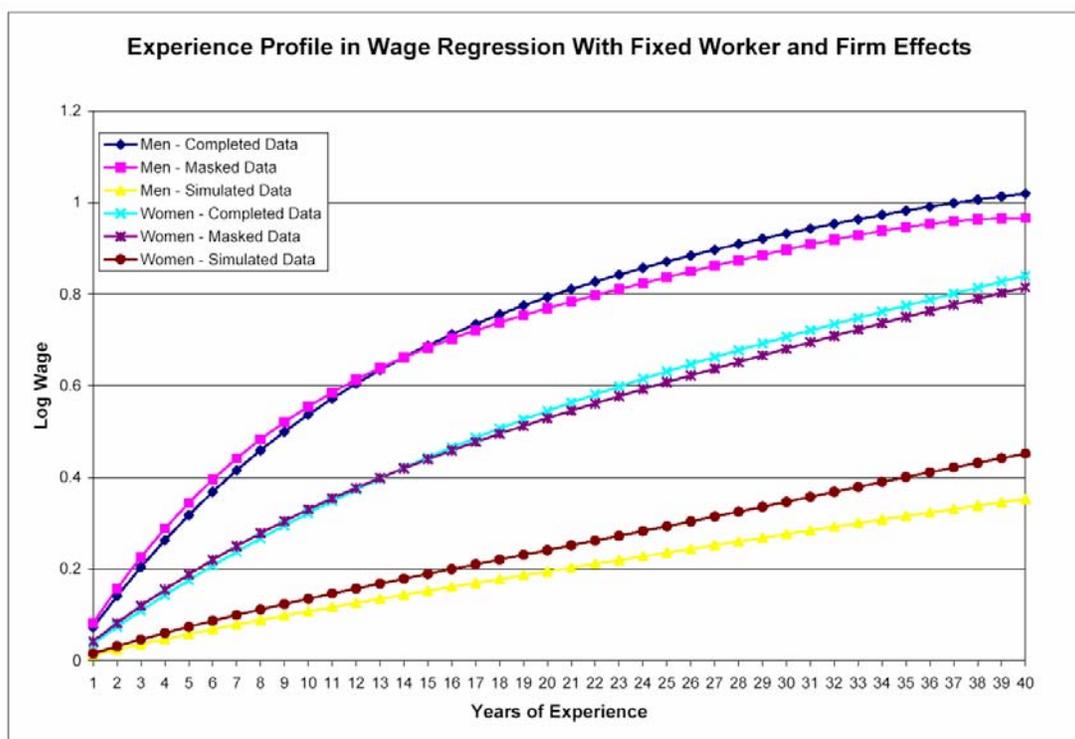
(Source: Abowd and Woodcock 2001)
Figure 3

b) Reputational costs

Another very real cost associated with outside researcher access to national statistical institutes is that of reputation. The production of official statistics is the mandated reason for their existence – and the typical agency expends enormous effort making sure that published statistics with their imprimatur are the national gold standard. As a result, each agency is understandably concerned that research results using data with their imprimatur, and without their expertise, could be misconstrued as "official" and be misused.

It is possible to manage this type of damage. The World Bank's Living Standards Measurement Survey (http://www.worldbank.org/lsms/) has extensive tutorials, software packages and "how-to" manuals to make sure that researchers working with similar datasets know what they're doing. An alternative approach was taken in the U.S. in the form of the recent "Information Quality Act" which requires the U.S. Office of Management and Budget to develop government-wide standards for data quality. Interestingly, that act distinguishes between "ordinary" and "influential" information – the latter including "influential scientific, financial or statistical information" that will "have a clear and substantial impact on important public policies or important private sector decisions" (67 FR 8452). Even more tellingly, influential information should be reproducible by qualified third parties (though exceptions apply).

c) Disclosure of respondent identities

The ultimate cost to an agency is for an external researcher to disclose the identity of a business or individual respondent. While the penalties for this are typically substantial – ranging up to 10 years in jail and a $250,000 fine in the U.S. – the consequences of such a breach could be devastating to respondent trust and response rates. As trust in the government appears to be declining, statistical agencies might well also be concerned that respondent trust in their ability to protect respondent confidentiality is declining – and that this might only be exacerbated by permitting widespread researcher access

There has been some research attempting to quantify the order of magnitude of the relationship between trust and response rates, and the trends over time in the U.S. (by Eleanor Singer for respondents to demographic surveys, and Nick Greenia for respondents to economic surveys). Indeed, a resolution was adopted at a UN/ECE confidentiality workshop in Skopje, Macedonia in 2001 to move forward with a joint European endeavor to quantify the effect of researcher access on perceptions, but I am not clear on how much progress has been made in actual implementation.

4. Putting the benefits and costs together

While our cataloguing of the costs and benefits of data confidentiality has not been encyclopedic, we believe that these capture the main economic factors of interest. Statisticians have formalized the interplay of these costs and benefits as a tradeoff between disclosure risk and data utility.[11] This is an important advance because it provides a framework in which the tradeoff between disclosure risk and data use can be quantified, thus allowing data providers to be efficient in the economic sense—getting the most data utility for a given amount of confidentiality protection.

The statistical framework provides a method for assessing whether or not a proposed protection technique provides the most disclosure limitation for a given level of data utility. Application of these methods promotes greater efficiency because, in economist's language, it allows the data provider to get closer to the production possibility frontier. Some methods in current use can be shown to be dominated—there exists a method with the same disclosure limitation that permits more data utility.[12]

Eliminating dominated disclosure limitation methods is important but it is not a complete analysis of the problem. Data providers must also apply economic decision making to the mix of dissemination methods for a given data product. Assuming that the risk-utility assessment has been properly applied to each of several dissemination methods (public use data, licensing, remote access, and research data center, for example), how does an agency decide whether to devote more resources to one method versus another. This decision is an example of optimal portfolio theory.[13] Any two data protection methods are correlated in their risk of disclosure of confidential information, but not perfectly. Combining the two methods can, then, produce greater data utility for any given level of disclosure risk in exactly the same way that an investor

---

[11] See Duncan et al. "Disclosure limitation methods and information loss for tabular data," in Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies edited by P. Doyle, J. Lane, J Theeuwes and L. Zayatz, (North Holland, Amsterdam, 2001), pp. 135-166.
[12] See Dobra *et al,. op. cit.*
[13] See H. Markowitz *Portfolio Selection: Efficient Diversification* (1959).

can achieve greater expected return for any given level of investment risk by combining the risky assets into a portfolio.

The application of optimal portfolio balance to the choice of data protection and dissemination methods provides the potential to answer two very important questions that arise directly from our cost and benefit analysis above. First, what is the overall disclosure risk of the mix of dissemination technologies? Second, what is the correct decision rule for moving information between a public use file and a restricted-access file? We consider these questions in turn.

The overall disclosure risk for a combination of dissemination methods is the expected cost from the proposed combination, not the sum of the disclosure risks from each of the methods taken individually. This point can be seen most clearly by comparing a public use micro data file with a supervised access protocol like a research data center. If there is only one variable in the public use file and there are 10 variables on the confidential then, for any given level of data utility there will be much more use of the research data center than of the public use file. So, the expected costs associated with the disclosure limitation will be dominated by the costs of running the research data center and risks associated with the accidental or malfeasant release of some confidential information from this modality. Again for any given level of data utility, adding a variable to the public use file and tightening the access to the research data center shifts expected cost of disclosure limitation from the supervised facility to the properties of the public use file, which must be controlled through investments in statistical methods to limit the identification and attribute disclosure risk in the two variables as compared to the single variable. The overall disclosure risk can actually decrease because there are lower expected costs in the supervision of the research data center.[14]

We can now consider the decision rule for moving information between the public use file and the research data center. The addition of information to public use files has a measurable impact on the disclosure risk. Sometimes this is low: for example, measures of benefit entitlement in a national program. Such measures don't depend on geography and have statutory minima and maxima. Adding such measures to a public use file and simultaneously eliminating the research data center use of the confidential data to create such variables for each study provides an increase in data utility from the public use file and a decrease in data utility from the research data center. That's the benefit tradeoff. At the same time it provides a change in the overall disclosure risk that depends upon how much extra information is contained in the benefit measure above what was contained in the original public use file and upon how much access to the research data center is reduced. That's the cost tradeoff. The new statistical methods for assessing disclosure risk and data utility can be used to quantify both the benefit and the cost side of this tradeoff. We suspect that these methods will reveal that there are improvements available from adjusting the dissemination mix in this case.

Sometimes there is considerable disclosure risk associated with a variable in a public use file. Exact birth dates and detailed geography are examples for household data. Exact industry and detailed geography are examples for business data. Improvements in information technology have increased the disclosure risk associated with the public use versions of these variables. This has resulted in increased use of restricted-access protocols like research data centers. As in the

---

[14] Michael Hurd originally suggested this argument to us. We have attempted to formalize it.

national benefit example above, disclosure risk assessments can be used to quantify the reduction in risk in the public use file from restricting the geography and the increase in disclosure risk from making the geography available in a research data center (holding constant the data utility). These same methods can be used to measure the change in data utility from this restriction in the public use file and associated increase in the research data center use. We suspect that there are potential portfolio gains to this rebalancing also.

5. Conclusion

It is clear that statistical agencies will increasingly be challenged to provide more access to micro-data. This pressure provides a chance to fulfill a critical societal mission. However, since increased access does not come without increased costs, it would seem reasonable to try to control these costs by combining research efforts. Some areas in which joint research and development might provide substantial dividends, for example, would be:

1. the creation of inference-valid synthetic datasets
2. the protection of micro-data that are integrated across several dimensions (such as workers/firms/geography)
3. the quantification of the risk/quality tradeoff in confidentiality protection approaches
4. the effect on response rates of increased micro-data access.