Title: Methods for Protecting the Confidentiality of Firm-Level Data: Issues and Solutions*

Lars Vilhuber

This version: 2013-03-28 v3b

1. Introduction

This report will provide an overview of methods used by statistical agencies to encourage, support, and enhance research access to data for the purpose of generating new knowledge. Quite a few reports and scientific articles have addressed the issue before, and we will be highly indebted to that literature. To a summary of that literature, we hope to provide some recent developments and experiences derived from a decade of working with systems that increase access as both researchers as well as data providers. The report will focus on the data provided by statistical agencies, but it should be understood that government agencies other than a National Statistical Office (NSO) may acquire that function. While excluding the legal background limiting or permitting such data collection and provision, we will highlight some alternate sources and methods, prior to concluding.

By its very nature, releasing information increases the amount of data on persons, households, and firms available to society, sometimes to the point of infringing on the privacy expectations of those entities. Statistical agencies and their partners aim to balance the risk of that happening with the benefit of knowledge gained from the increased information, for the larger benefit of society. Abowd (2013) provides a overview over the economics of such an arbitrage. This report, and most of the literature, provide little guidance to what the optimal balance may be, and we will remain silent on the final decision, which remains a political decision, rather than a technical or statistical decision.

The report will in general describe protection and access mechanisms in general terms, without specific reference to business or establishment data (henceforth referred to as "business data" for simplicity), since most such mechanisms apply to business, establishment, individual, and household data symmetrically. We will point out the issues that apply to the access to business data specifically at appropriate points in the discussion.

2. Known methods

There is a broad consensus on the scope of making data from National Statistical Offices (NSOs) available to a "research community" however defined (Nations; Weinberg et al. 2007). First, and generally foremost, NSOs provide ***public-use data***, typically tabulations of the underlying microdata by some level of geography or industry. The core principle is that statistical disclosure limitation (SDL) of some sort has been applied to the underlying microdata to protect the confidentiality of respondents. Such SDL may include the following:

* This document was originally prepared for Industry Canada in 2013 (Contract number 5025993), and subsequently also published with permission as part of the "Network to Study Productivity in Canada from a Firm-Level Perspective" as NSPCFP Working Paper 2013-03. This version: 2013-03-28v3.

2.1. Publication of tables and data cubes (multi-dimensional tables).

In this scenario, the underlying data are aggregated across several characteristics. The respondent data are protected by a variety of mechanisms, such as

- suppression of data cells that do not meet certain criteria, including complementary data cell suppression
- swapping of data cells prior to tabulation. Examples include the US Census Bureau's Decennial Census and American Community Survey
- noise-infusion (Evans et al., 1998), which adds statistical noise to the underlying microdata, protecting small cells while maintaining adequate precision in large cells. Examples include the US Census Bureau's QWI (Abowd et al. 2009) and CBP (http://www.census.gov/econ/cbp/methodology.htm)
- use of synthetic data to supplement or replace sensitive data at the pre-tabulation stage, in order to protect the confidentiality of certain populations (Hawala 2008; Abowd et al. 2012)

2.2. Public-use microdata

An alternative to aggregation is the provision of **public-use microdata**, in which the underlying data has been protected through a variety of mechanisms:

- sampling either during data collection (surveys) or post-collection (censuses), or both (ACS)
- swapping of data cells (again)
- variable suppression
- top- and bottom-coding
- re-categorization and coarsening, in particular of geographic and industry identifiers
- provision of partially (Kennickell 1998) or fully synthetic, analytically valid microdata (Kinney et al. 2011; Abowd, Stinson, and Benedetto 2006). We will get back to this method in more detail later.

2.3. Access to confidential data

Improvements in statistical procedures have increased the level of detail available in public-use tabulations, replacing the suppression of sensitive cells with detailed but protective statistics in small cells. For instance, the QWI tabulates cells at the level of “county by detailed NAICS industry (NAICS4) by two demographic worker characteristics”, describing the employment stocks and flows in detail in several thousand cells for each and every of the over 3000 counties in the United States, even though a significant number of cells are still suppressed. (Abowd et al. 2009)

However, tabulations in general may not provide the level of detail required for detailed analysis, in particular of marginal effects. Public-use microdata may lack certain elements that cannot be feasible protected, such as longitudinal earnings and work histories (Abowd and Woodcock 2001). Even the provision of very detailed microtabulations or public-use microdata may not be sufficient to inform certain types of research questions. In particular for business data, the thresholds that trigger SDL methods are met far more often than for individuals or households. In those cases, the research community needs access to (confidential) microdata. United Nations Economic Commission for Europe (2007, pg. 4) cite three key reasons why access to microdata may be beneficial:

(i) “microdata permits policy makers to pose and analyse complex questions. In economics, for example, analysis of aggregate statistics does not give a sufficiently accurate view of the functioning of the economy to allow analysis of the components of productivity growth;

(ii) access to microdata permits analysts to calculate marginal rather than just average effects. For example, microdata enable analysts to do multivariate regressions whereby the marginal impact of specific variables can be isolated;
(iii) broadly speaking, widely available access to microdata enables replication of important research”

Several methods are currently used by NSOs and other data collecting agencies to provide access to confidential data. The following sections will describe each of them in turn. I draw on Weinberg et al. (2007) for much of this section.

2.3.1. Licensing

Licensing is often used in order to provide access to restricted-use microdata. In general, the detail in these files is greater than in an equivalent (or related) public-use file, and may allow for disclosure of confidential data if exploited, but such files tend to have several levels of disclosure avoidance methods applied to them as well. For instance, the NLSY provides both public-use files and licensed files. The licensed files have more detailed geography on respondents (county, rather than Census region), but do not have the most detailed geography (GPS coordinates or exact address). Generally, the legally enforceable license imposes restrictions on what can be published by the researchers, and restricts who can access the data, and for what purpose. In the United States, some surveys (NCES, NLSY, HRS) use licensing to distribute portions of the data they collect on their respondents, and commercial databases (COMPUSTAT, etc.) are also distributed in that fashion. Penalties for infractions range from cutting off future funding (example: HRS) to monetary penalties. Sometimes, the licensing arrangement also requires that the data recipient provide a secure data enclave - often a stand-alone computer, but data enclaves such as CRADC or NORC may be acceptable.

2.3.2. Statistical data enclaves

Statistical data enclaves, or Research Data Centers, are secure computing facilities that are used to provide full or nearly full access to confidential microdata to researchers, while putting restrictions on what content can be removed from the facility. In contrast to licensing arrangements, which are self-monitoring as to statistical output, statistical data enclaves tend to have more forceful output monitoring, typically by staff of the data provider. Access is in general for approved users only, a sometimes lengthy approval process is standard.

Statistical data enclaves can be central locations, in which a single location at the statistical agency is made available to approved researchers (in the US, NCHS and BLS follow this model; in Canada, some datasets at Statistics Canada are accessed through this mechanism), or may be distributed geographically (demographic data in Canadian RDCs are distributed; each facility has a discrete copy of the data).

Some facilities are hybrid facilities, where the statistical processing occurs at a central location, but secure remote access facilities are distributed geographically. The U.S. Census Bureau's RDCs work this way since the early 2000s. A central computing facility is located on Census Bureau premises at its headquarters, and secure remote access is obtained by researchers at designated sites throughout the county. Each of the designated sites itself is actually a Census Bureau facility, located on university premises. In France, the Centre d'accès sécurisé distant aux données (CASD) has a secure central computing facility, and allows for remote access through secure devices from designated university offices (which satisfy certain physical requirements). Both the Census Bureau and CASD also host data from other data providers, through collaborative agreements.

2.3.3. Remote access

We mentioned remote access through secure facilities in the previous section. However, two other alternative remote access mechanisms are often used: manual and automatic remote processing. We refer to manual remote processing when the remote "facility" is a staff member of the data provider. This can be as simple as sending programs in by email, or finding a co-author who is an employee of the data provider. The United States' NCHS, Germany's IAB, and Statistics Canada provide this mechanism. Generally, because of the high labor intensity of this mechanism, costs are incurred for users prevailing themselves of this service.

More sophisticated mechanisms automate portions or all of the data flow. For instance, programs may be executed automatically based on email or web submission, but disclosure review is performed manually. Fully automated mechanisms, such as LISSY (Luxembourg), ANDRE(NCHS), DAS (NCES), Australia's RADL and the Census Bureau's Advanced Query (AQ) and Microdata Analysis System (MAS) systems, generally limit what the users can do to certain statistical procedures and languages for which known automated disclosure limitation procedures have been implemented.

Of the known systems surveyed above, only Australia's RADL systems and the Bank of Italy's implementation of LISSY (Bruno et al, 2008) seem to provide access to business microdata through automated remote processing facilities.

2.3.4. Synthetic data

We mentioned earlier, in the section on public-use microdata, the feasibility of partially or fully synthetic microdata. While the Survey of Consumer Finance (Kennickell 1998)has most notably used this method to provide access to respondent attributes that would otherwise be too sensitive to provide to users of its *public-use* survey data, other synthetic datasets have not yet been made available as freely accessible public-use dataset, although that remains the ultimate goal. The use of synthetic data was first proposed in 1993 (Rubin 1993; Little 1993) as a way to provide tabulations or microdata when the underlying data could not be adequately protected. In fact, the 1990 Census implemented what it called "blank and impute" method in the preparation of its public-use tabulations (Weinberg et al. 2007). The first fully-synthetic microdata file, the SIPP Synthetic Beta (SSB) was released in 2007 (Abowd, Stinson, and Benedetto 2006), which implemented the original idea of reproducing most of the statistical properties of the underlying confidential data while replacing all values with simulated values. The subsequent release of the Synthetic Longitudinal Business Database (SynLBD) in 2011 (Kinney et al. 2011) is, to our knowledge, the first release of analytically valid synthetic business microdata. Both of these datasets, while satisfying the disclosure-protection criteria of the U.S. Census Bureau, have not yet been released as freely accessible public-use microdata files, and implement a form of licensing. Currently, the Synthetic Data Server project at Cornell (http://www.vrdc.cornell.edu/sds/), for which I am the Principal Investigator, houses these two synthetic data sets, providing controlled but full remote access to these restricted-access datasets. Researchers have full access to statistical software, but cannot remove data from the system. No disclosure avoidance review is performed on the results, though, and personnel from the Census Bureau releases analysis results to the researchers with a very quick turnaround. Users can choose to validate their results against the confidential data, and are strongly encouraged to do so. The results returned from running the researchers' models against the confidential data are, of course, subject to disclosure avoidance review, but because they are model-based, generally do not create a large disclosure risk. The overall workflow blends elements of licensed data use, remote access to data enclaves, and remote processing. We note that the two datasets available (SSB and SynLBD) differ substantially in both their

characteristics and the complexity of their generating process. The SSB starts with a longitudinally harmonized version of multiple SIPP panels, and then synthesizes 93 out of 95 variables in a complex SRMI model. The SynLBD starts with the longitudinally combined data from the Business Register called the Longitudinal Business Database (LBD) (Miranda and Jarmin 2002), and synthesizes 9 out of 28 variables (and drops the other 19 variables). However, in part due to the simpler process required to generate the SynLBD, we have started a project to create a German version of the SynLBD, using the same methodology, and have explored the idea for Canadian data. Since the creation of the Synthetic Data Server in 2010, xx projects have used its services (several started prior to that date on a predecessor server that housed only an earlier version of the SSB). A newer version of the SSB is scheduled to be made available in April 2013, and updated (2.1) and enhanced (3.0) versions of the SynLBD are in active development.

Additional uses of synthetic data can also lead to more detailed public-use tabulations. The Census Bureau released a new public-use data product with unprecedented detail on the relation between workplaces and residences, called OnTheMap (http://onthemap.ces.census.gov), based on synthetic data on the distribution of residences. The data (and an associated application) provides details on worker and firm characteristics for tract and block-level, with no suppressions and a high degree of analytic validity (report here). The data and application were the result of a multi-year effort involving stakeholders from multiple government agencies, and the development of novel protection mechanisms (Machanavajjhala et al. 2008). Synthetic data were also used in the release of ACS tabulations and public-use microdata, protecting the data on a particular sub-population (residents of group quarters), due to the sparsity of such locations.

2.3.5. Temporary or de facto employment of researchers

If the agency has data it wishes to be analyzed, it may be able to contract with researchers in such a way that they are legally considered employees, with access to the data, and possibly to analytical systems in-house at the agency. This method is often combined with one of the other methods. Researchers wishing to access the Canadian Research Data Centers sign contracts that result in them being treated as “deemed employees” without pay. The Internal Revenue Service sometimes uses similar methods. Researchers who use the U.S. Census Bureau sign agreements that commit them to legal obligations equivalent to that of employees, without being considered de-facto employees (“special sworn status”). Depending on the legal framework, this can provide researchers with both the obligations and access privileges of regular paid employees.

3. Potential users of data

Important distinction about the ultimate user group: general public, general-purpose publications (newspaper, general circulation magazines), policy analysts with high-level goals, scientific uses with detailed models.

4. Examples from select institutions

4.1. U.S. Census Bureau

The U.S. Census Bureau uses a wide variety of methods to enhance research access to data. Internships, post-doctorate fellowships, temporary employment of academics (both in management and research positions), and extensive research contracts are all in continuous use. The RDC network, of which the Census Bureau supports a large fraction of the cost, is spreading (as of March 2013, 15 locations), and

a new research network called the National Science Foundation (NSF)-Census Research Network (NCRN, see www.ncrn.info), allow for varying levels of access. New methods are developed and tested both in-house (Microdata Analysis System (MAS), Advanced Query System) and using outside researchers (QWI, OnTheMap, Synthetic LBD). Regular reports (U.S. Census Bureau 2012) highlight the advances made by researchers using the different access mechanisms. Researchers who access the data through the RDC network or through contracts are subject to the same legal sanctions that regular Census Bureau employees are ($250,000 or 5 years of prison for illegal disclosures), regardless of how and if they are remunerated. Annual training is compulsory for all, and the training for external researchers is the same as for Census Bureau employees.

4.2. Bureau of Labor Statistics

The Bureau of Labor Statistics only runs a Research Data Center at its headquarters in DC. Administrative data it collects can only be accessed on-site. However, the BLS outsources much of its surveys to subcontractors (including the Census Bureau), and some of those use licensing to allow for off-site access to restricted-use data. Most surveys are made available as public-use microdata (NLSY, CPS, etc.). Business-level data from administrative sources (QCEW) and surveys (JOLTS, etc.) are made available only in the form of aggregated time-series, or at the HQ RDC. Resources at the HQ RDC are limited, and application deadlines infrequent (4 times a year). The BLS publishes regular research briefs, typically by in-house researchers, but also sponsors post-doctoral fellowships specifically to encourage new research.

4.3. National Center for Health Statistics

Although NCHS does not have firm-level data, it is worthwhile to point out the mechanisms it uses, as it covers a wide-variety of research access methods discussed here. Detailed public-use microdata are available for many datasets, but custom extracts of the underlying confidential data can be made available (a) through person-based (“staff assisted”) remote processing (b) through email-based remote processing (ANDRE) (c) through a onsite Research Data Center in Atlanta, GA (d) through an agreement with the Census Bureau's RDC network in all 15 locations of that network. NCHS charges researchers per-day fees to access the different systems.

4.4. IRS

The Statistics of Income (SOI) division of IRS has more recenlty expanded access through its Collaborative Research Program . A call for proposals is occasionally made public, identifying projects that IRS-SOI is willing to entertain. Researchers always collaborate with SOI staff (no self-guided research). Methods of access range from custom tabulations to temporary non-remunerated staff positions. In total, 20 projects were accepted in 2011 out of 51 submissions. Out of those, more than half involved the researchers becoming temporary employees or engaging in unpaid research contracts. The IRS does not have a designated RDC.

4.5. Bank of Italy

The Bank of Italy's provision of access to business microdata through LISSY, the Luxembourg Income Study's remote processing system, is worth highlighting, since it is, to our best knowledge, the only use of such a system for the purpose of granting access to confidential business microdata (Bruno, D’Aurizio, and Tartaglia-Polcini 2008). The authors note that “Firms’ privacy is safeguarded by forbidding potentially confidentiality-breaking programme statements and by denying the visualisation

of individual data. Data confidentiality is protected by removing key identifiers from the database and by trimming data in the right tail of the distribution." (Bruno, D'Aurizio, and Tartaglia-Polcini 2008). However, much manual checking is still involved, and the system may not be replicable elsewhere.

4.6. IAB

We highlight the German Institute for Employment Research, a unit of the German Federal Employment Agency, because they are the only data provider that has succeeded in tackling the thorny issue of cross-national legal enforcement of researcher access to their lawyers' satisfaction. Put differently, they are currently the only institution that has thin client-based remote access between RDCs on both sides of the Atlantic (RDC in RDC approach). The prototype was launched at the University of Michigan, and has been expanded to include Cornell University and University of California at Berkeley this year (opening of the Cornell site is expected in April 2013). The thin client approach is combined with licensing of "scientific use files" to North American researchers (with some access restricted through the use of Cornell's CRADC). US researchers will be able to access the same remote compute center in Germany as German and other European researchers have accessed from several university-based thin clients for several years now. The success of the IAB is notable, since other European countries have ruled out such access, due to the different legal environments and the inability to enforce contracts.

5. Discussions of alternate access mechanisms

In this section, we will discuss the specific case of Industry Canada, and options it may have to increase researcher access to business data. In this, we take for granted the current and nascent access to some data through CEDR, but explore additional avenues, as well as the costs and risks such strategies might entail.

5.1. Outreach and direct support

A significant avenue to enhance new research using Canadian data, conditional on current access models, is to actively reach out to foreign researchers. In general, foreign researchers cannot easily access the restricted-access datasets due to both residency requirements and cost. By inviting and sponsoring researchers to come to Canada, either at Industry Canada or at Canadian universities, the breadth of ideas that are tested against Canadian data is greatly increased. This may go hand-in-hand with an expansion of internal statistical capabilities that are tailored to the needs of Industry Canada. The German IAB has used local collaborations with universities as well direct funding of visiting post-graduate researchers to further research using their in-house data, even when such data is not available through other access mechanisms. Researchers are located on IAB's premises, and interact with IAB research staff for periods ranging from days to months. New remote access locations in the United States and elsewhere in Europe have made such outreach more robust.

Support for Canadian researchers, both for research visits to Industry Canada as well as on focussed grants to Statistics Canada, in-house statistical internships by students (economists, statisticians) may also be useful.

The use of commercial databases (Compustat, Bloomberg, etc.) is generally expensive for universities. Providing grants or subsidies to use these data sources, in augmentation to or in replacement of data provided by Statistics Canada through its different data publication outlets, may also enhance research into Canadian issues.

5.2. Self-provision of access to data

With the declining cost of secure remote access, and the increased robustness and acceptance of secure computing models, it may be feasible to provide data not otherwise available, to researchers on-premise, through licensing models, or through remote processing capabilities. This applies in particular for novel data derived from administrative sources not yet provided through Statistics Canada, or surveys sponsored by Industry Canada. The costs for self-provision may be considerable (contract enforcement, IT infrastructure) but need to be placed in the context of alternative methods of providing data.

5.3. Provision of data through Canadian RDC nework

Secure data enclaves are sometimes used by multiple data providers, for instance in the United States (Census Bureau, NORC Data Enclave, CRADC) and France (CASD0. There are substantial benefits to providing data from multiple providers through a single secure location (Committee on National Statistics, 2005; page 77). Many of the Canadian RDCs have themselves roles that go beyond the access provision to data provided by Statistics Canada. For instance, the Québec Inter-University Centre for Social Statistics (CIQSS) also provides access to data provided by the Institut de la statistique du Québec, including entreprise microdata (although access procedures are different than for demographic and household data). All RDCs have extensive experience in providing secure access, and in managing the project proposal process. This experience and the facilities may be of use to providing access to confidential microdata not presently available through any Statistics Canada facility.

5.4. Support for research into novel statistical approaches

This overview would not be complete without highlighting the potential benefits of investing in new methods of creating new public-use data products, where the confidentiality protection uses novel methods (noise infusion, partial or full synthesis). Advances in the literature on statistical disclosure limitation have lead the U.S. Census Bureau to release products with levels of detail that would have been impossible to achieve with “traditional” disclosure protection mechanisms. We highlight in particular the synthesis methods underlying OnTheMap ((Machanavajjhala et al. 2008), which allows for highly detailed geography to play a role where it is needed, even when entities are sparsely distributed, and the attempts to create fully synthetic business microdata (Kinney et al. 2011), which aim for analytic validity along narrow dimensions, but also open the door for efficient remote processing. We note that the synthetic business microdata described in Kinney et al. 2011 are different from the “dummy” files sometimes labelled “synthetic”, because the former achieve limited analytic validity along a few dimensions, greatly enhancing the early exploratory work that research relies upon. The use of partially synthetic data in ACS (Hawala 2008) has a direct corollary for business data: when data are too sparse to release due to data confidentiality concerns, it may be possible to replace only the attributes of that sub-population with draws from synthetic data. In each case, the new confidentiality protection methods have also improved the proof of proper protection, by relying on hard metrics rather than weak secrecy of the protection methods.

6. References

Abowd, John M, Kaj Gittings, Kevin L McKinney, Bryce E Stephens, Lars Vilhuber, and Simon Woodcock. 2012. “Dynamically Consistent Noise Infusion and Partially Synthetic Data as Confidentiality Protection Measures for Related Time-series.” In *FCSM*.

Abowd, John M, Bryce E Stephens, Lars Vilhuber, Kevin L Mckinney, Marc Roemer, Simon Woodcock, and Fredrik Andersson. 2009. “The LEHD Infrastructure Files and the Creation of the

Quarterly Workforce Indicators." In *Producer Dynamics: New Evidence from Micro Data*, ed. Mark J. Roberts, Timothy Dunne, and J. Bradford Jensen.

Abowd, John M, Martha Stinson, and Gary Benedetto. 2006. *Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project*.

Abowd, John M, and Simon D Woodcock. 2001. "Disclosure Limitation in Longitudinal Linked Data, Appendix A: Recent Research on Disclosure Limitation." In *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, ed. Laura M Zayatz, Julia I Lane, Pat Doyle, and Jules J M Theeuwes, 215–277. New York: Elsevier Science.

Bruno, Giuseppe, Leandro D'Aurizio, and Raffaele Tartaglia-Polcini. 2008. "Remote Processing of Firm Microdata at the Bank of Italy."

Evans, Timothy, Laura Zayatz, and John Slanta, 1998. "Using Noise for Disclosure Limitation of Establishment Tabular Data." *Journal of Official Statistics*.

Hawala, Sam. 2008. "Producting Partially Synthetic Data to Avoid Disclosure." In *JSM 2008 - Section on Government Statistics*, 1345–1350.

Kennickell, Arthur. 1998. *Multiple Imputation in the Survey of Consumer Finances*.

Kinney, Satkartar K, Jerome P Reiter, Arnold P Reznek, Javier Miranda, Ron S Jarmin, and John M Abowd. 2011. "Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database." *International Statistical Review* 79 (3) (December): 362–384.

Little, Roderick J A. 1993. "Statistical Analysis of Masked Data" 2 (2): 407–426.

Machanavajjhala, A, D Kifer, J Abowd, J Gehrke, and L Vilhuber. 2008. *Privacy: Theory Meets Practice on the Map. 2008 IEEE 24th International Conference on Data Engineering*. Vol. 00. Ieee. doi:10.1109/ICDE.2008.4497436.

Miranda, Javier, and Ron Jarmin. 2002. *The Longitudinal Business Database*.

Nations, United. *UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE Managing Statistical Confidentiality & Microdata*.

Rubin, Donald B. 1993. "Satisfying Confidentiality Constraints Through Use of Synthetic Multiply-imputed Microdata (Discussion: Statistical Disclosure Limitation)" 9: 461–468.

U.S. Census Bureau. 2012. *Center for Economic Studies and Research Data Centers Research Report : 2010 and 2011*.

Weinberg, Daniel H, John M Abowd, Sandra K. Rowland, Philip M. Steel, and Laura Zayatz. 2007. *Access Methods for United States Microdata*.