A Web–Based Query System for Disclosure–Limited Statistical Analysis of Confidential Data

This document may not be distributed without permission of the National Institute of Statistical Sciences.

1 Introduction

Responding directly to the Federal government's role as the nation's largest gatherer and consumer of data, we propose a cross-disciplinary collaboration among Federal agencies and academic researchers to design and build a (prototype) World Wide Web-based, history-dependent query system. By disseminating information and knowledge derived from data, rather than data themselves, the system will preserve confidentiality of Federal data bases and the privacy of those to whom the data pertain. The system will support and leverage core missions of Federal agencies as well as commercial and other public uses of Federal data.[1] The contexts include formulation and evaluation of policy at national, state and local levels, crisis management (Schorr & Stolfo, 1997), and protection of the nation's infrastructure (PCCIP, 1997).

The research advances in computer science, social science and statistical science necessary to make such a system work are the thrust of the proposal. The specific issues are deep, but concrete and finite, ranging over multiple disciplines. **Computer science** will be used to formulate fundamental abstractions that allow the system to process queries and maintain a complex, dynamic query history data base (QHDB). **Statistics** will provide models for the disclosure risk reflecting such factors as the query history and user behavior, together with strategies to reduce high risks. **Social science domain knowledge** from the disciplines to which the data pertain — in our case, demography, economics, the environment and health — will be central in determining what statistical analyses are provided, selecting core analyses whose results will be available to all users, and defining risk. **Visualization** tools will enable deeper understanding of disclosure risk, provide graphical responses to queries and support evaluation of usage and performance of the system. **Software and systems engineering** will integrate the components into a functioning system.

The disciplines merge throughout, but most sharply in dealing with metadata needed for the system to function, the multiplicity of roles of the QHDB, and iteration between computation of disclosure risk and selection of risk reduction strategies.

Scale is another cross-cutting theme. Many existing techniques (e.g., for reducing disclosure risk) are untried at the scale (complexity of analyses, dimension and size of data sets, number of users) of the proposed system. One key challenge of the research, therefore, is simply to make the system work.

A cross-disciplinary, multi-institution research team, led by the National Institute of Statistical Sciences (NISS), will address the issues in an integrated manner, a crucial element for success. The team comprises leading researchers from NISS, Carnegie Mellon University (CMU), Kansas State University (KSU), Los Alamos National Laboratory (LANL), MCNC,[2] Ohio State University (OSU) and the University of Maryland College Park (UMCP). Four leading Federal statistics agencies — the Bureau of the Census (Census), the Bureau of Labor Statistics (BLS), the National Agricultural Statistics Service (NASS) and the National Center for Health Statistics (NCHS), as partners in the proposal, will provide access to data and personnel to participate in development and evaluation of the system. A senior staff member from each will serve on the project's User Advisory Committee (§9.3).

[1] Limited agency resources simply cannot meet burgeoning demands to access data for research purposes.

[2] Formerly, the Microelectronics Center of North Carolina.

In §2 we summarize the setting, goals and impact of the research. A description of system operation and architecture appears in §3. Research issues, organized by system function, are described §4–7. Some longer term questions appear in §8. Project management is discussed in §9.

2 Setting, Goals and Impact of the Research

The system will process queries from networked users of a Federal target data base (TDB),[3] performing and reporting statistical analyses that *extract knowledge from the data but preserve confidentiality*. The system's distinguishing characteristic is *history-dependence*: the response to each query will depend on the history of previous queries and responses.

2.1 Setting: Disclosure of Federal Data

The Federal government is uniquely responsible for "protecting the interests of data subjects through procedures that ensure privacy and confidentiality, enhancing public confidence in the integrity of statistical and research data, and facilitating the responsible dissemination of data to users" (Duncan, *et al.*, 1993). Deep public concern about privacy (Diffie & Landau, 1998; NRC, 1997) is subjecting agency practices for protecting confidentiality (identities of individuals and establishments, and sensitive data elements such as mental illness) to increasing scrutiny. New technologies expose data to both external attacks and abuse by legitimate users.

Confidentiality and privacy are violated by disclosure: the "[dissemination] of data in a manner that would allow public identification of the respondent or would be in any way harmful to him [*sic*]" (PCFS, 1971), or "inappropriate attribution of information to a data subject" (FCSM, 1994); see also CMPS (1997) and (Duncan, *et al.*, 1993). Disclosure can occur in multiple ways (§6), typically by record linkage from the TDB to an external data base (EDB)[4] or by personal knowledge.

2.2 Goals

The project has two high-level goals. The first is to use information technology and statistics to disseminate information and knowledge contained in data without compromising confidentiality of the data themselves. The second goal is to deploy a laboratory and testbed to (1) Develop and compare different approaches to disclosure limitation; and (2) Understand how a live user community, with widely varying demands and sophistication, employs the system, and how it performs in response.

To this end, two versions of the system will be built; see Table 1. Version 1, to be finished within twelve months, will have incomplete capabilities, will demonstrate proof-of-concept. Version 2, to be completed by the end of the three-year project, will have greatly increased functionality: multiple user classes, multiple query modes, user-provided data bases, regressions as analyses, entity-based risk evaluation, multiple risk reduction strategies and interactive visualizations.

2.3 Impact

In the past, confidentiality has been preserved by two principal strategies. *Restricted access*, exemplified by the Census's Regional Data Centers, limits who can use the data, by allowing (physically controlled) access only after detailed vetting of users and proposed analyses (Jabine, 1993a,b). Similarly, the National Center

[3]For example, the National Health Interview Survey; see §3.1.

[4]Software for this purpose, for example, *AutoMatch* (MatchWare, 1998), is available commercially.

for Education Statistics (NCES) licenses data sets to approved users. *Restricted data*, by contrast, allows public access to a subset of the TDB on which no analysis can compromise confidentiality.[5]

The proposed system will lessen the need for both. It will allow agencies to expand access to data at the same time as enlarging what can be learned from the data. Equally important, the system will avoid the perception of disclosure, thereby facilitating collection of high quality data. (It will for example, mitigate disincentives to participate in surveys for fear of loss of privacy.)

Specific examples: the BLS will be able to provide researchers analyses for specific geographical areas, industries, occupations, products, and occupation injuries and illnesses that are not available currently; researchers can perform state-level analyses on NCHS data, detecting disease epidemics earlier than now possible; multiple Federal agencies, as recommended in PCCIP (1997), can merge and share information about possible intrusions into their information infrastructures.

Moreover, the system complements and extends existing systems, including the Data Access and Dissemination System (DADS), under development by the Bureau of the Census, and Tau Argus, developed by Statistics Netherlands. DADS will be a static dissemination system for tables from the 2000 Census, with all disclosure limitation (data swapping and restrictions on geography, sensitive attributes and table dimension, size and sparseness; see Moore (1996)) performed beforehand. Tau Argus, which also is static, limits disclosure by combining local suppression with global recoding (Willenborg & Hunderpool, 1998).

3 System Operation and Architecture

How the system will operate — extracting low-disclosure-risk knowledge from the TDB as it responds to user queries — appears in §3.1. The overall system architecture is described in §3.2. For this discussion, the TDB can be thought of as containing only character data, with entities (rows) representing individuals or establishments and attributes (columns) that are categorical or numerical.

3.1 Operation

Key steps in processing and responding to queries are:

Metadata: Via a Web browser interface, the system will convey to users metadata on the TDB — its format and higher-level information about the context and quality of the data.

Handling Queries: Employing Web browser–based clients, users will formulate queries, which the client software will reduce to a standard form. Queries may request analyses that integrate user data bases (UDBs) with the TDB, in which case the system will access and process metadata on them.

Inference: Statistical analyses (tables, logistic regressions and linear and log-linear models) will be performed by an inference engine.

Risk Evaluation and Reduction: Using a detailed QHDB and statistical models, the disclosure risk associated with the query will be estimated, accounting for the query itself, previous analyses and EDBs that constitute paths for record linkage. The specific EDBs depend on the TDB and the query. Also, the distinction between EDBs and UDBs is situational: what in one setting is an EDB might in another be a UDB.

Iteration: Should the estimated disclosure risk be too high, the system will implement statistically–based strategies to reduce it to an acceptable level, iterating if necessary.

[5] Sampling or perturbations of the data (Marsh, *et al.*, 1994; Chen & Keller-McNulty, 1998) may be involved as well.

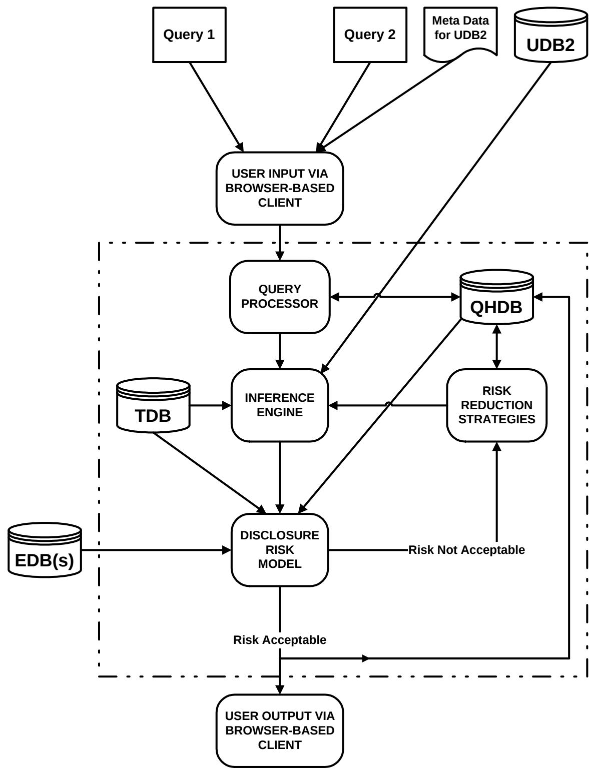

Figure 1: Overall View of the System. Both the browser-based client and the secure server are shown; the latter is contained within the dashed box. Two queries are depicted: Query 1 involves only the target data base, while Query 2 involves a user data base and associated metadata. Visualization and evaluation functions are not shown.

Response: Output provided via the Web interface will consist of results of analyses (tables or regression coefficients), supplementary statistical information (standard errors of estimators, significance levels for tests or "studentized" residuals from a linear model), and information about risk reduction strategies that were applied.[6]

Figure 1 illustrates pictorially, but omits two other essential functions:

Visualizations to understand and reason about risk, as responses (graphs and plots) to queries, as risk reduction strategies, and for monitoring system usage and performance.

Evaluation to understand how the system is used, how it performs operationally and in protecting confidentiality, and whether it meets users' needs.

The National Health Interview Survey (NHIS) (NCHS, 1998), conducted annually, illustrates key items.

TDB. The five main tables of the NHIS contain information about households, individuals (linked to households), health conditions, doctor visits and hospitalizations (the latter three linked to individuals). Numbers of records range from 11,000 to 116,000, of approximately 100 attributes each. The NHIS contains (geo-coded) information at zip code level, but only data aggregated to multi-state regions are released currently.

[6] The latter is essential for proper interpretation of the results.

Metadata consists primarily of descriptions of variables, together with some sampling weights. Metadata are available in hard copy form and on CD-ROMs containing public use files.

Queries of interest to the user community for the NHIS are principally tables and regressions. Currently, queries at the state level — for example, to compare disease incidence — are not permitted; the proposed system will allow informative responses to them.

Particular attention focuses on analyses that exploit geographical detail without creating disclosures. For example, suppose that the TDB contains state-level attributes $a_1, \ldots, a_K$. Then a user may submit a query requesting the linear regression

$$a_1 = b + \sum_{j=2}^{K} c_j a_j + \sum_{\ell=1}^{M} c_\ell^{\mathrm{UDB}} a_\ell^{\mathrm{UDB}}. \tag{1}$$

where $a_1^{\mathrm{UDB}}, \ldots, a_M^{\mathrm{UDB}}$ are additional state-level attributes contained in a UDB.[7] As compared to a query against only the TDB, (1) presents additional risks of disclosure.

EDBs relevant to (record-linkage) disclosure from the NHIS contain health, physician or hospitalization data from other sources: the Census, the National Cancer Institute and insurance companies.

3.2 Client–Server Architecture

In order that the system serve customers in near-real-time at the anticipated scale of hundreds to thousands of queries per day, the high-level architecture will be Unix server – Web client. Clients may be PCs, Macintoshes, X-terminals or Unix machines.

Client Software. Primary functionality of the client software will be (1) Formation of queries; (2) Translation of queries to standard format; and (3) User output. The software will be built around existing Web browsers; Netscape's release of the source code for Navigator™ (Netscape, 1998) makes this feasible.

Building on existing knowledge in human–computer interaction (HCI) (Mullet & Sano, 1995; Shneiderman, 1998) and NISS research on Web–based tools for dissemination of data and scientific documents (Eick, *et al.*, 1997a,b), the Web interface will *both facilitate and constrain* construction of queries by users. Version 1 of the system[8] will constrain queries by allowing only those constructed via mouse-based input, as in Figure 2.[9] Version 2 will allow a combination of mouse-based input, command line input and uploaded command files containing, for example, FORTRAN, SAS® code and SQL (Van Der Lans, 1998).

Details of how the interface will make the query history available as guidance for query construction must be worked out; visualizations of "query space" (§7.1) will play a central role.

The interface will convey to the user the results of statistical analyses, using character-based output (tables, values of regression coefficients, standard errors) and graphical displays.

Server Software will be a combination of compiled C/C++ code, Perl and `cgi-bin` scripts, Java and a high–performance data base such as Oracle®. The system will be modular and extensible, so that additional statistical analyses or new risk reduction strategies can be inserted as desired.

Interoperability with current and planned Federal information systems is essential. In particular, the planned Census Metadata Repository is likely to serve as the initial source of metadata for UDBs.[10]

---

[7] Which the system must then access.

[8] See §9.4.

[9] DADS plans a graphical interface with similar functionality. Version 4 of S-Plus® allows point-and-click building of regression models.

[10] Obviating the need for the system to deal with UDBs of unverified structure and content.

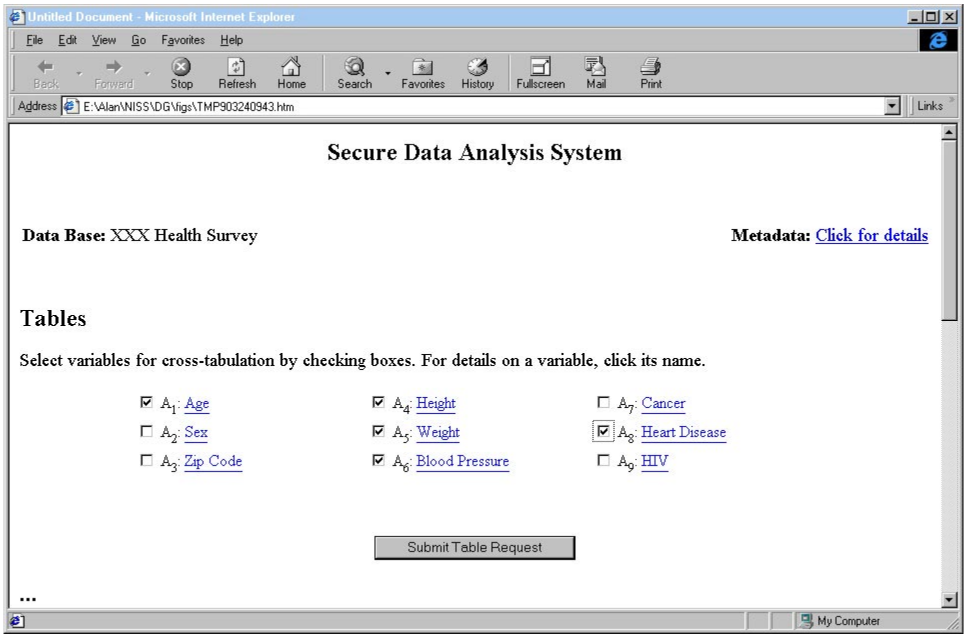

Figure 2: Prototype HTML interface, showing how TDB metadata might be conveyed and queries constructed. The query TAB[a_1, a_4, a_5, a_6, a_8] is requested.

Real-time operation also raises more subtle issues. For example, concurrency conditions must be detected and resolved, lest confidentiality be broken by multiple, simultaneous queries.

Security. An on-line Federal computer system allowing access to confidential data is a certain target of hackers. While security *per se* is not a major thrust of this project, nevertheless the right abstractions and "hooks" for a range of protective measures must be present. These measures include system security (firewalls), encryption and user authentication (Adams & Wortmann, 1989; Gusfield, 1988; Schlöere, 1981), and detection of anomalous user behavior (Anderson, *et al.*, 1995; DuMouchel & Schonlau, 1998a,b; Lunt & Jagannathan, 1988).

User Classes and needs must be characterized and modeled in order that disclosure risk be estimated properly and risk reduction strategies (RRSs) chosen appropriately. In Version 2, the system will support multiple user classes, for example, "public" users, "vetted" researchers and government agencies other than the owner of the TDB.[11]

User privileges will be class-dependent. Our approach will emphasize queries, risks and responses. For example, users can be distinguished on the basis of whether queries will be answered that extend the "revealed" portion of query space (§5.1).

Core Analyses. The history-dependent nature of the system raises the possibility that numerous or "unusual" queries could drive it to a state from which queries of broader interest could be answered either not at all or only by imposing very severe risk reduction strategies. To prevent this, we will construct a set of core analyses, whose results — possibly involving disclosure limitation — will be available to all users from the start. The research challenge is to balance (domain-dependent, as determined by expert social scientists) measures of the "broad interest" of queries against disclosure risk.

[11] "Super-users" with unrestricted access to the TDB (for example, employees of the owning agency) can also be accommodated.

Inference Engine. The statistical analyses requested by users will be performed by the inference engine. Absent the need for risk reduction, queries will be responded to as follows:

Tables: A query requesting TAB$[a_1, \ldots, a_K]$ receives this table.

Logistic Regressions: A query requesting LGR$[a_1, \ldots, A_K]$ receives at least estimated coefficients and their standard errors.

Linear Regressions: The response to a query requesting LIN$[a_1|a_2, \ldots, a_K]$ consists, minimally, of estimated coefficients $\hat{b}$ and $\hat{c}_j$, measures of model "fit" and standard errors for the estimates.

Support for Evaluation. The system will collect voluminous data about itself, in order to evaluate usage and performance (§9.5), including time-dependent numbers of users, response times to queries, and usage patterns. Some of these can be handled by the data base engine or the server operating system (e.g., HTTP log files), while instrumentation to collect others will be built into the system infrastructure. Still others (e.g., usage patterns) will be recorded by the client software, building on recent research on instrumentation and usage of Web browsers (Clemente, 1998; Cunha, *et al.*, 1995; Tauscher & Greenberg, 1997).

4 Metadata

For our purposes, *metadata* means "data and information about data." Minimally, metadata for a data base documents its format — the numbers of cases, and the number and definitions of attributes. Metadata may also contain higher-level information about the context, quality, condition, or characteristics of the data or the processes by which they were generated.

Different issues arise for each class of data bases. For the TDB, research issues are how metadata will be organized, represented and transmitted to users, how best to incorporate domain knowledge, and characterizing the ways in which metadata (for example, primary sampling units or sampling weights for a survey) may impact disclosure risk.

Metadata for UDBs (and, equally, EDBs) raises additional issues of organization and representation. Abstractions will be developed for UDB metadata that allow their integration into statistical analyses, as in (1), and calculation of disclosure risk. Some components of this abstraction, such as data base schema, are straightforward; others — for example, sampling frames and aggregation — are not. Added complications are elicitation of location, structure and metadata for UDBs.[12] New abstractions are necessary, as well, to encapsulate relationships between UDBs and the TDB.

5 Query Processing

A *query* specifies a statistical analysis to be conducted on the TDB. In Version 2, available analyses will be tables, logistic regressions and linear regressions (and visualizations; see §7), some of which may involve UDBs; Version 1 will allow only tables.

5.1 Parsing and Translation

Let $a_j(i)$ denote the value of attribute j for record i, with $a_0(\cdot)$ the primary key of the TDB. There are two principal abstractions: queries themselves and the query space $\mathcal{Q}$ in which they lie. For the former, preliminary forms are:

[12] For security, UDBs will be accessed by the system rather than uploaded by users.

Tables. For categorical attributes (say) $a_1, \dots, a_K$, TAB$[a_1, \dots, a_K]$ is the associative array, indexed by attribute categories, whose values are the numbers of elements in the TDB falling in those categories:

$$\text{TAB}[a_1, \dots, a_K](c_1, \dots, c_K) = \#\{i \in \text{TDB} : a_1(i) = c_1, \dots, a_K(i) = c_K\}\,. \tag{2}$$

Statistical analyses, risk evaluation and risk reduction may be based on multinomial models for tables: entities are allocated independently to cells, each falling in cell i with probability p_i.

Tables for numerical attributes are constructed by quantizing the attribute ranges, which effectively converts the attributes to categorical variables.

Logistic Regressions. The logistic regression LGR$[a_1, \dots, A_K]$ is the generalized linear model (Venables & Ripley, 1994)

$$\text{logit}(p) = b + \sum_{j=1}^{K} b_j a_j$$

for the probability p of a binary response as a function of the attributes, where $\text{logit}(p) = \log[p/(1-p)]$. As models for cell probabilities, logistic regressions overlap tables significantly.

Linear Regression. The linear regression LIN$[a_1|a_2, \dots, a_K]$ is the model

$$a_1(i) = b + \sum_{j=2}^{K} c_j a_j + \varepsilon_i \tag{3}$$

for one numerical attribute as a function of others. In (3), the ε_i are random errors, whose statistical structure constitutes part of the query. Variations include forcing $b = 0$ and transformations of the variables.

Implicit in the abstraction of a query and the notation above is a standardized format to which all queries, regardless of mode, are translated. As discussed in §3.2, this will be primarily a client-side function, which is embedded in the HTML form underlying Figure 2.

Query processing by the server will require strong interplay between computer science and statistics. In particular, the system must recognize that differently posed queries are equivalent mathematically, a process made more difficult by the fact that one query may be equivalent to a set of other queries. A simple example, with $\#(S)$ the cardinality of a set S, illustrates: suppose that Q and T are sets, say of individuals, and that $\#(Q)$ is too small to be released.[13] Then the query processor must recognize that releasing $\#(Q \cup T)$, $\#(Q \cup (\sim T))$, $\#(T)$ and $\#(\sim T)$, none of which may be small, releases $\#(Q)$ as well:

$$\#(Q) = \#(Q \cup T) + \#(Q \cup (\sim T)) - \#(T) - \#(\sim T)\,.$$

Abstractions of the query space $\mathcal{Q}$ will be constructed that enable definition of a "frontier" of (as-yet) unanswered queries, in the sense that answering them would increase the amount of released data. The approach will be to combine previous research on partial orderings of queries (Cox, 1980, 1981, 1987) with corresponding data structures (Moore & Lee, 1998). The same abstraction also leads to methods to select risk reduction strategies, to define user classes (§3.2), and to visualize risk (§7.3).

5.2 Query History Data Base

Because of the multiplicity of functions it must support, the QHDB is the most complex part of the system. The research issues are how to build it at all, and then how to build it well.

[13] Because this would compromise the confidentiality of information associated with the individuals in Q.

For system journaling and auditing, the minimal information that must be stored for each query is the user identity, the date and time, the standardized form, the UDBs involved, risk models employed and EDBs considered in computing disclosure risk, risk reduction strategies (if any) invoked, and the response[14] to the user. Because UDBs may be transient, merely recording their identities may not suffice; instead, metadata and even data elements accessed must be recorded in the QHDB.

Query parsing requires ascertaining whether a query, or an equivalent query or even set of queries (see §5.1), exists in the QHDB. What kinds of queries to the QHDB itself best support this function is a research issue to be addressed.

Disclosure risk computation for a current query q requires access methods that can determine which components of the QHDB are germane for models such as (4) below.

Response integrity must also be assured: if q (or an equivalent) has been posed before, then the system must respond *exactly* as it did previously. Because some risk reduction strategies involve randomization (§6.2), the QHDB must contain either the exact response or information (a random number seed) sufficient to reconstruct it.

Analysis of risk estimation models and risk reduction strategies, in order to understand which strategies (singly or in combination) are effective in what circumstances, is an essential testbed function of the system. This requires that the QHDB be accessed during selection of strategies, and also that the risk reduction effected by a strategy be stored. To evaluate models and strategies, it is as important to know when they do not work as when they do work, so the entire risk reduction iteration process (§6.3) will be maintained.

6 Risk Evaluation and Reduction

The TDB is vulnerable to disclosure of three kinds: (1) Identity disclosure, the identification of an individual or establishment (the primary key of the TDB) from other data;[15] (2) Attribute disclosure, the release of sensitive attributes (e.g., health status or income) of individuals or establishments, with or without identity disclosure; and, more generally, (3) Inferential disclosure, a statistical estimate of a sensitive attribute value with unacceptably low uncertainty (Duncan & Lambert, 1986, 1989; Keller & Bethlehem, 1992; Lambert, 1993; Skinner, 1992; Willenborg & de Waal, 1996).

Ordinarily disclosure risk is expressed as a probability, which can interpreted on either the population level (fraction at risk) or entity level (probability of disclosure). Query-by-query, risk must be estimated (§6.1) and, if necessary, reduced (§6.2).

6.1 Models for Disclosure Risk

Models of the form

$$\mathrm{DR} = f(q, \mathcal{C}, \mathcal{H}_Q, \mathrm{EDB}_1, \ldots, \mathrm{EDB}_k) \tag{4}$$

will be developed that estimate disclosure risk as a function of the query q itself, the core analyses $\mathcal{C}$, the query history $\mathcal{H}_Q$ and domain-dependent external data bases $\mathrm{EDB}_1, \ldots, \mathrm{EDB}_k$. Exemplifying the issues for a history-dependent system, risk estimates must account for increasingly tight bounds on categorical variables that can be derived from successive high-dimensional, complex, hierarchical tables (Duncan & Mukherjee, 1991; Fienberg, 1998).

Characterization of the variability of risk estimates is essential.

[14] In Version 1, HTML pages.

[15] Through record linkage or personal knowledge.

Disclosure risk models for queries involving tables will build on existing population-level models (Blien, *et al.*, 1992; Chen & Keller-McNulty, 1998; Fienberg & Makov, 1998; Greenberg & Zayatz, 1992; Keller-McNulty & Unger, 1993, 1998; Samuels, 1998; Unger & Keller-McNulty, 1990; Willenborg & de Waal, 1996). Such models assume that (one-to-one) links between the population entities and TDB records are established by a set of identifying attributes that form a k-dimensional cross-classification of the population, using knowledge of these attributes from EDBs. In the process, previous research on population uniques[16] will be extended to measure disclosure risk for "small" groups of subjects.

System-assisted procedures will be constructed that formalize the role of domain knowledge in calculating risk. To illustrate the issues, in a table indexed by doctors, patients and treatments: TAB[doctor, patient] and TAB[doctor, treatment] may be innocuous, but TAB[patient, treatment] may be highly sensitive.[17]

Research is also necessary to incorporate "indirect" disclosure in models such as (4), via release of seemingly benign information, such as sampling weights, that may be related to sensitive, non-released attributes (de Waal & Willenborg, 1997).

In Version 2, departing from past emphasis on population-level measures of risk, entity-risk analogues of (4) will be constructed. These will extend models (Fuller, 1993) for the release of perturbed microdata when EDB_1 (say) contains a corresponding subset of unperturbed data. Such models, which are highly subjective and computationally intensive, estimate the probability of correctly matching a released record within EDB_1. As a first step, population risk models will be used to identify which attribute sets would most damaging if known to the query poser. Such information can then be used to estimate entity-level risk.

Version 2 will also incorporate user a history $\mathcal{H}_U$ in (4). Difficult issues arise, including assumptions about or models of the behavior of legitimate users and intruders (Anderson, *et al.*, 1995; Fienberg, *et al.*, 1997) and statistical analysis of audit trails (Lunt, 1988; Lunt & Jagannathan, 1988).

Finally, versions of (4) for loss functions other than probability of disclosure must be developed. Ubiquity of EDBs may render almost any record linkable to some data base, and the relevant criteria become the cost or effort to effect a disclosure or the associated harm (Lambert, 1993).

6.2 Risk Reduction

Strategies to reduce the risk of disclosure operate on the data extracted from TDB itself, on the results of analyses, or by changing the analyses themselves.

Methods operating on the TDB itself alter it locally but preserve selected global characteristics. For tables, this typically means leaving low-dimensional marginal distributions unaltered. Many methods are known (FCSM, 1994; Fienberg, 1994; Fienberg, *et al.*, 1996; Duncan & Lambert, 1986; Willenborg & de Waal, 1996): swapping values from one record to another, jittering attributes by addition of random noise, and sampling records from the TDB. A primary research issue will be to evaluate these methods in a "live" setting.

Another question to be confronted is how the response to a query and selection of RRSs are to depend on that user's privileges and query history.

To quantify the risk reduction associated with a given query and RRS, we will use extensions (to be developed) of (4) of the form

$$\text{DR} = f(q, \text{RRS}, \mathcal{C}, \mathcal{H}_Q, \text{EDB}_1, \ldots, \text{EDB}_k). \tag{5}$$

To achieve this, parameterizations of RRSs are necessary. One formulation allowing this is that for numerical

[16] Cells with count one in the k-dimensional cross-classification.

[17] If, for example, one of the treatments is for AIDS.

data, many methods[18] can be characterized as *matrix masking*: a n-record, p-attribute TDB is replaced by

$$\mathrm{TDB}^* = A \times \mathrm{TDB} \times B + C,$$

where A, B and C are matrices, possibly involving randomness.

Risk reduction for linear models constitutes a new line of investigation. Global characteristics to be retained are low-order moments, such as means, standard deviations and correlations. Risk itself is more complicated than for tables, since it depends on both the estimated regression coefficients $\hat{b}$ and $\hat{c}_j$ and the diagnostic information associated with fit of the model, such as residuals.[19] New concepts and tools are needed to treat it.

A cross-cutting theme of this component of the research will be to quantify effects of risk reduction on uncertainties and on the conclusions of the analyses. Such understanding is essential in order that analyses be interpreted properly, and feeds back in subtle ways to how statistical analyses are performed.[20]

6.3 Iteration

As shown in Figure 1, risk evaluation and risk reduction may need to be iterated until the risk of disclosure is acceptably low. The process may also involve the inference engine, if analyses themselves are modified. In addition to being effective, iteration must not degrade system response.

Broadly, our approach will be to select RRSs that minimize distortion of the data[21] subject to disclosure risk being acceptably low. One part of the loop — *estimation* of risk reduction — employs models such as (5). For the second part — *selection* of candidate RRSs — we propose an expert system that reduces risk by one of two means. In Version 1, risk reduction will be effected via adjustment of parameters of RRS, for example, increasing the variance of additive noise or the degree of aggregation in reported results. Version 2 will also allow selection of an alternative strategy (and its parameters) on the basis of domain knowledge and the QHDB.

7 Visualization

We will use visualizations as responses to queries (§7.1), to reduce risk by conveying qualitative information without disclosing numerical details (§7.2), and to portray and reason about risk itself (§7.3) and system usage (§7.4).

7.1 Visualizations as Responses

Given current and emerging 3D graphics capabilities of personal computers, visualizations can play strong roles as responses to queries. Examples include histograms, scatter plots, other exploratory visualizations (Cleveland, 1993; Tufte, 1983) and plots of fitted regressions.

The system will incorporate interactive visualizations employing ActiveX™ components developed by Visual Insights (Visual Insights, 1998). These components, which run in standard Web browsers, include bar charts and histograms, illustrated in Figure 3, as well as scatter plots, data sheets and graph representations. Views referencing the same data are linked automatically.

---

[18] Including record or attribute suppression, swapping, jittering and aggregation.

[19] Ordinarily, the residuals $a_1(i) - \hat{b} - \sum_{j=2}^{K} \hat{c}_j a_j(i)$ are used to assess model fit, but knowledge of some attributes $a_j(i)$ and the residual for record i can lead to a predictive disclosure of $a_1(i)$.

[20] For example, a linear regression in which predictors are jittered should be performed using errors-in-variables techniques.

[21] By preserving global characteristics; see §6.2.

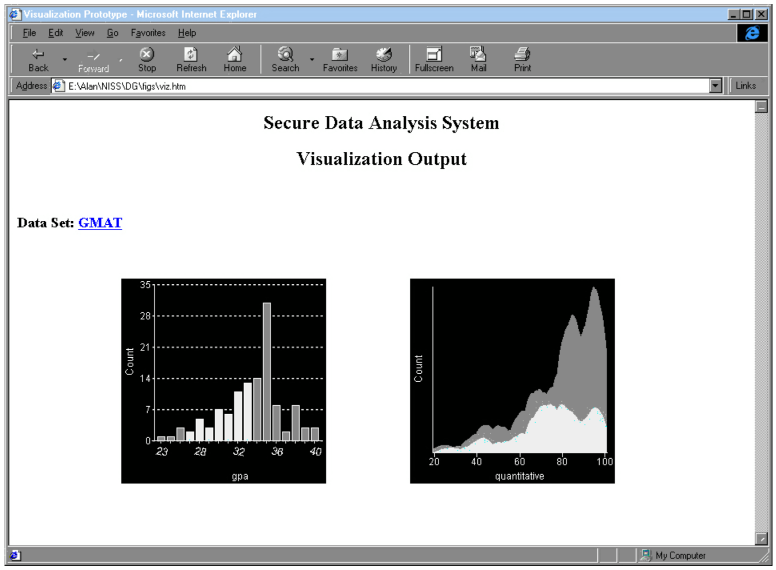

Figure 3: Example of Visualization Output, using VisualInsights™ ActiveX™components. Linked views of two attributes from a single data set are shown: selecting and highlighting a subset of the data in one highlights the same subset in the other. *Left:* Bar chart. *Right:* Histogram.

To provide visualizations to users requires an abstraction of their risk that can lead to analogues of models such as (4). Effectively no relevant previous research exists. A key research question: if q is a visualization query, is its risk as an image different from that of the data stream that creates the image? If so, then one strategy would be to transmit only GIF images generated dynamically by the server.

7.2 Visualization as a Risk Reduction Strategy

An entirely novel role of visualization is as a mechanism for risk reduction. While directly applicable previous research is lacking, promising techniques are not. In particular, we will evaluate the efficacy of technologies used in other settings to distort shapes in order to enhance clarity (Card, Mackinlay & Shneiderman, 1998; Munzner, 1997; Schaffer, *et al.*, 1996), but meant in this case to illuminate qualitative structure without divulging numerical details.

7.3 Visualization of Risk

The tools to be developed will support deeper understanding of the behavior of disclosure risk as a function of the query and the query history. Because the characteristics to be visualized are not inherently physical, the primary need is for compelling metaphors, principally for the query space $\mathcal{Q}$.

Initial approaches, building on existing tools (Becker, *et al.*, 1995; Fruchterman & Reingold, 1991; Wills, 1997), will treat queries for tables by means of network representations in which nodes correspond to queries and links to measures of “relatedness,” for instance, common attributes and attribute values. How to map query properties onto node characteristics (such as size, value, orientation, texture, shape and position (Bertin, 1981; Mullet & Sano, 1995)) and measures of relatedness onto link characteristics (width and color (Fortner and Meyer, 1997)), as well as the scale of our problems (dozens of link characteristics), raises needs that cannot be met with existing technologies. Deeper questions will be addressed for heterogeneous sets of queries containing, for example, both tables and regressions.

7.4 Visualization of Usage and Performance

Visualizations of usage and performance are essential to understanding, evaluation and improvement of the system. Those to be developed will merge network-oriented visualizations of the system's query history (§7.3), NISS research on characterizing computer usage via command sequences (DuMouchel & Schonlau, 1998a,b), and other existing visualization technologies for computer and network usage (Becker, *et al.*, 1995; Eick & Wills, 1993). Query sequences that select, filter or restructure information will receive particular emphasis. Previous research on information access and information spaces (Card, Mackinlay & Shneiderman, 1998; Fischer & Stevens, 1991; Marchionini, 1995) constitutes one thread of input.

8 Long Run Research Issues

Although complete resolution of these is beyond the scope of the proposed research, the system design must allow them to be accommodated at some future time.

Dynamic data bases containing either longitudinal (time-accumulating) or time-evolving data (McLeish, 1989; Unger & Keller-McNulty, 1990) require new abstractions of risk, place increased demands on the QHDB and create greater difficulty to maintain consistency of responses.

Multimedia data raise new metadata challenges (Sheth & Klas, 1998), and require new abstractions for risk, new models for risk and new strategies for risk reduction.

9 Management Plan

Primary management issues are personnel (§9.1), a User Advisory Committee (UAC) to ensure that the project meets agency and user needs (§9.3), and schedule (§9.4).

9.1 Personnel

Alan F. Karr, Associate Director of NISS, will be project director. At NISS, Karr has successfully initiated and managed projects of the same scientific, geographic and financial scope as this one. In addition to ensuring proper resource allocation and strong communication among project activities, Karr will lead research on visualization.

Four participants will have oversight responsibility for specific components of the research: **George T. Duncan** (Statistics, CMU), Task Leader for Risk Reduction; **Sallie Keller-McNulty**, (Statistics, LANL); Task Leader for Disclosure Risk Modeling; and **Latanya A. Sweeney** (Center for Automated Learning, CMU), Task Leader for System Design; and **Laura C. Vandiver** (MCNC), Task Leader for System Implementation.

Other senior participants are **Stephen E. Fienberg** (Statistics, CMU); **Michael J. Franklin** (Computer Science, UMCP); **Sousan Karimi** (MCNC); **Alan J. Saalfeld** (Civil and Environmental Engineering, OSU); and **Elizabeth A. Unger** (Computer Science, KSU).

Working closely with Karr and Sweeney, MCNC, which is adjacent to NISS, will provide the data base, network and programming expertise needed to create functioning systems.

Through Duncan, Fienberg and Sweeney, the project will access relevant expertise (for example, on data structures, as in Moore & Lee (1998)) from the Center for Automated Learning at CMU.

9.2 Federal Agency Partners

Partner Federal agencies[22] are:

Bureau of the Census will be the coordinating Federal agency. Census personnel will test and evaluate the system, and will also ensure that the system developers have access to the data needed for this project.

Bureau of Labor Statistics will provide data sets for pilot implementations. BLS staff will provide information on the data and rules currently being used in disclosure avoidance, develop examples of user queries, and review proposed output. The Behavioral Science Research Center will work with NISS to develop system evaluation protocols.

National Agricultural Statistics Service will be the lead agency for incorporating special requirements associated with geo-coded environmental data, will evaluate the system in this context, and will provide access to needed data.

National Center for Health Statistics will provide NISS data and documentation from the NHIS to serve as a pilot TDB; make available staff to explain the survey, its data and documentation; and evaluate confidentiality protections provided by the system.

9.3 User Advisory Council

The UAC will ensure that agency and user interests are represented throughout the project. It will also be a strong source for relevant domain knowledge. The UAC will meet regularly,[23] by teleconference, with the Project Director and Task Leaders. At least two face-to-face meetings will be held each year, in Washington.

The UAC will be co-chaired by **Martin H. David** (Economics, University of Wisconsin) and **Sarah M. Nusser** (Statistics, Iowa State University). Other members will be **Stephen Cohen** (BLS), **Carol House** (NASS), **Alvan Zarate** (NCHS) and **Laura V. Zayatz** (Census).

9.4 Schedule

Table 1 shows the primary characteristics of the two versions of the system. These define the schedule for the project. In particular, Version 1 specifies a set of clear, achievable milestones.

9.5 Evaluation

Evaluation will be the shared responsibility of project management and the UAC, through which agency input described in §9.2 will flow. We take this activity most seriously: to succeed the system must meet the needs of its users effectively and efficiently. Formal evaluations will take place at the end of each year of the project, as well as when test versions are implemented on specific TDBs.

10 Results from Prior NSF Support

Two ongoing NISS projects contain elements that point to the proposed research.

Software Engineering.[24] This is a pilot effort targeted at making scientific — by means of quantification, measurement and prediction — the hitherto elusive phenomenon of code decay. Participants are

[22]Letters of support are attached to this proposal.

[23]Initially, bi-weekly, and monthly as the project progresses.

[24]SBR–9529926, *Code Decay in Legacy Software Systems: Measurement, Models and Statistical Strategies,* Alan F. Karr, Principal Investigator.

Item	Version 1 (12 Months)	Version 2 (36 months)
User Classes	One	Other Agencies, Vetted Researchers, Public
Query Modes	HTML Forms	HTML Forms, Fortran, SAS® code, SQL
Query History Data Base	Oracle®	Oracle®
User Data Bases	No	Accessed via Census Metadata Repository
Analyses	Tables	Tables, Logistic Regressions, Linear Models
Visualizations	Static	Interactive
Risk Evaluation	Population-Based	Entity-Based
Risk Reduction Strategies	Suppression	Perturbation
Iteration	Parameters of Single Strategies	Multiple Strategies
Client Software	Existing browsers; HTML	Modified browsers; Java
Server Software	Perl; `cgi-bin`	C++
Data Base Engine	Oracle®	Oracle®
Data Base Access	SQL via Perl	JDBC; CORBA

Table 1: Versions 1 and 2 Compared Along Key Dimensions.

computer scientists, behavioral scientists, software engineers and statisticians from Lucent Technologies, Duke University, UMCP, University of North Carolina at Chapel Hill (UNC) and NISS.

Principal contributions to date:

- Tools for analysis and visualization of large-scale software data[25] (Eick, *et al.*, 1997a,b, 1998b; Graves, *et al.*, 1997b; Mockus, *et al.*, 1998).
- Quantified code decay indices that provide definitive evidence that code decay is real (Eick, *et al.*, 1998a), identify effects of code decay on software quality (Graves, *et al.*, 1997a) and quantify the deleterious impact of decay on the effort to make future changes (Graves & Mockus, 1998).
- A critical event history linking organizational and version management data (Staudenmayer, *et al.*, 1998).

Network Intrusion.[26] This collaboration with AT&T Research, involving computer scientists, network engineers and statisticians, is targeted at developing methods for on-line detection of network-based intrusion into computer systems.

The research leverages significant AT&T expertise in fraud detection for telephone systems, but the problem is more subtle. From the human side, there are multiple intruder motives. For organizations, losses are hard to quantify. Rather than observable in progress, intrusions are often not detected until the "damage is done." The massive scale of the data (gigabytes per day in large organizations) precludes analysis even if all data could be retained.

The main thrust of the research to date have been methods to detect anomalous user behavior and characterize usage on the basis of command histories, using transitions (DuMouchel & Schonlau, 1998a,b), user-unique commands and compression methods.

[25] In this case, a version management system containing the change history of some 200,000,000 lines of source code and headers.

[26] DMS–9700867, *Pilot Projects to Explore Large Data Sets.* Alan F. Karr and Jerome Sacks, Co-Principal Investigators.

References

Adam, N. R., and Wortmann, J. C. (1989). Security-control methods for statistical databases: A comparative study. *ACM Computing Surv.* **21** 515–556.

Anderson, D., Lunt, T. F., Javitz, H., Tamaru, A, and Valdes, A. (1995). Detecting unusual program behavior using the statistical component of the Next-Generation Intrusion Detection Expert System (NIDES). Technical Report, SRI, Inc.

Becker, R. A., Eick, S. G., and Wilks, A. R. (1995). Visualizing network data. *IEEE Trans. Visualization and Graphics* 16–28.

Bertin, J. (1981). *Graphics and Graphic Information Processing*, Walter de Gruyter & Co., Berlin.

Blien, U., Wirth, H., and Müller, M. (1992). Disclosure risk for microdata stemming from official statistics. *Statist. Neerland.* **46** 69–82.

Business Week. (1998b). The new ratings game. *April 27* 73–78.

Card. S. K., Mackinlay, J., and Shneiderman, B. (1998). *Readings in Information Visualization*. Morgan Kaufmann, San Francisco.

Chen, G., and Keller-McNulty, S. (1998). Estimation of identification disclosure risk in microdata. *J. Official Statist.* **14(1)** 79–95.

Clemente, P. (1998). *The State of the Net: The New Frontier*. McGraw–Hill, New York.

Cleveland, W. S. (1993). *Visualizing Data*. Hobart Press, Summit, NJ.

Committee on Maintaining Privacy and Security in Health Care Applications of the National Information Infrastructure. (1997). *For the Record: Protecting Electronic Health Information.* National Academy Press, Washington.

Cox, L. H. (1980). Suppression methodology and statistical disclosure control. *J. Amer. Statist. Assoc.* **75** 337–385.

Cox, L. H. (1981). Linear sensitivity measures in statistical disclosure control. *J. Statist. Plann. Inf.* **5** 152–164.

Cox, L. H. (1987). A constructive procedure for unbiased controlled rounding. *J. Amer. Statist. Assoc.* **82** 520–524.

Cunha, C R., Bestavros, A., and Crovella, M. E. (1995). Characteristics of WWW client-based traces. Technical Report TR-95-010, Department of Computer Science, Boston University. Available on-line at `cs-www.bu.edu/faculty/crovella/paper-archive/TR-95-010/paper.html`

Denning, D. E. (1982). *Cryptography and Data Security.* Addison-Wesley, Reading, MA.

Denning, D. E., and Denning, P. J. (1998). *Internet Besieged: Countering Cyberspace Scofflaws*. ACM Press, New York.

de Waal, T., and Willenborg, L. (1997). Statistical disclosure and sampling weights. *J. Official Statist.* **13(4)** 417–434.

Diffie, W., and Landau, S. (1998). *Privacy on the Line: The Politics of Wiretapping and Encryption.* MIT Press, Cambridge, MA.

DuMouchel, W., and Schonlau, M. (1998a). A fast computer intrusion detection algorithm based on hypothesis testing of command transition probabilities. *Proc. KDD 98*. (To appear)

DuMouchel, W., and Schonlau, M. (1998b). A comparison of test statistics for computer intrusion detection based on principal components regression of transition probabilities. *Proc. 30th Symp. Interface of Computing Sci. and Statist.* (To appear)

Duncan, G. T., Jabine, T. B., and de Wolf, V. A., eds., (1993). *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics.* Panel on Confidentiality and Data Access, Committee on National Statistics, National Academy Press, Washington.

Duncan, G. T.. and Lambert, D. (1986). Disclosure-limited data dissemination. *J. Amer. Statist. Assoc.* **81** 10–28.

Duncan, G. T. and Lambert, D. (1989). The risk of disclosure for microdata. *J. Business and Econ. Statist.* **7** 207–217.

Duncan, G. T., and Mukherjee, S. (1991). Microdata disclosure limitation in statistical databases: query size and random sample query control. *Proc. 1991 IEEE Conf. Security and Privacy* 278–287.

Eick, S. G., and Wills, G. J. (1993). Navigating large networks with hierarchies. *Visualization '93 Conference Proceedings* 204–210.

Eick, S. G., Graves, T. L., Karr, A. F., and Mockus, A. (1997a). A Web laboratory for software data analysis. *World Wide Web*. (To appear)

Eick, S. G., Mockus, A., Graves, T. L., and Karr, A. F. (1997b). Web-based text visualization. In Bandilla, W., and Faulbaum, F., eds. *Softstat '97 Advances in Statistical Software* **6** 3–10.

Eick, S. G., Graves, T. L., Karr, A. F., Marron, J. S., and Mockus, A. (1998b). Does code decay? Assessing the evidence from change management data. Submitted to *IEEE Trans. Soft. Engrg.*

Eick, S. G., Graves, T. L., Karr, A. F., and Mockus, A. (1998b). Visualizing software changes. (In preparation)

Federal Committee on Statistical Methodology (1994). *Report on Statistical Disclosure Limitation Methodology.* Statistical Policy Working Paper 22. Subcommittee on Disclosure Limitation Methodology. Office of Management and Budget, Executive Office of the President, Washington.

Fienberg, S. E. (1994a). Conflicts between the needs for access to statistical information and demands for confidentiality. *J. Official Statist.* **10** 115–132.

Fienberg, S. E. (1998). Fréchet and Bonferroni bounds for multi-way tables of counts with applications to disclosure limitation. *SDP '98*, Portugal, March, 1998.

Fienberg, S. E., and Makov, U. E. (1998). Confidentiality, uniqueness and disclosure limitation for categorical data. *J. Official Statist.* (To appear)

Fienberg, S. E., Makov, U. E., and Sanil, A. P. (1997). A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data. *J. Official Statist.* **13** 75–89.

Fienberg, S. E., Steele, R. J., and Makov, U. E. (1996). Statistical notions of data disclosure avoidance and their relationship to traditional statistical methodology: data swapping and log-linear models. *Proceedings of Bureau of the Census 1996 Annual Research Conference*. US Bureau of the Census, Washington, 87–105.

Fischer, G., and Stevens, K. (1991). Information access in complex, poorly structured information spaces. *Proc. ACM CHI '91 Human Factors in Computing Systems*, 63–70.

Fortner, B., and Meyer, T. E. (1997). *Number by Colors*. Springer, New York.

Fruchterman, T., and Reingold, E. (1991). Graph drawing by force-directed placement. *Software – Practice and Experience* **21(11)** 1129–1164.

Fuller, W. (1993). Masking procedures for microdata disclosure limitation. *J. Official Statist.* **9** 383–406.

Gouweleeuw, J. M., Kooiman, P., Willenborg, L. C. R. J., and de Wolf, P.-P. (1998). Post randomisation for statistical disclosure control: Theory and implementation. *J. Official Statist.* **14**. (To appear)

Graves, T. L., Karr, A. F., Marron, J. S., and Siy, H. (1997a). Predicting software faults using change history. *IEEE Trans. Soft. Engrg.* (To appear)

Graves, T. L., Karr, A. F., and Mockus, A. (1997b). Modeling software changes. In Minder, C. E., and Friedl, H., eds., *Proceedings of the 12th International Workshop on Statistical Modelling*.

Graves, T. L.. and Mockus, A. (1998). Inferring change effort from configuration management databases. *Metrics 98*.

Greenberg, B. G., and Zayatz, L. V. (1992). Measuring risk in public use microdata files. *Statist. Neerland.*

46 33–48.

Gusfield, D. (1988). A graph theoretic approach to statistical data security. *SIAM J. Computing* **17** 552–571.

Gusfield, D. (1990). A little knowledge goes a long way: false detection of compromised data in 2-D tables. *IEEE Symposium on Research in Security and Privacy*, 86–94.

Jabine, T. B. (1993a). Procedures for restricted data access. *J. Official Statist.* **9(2)** 537–589.

Jabine, T. B. (1993b). Statistical disclosure limitation practices of United States statistical agencies. *J. Official Statist.* **9(2)** 427–454.

Keller, W. J., and Bethlehem, J. G. (1992). Disclosure protection of microdata: problems and solutions. *Statist. Neerland.* **46** 5–19.

Keller-McNulty, S., and Unger, E. A. (1993). Database systems: Inferential security. *J. Official Statist.* **9** 475–479.

Keller-McNulty, S., and Unger, E. A. (1998). a remote access database prototype for the release of information based on confidential data. *J. Official Statist.* (To appear)

Lambert, D. (1993). Measures of disclosure risk and harm. *J. Official Statist.* **9** 313–331.

Lunt, T. F. (1988). Automated audit trail analysis and intrusion detection: A survey. *Proc. 11th National Computer Security Conf.*

Lunt, T. F., and Jagannathan, R. (1988). A real-time intrusion-detection expert system. *Proc. 1988 IEEE Symposium on Security and Privacy*.

Marchionini, G. (1995). *Information Seeking in Electronic Environments*. Cambridge University Press, Cambridge, UK.

Marsh, C., Dale, A., and Skinner, C. (1994). Safe data versus safe settings: Access to microdata from the British Census. *Internat. Statist. Rev.* **62** 35–53.

MatchWare Technologies Corporation (1998). *AutoMatch*. Information available on-line at `www.matchware.com`.

McLeish, M. (1989). Further results on the security of partitioned dynamic statistical databases. *ACM Trans. Database Syst.* **4** 98–113.

Mockus, A., Eick, S. G., Graves, T. L., and Karr, A. F. (1998). A system for analyzing software data. (In preparation)

Moore, A., and Lee, M. S. (1998). Cached sufficient statistics for efficient machine learning with large datasets. *J. Artificial Intell.* **8** 67–91.

Moore, R. A. (1996). Preliminary recommendations for disclosure limitation for the 2000 Censes: Improving the 1990 confidentiality edit procedure. Report RR 96–06, Statistical Research Division, Bureau of the Census, Washington.

Mullet, K., and Sano, D. (1995). *Designing Visual Interfaces: Communication Oriented Techniques*. Prentice Hall, Englewood Cliffs, NJ.

Munzner, T. (1997). H3: Laying out large directed graphics in 3d hyperbolic space. *IEEE Inform. Visualization Conf. Proc.* 2–10.

National Center for Health Statistics (1998). National Health Interview Survey (NHIS). Available on-line at `www.cdc.gov/nchswww/about/major/nhis/nhis.htm`.

Netscape Communications, Inc. (1998). Source code for *Netscape Navigator*™. Available on-line at `www.mozilla.org/download`.

National Research Council, Committee on Maintaining Privacy and Security in Health Care Applications of the National Information Infrastructure (1997). *For the Record: Protecting Electronic Health Information*. National Academy Press, Washington.

President's Commission on Critical Infrastructure Protection (1997). *Critical Foundations: Protecting America's Infrastructures.* Washington, DC. Available on-line at www.pccip.gov/report_index.html.

President's Commission on Federal Statistics (1971). *Federal Statistics: Report of the President's Commission on Federal Statistics*, I. US Government Printing Office, Washington.

Samuels, S. (1998). A Bayesian, population–genetics–inspired approach to the uniques problem in microdata disclosure risk assessment. *J. Official Statist.* (To appear)

Schaffer, D., Zuo, Z., Greenberg, S., Bartram, L., Dill, J., Dubs, S., and Roseman, M. (1996). Navigating hierarchically clustered networks through fisheye and full-zoom methods. *ACM Trans. Computer–Human Interaction* **3(2)** 162–188.

Schlörer, J. (1981). Security of statistical databases: Multidimensional transformation. *ACM Trans. Database Syst.* **6** 95–112.

Schorr, H., and Stolfo, S. J. (1997). *Towards the Digital Government of the 21st Century*. Report from the Workshop on Research and Development Opportunities in the Federal Information Services (Arlington, VA, May, 1997). Available on-line at www.isi.edu/nsf/prop.html.

Sheth, A., and Klas, W., eds. (1998). *Multimedia Data Management: Using Metadata to Integrate and Apply Digital Media*. McGraw-Hill, New York.

Shneiderman, B. (1998). *Designing the User Interface: Strategies for Effective Human–Computer Interaction*, 3rd ed. Addison–Wesley, Reading, MA

Skinner, C. J. (1992). On identification disclosure and prediction disclosure for microdata. *Statist. Neerland.* **46** 21–32.

Slack, J. (1992). *A Model for Object-Oriented Database Integrity*. Ph.D. dissertation, Computer Science, Kansas State University.

Smith, S., Bergeron, R. D., and Grinstein, G. G. (1990). Stereophonic and surface sound generation for exploratory data analysis. *SIGCHI '90 Proc.* 125–132.

Staudenmayer, N. A., Graves, T. L., Marron, J. S., Mockus, A., Siy, H., Votta, L. G., and Perry, D. (1998). Adapting to a new environment: how a legacy software organization copes with volatility and change. *Proceedings of Fifth International Product Development Management Conference*.

Tauscher, L., and Greenberg, S. (1997). How people revisit Web Pages: Empirical findings and implications for the design of history systems. *Internat. J. Human Computer Studies* **47(1)** 97–138. Available on-line at www.cpsc.ucalgary.ca/grouplab/papers/97-HowUsersRepeat.IJHCS/RevisitArticle.html

Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT.

Unger, E. A., and Keller-McNulty, S. (1990). The deterrent value of natural change in a statistical database. *Proc. 13th Annual National Computer Security Conf.* **16(1)** 59–68.

Van Der Lans, R. (1998). *The SQL Guide to Oracle*. Addison Wesley, Reading, MA.

Venables, W. N., and Ripley, B. D. (1994). *Modern Applied Statistics with S-Plus*. Springer, New York.

Visual Insights (1998). Visual Insights ActiveX components. Information available on-line at www.visualinsights.com/components.

Willenborg, L., and de Waal, T. (1996). *Statistical Disclosure Control in Practice.* Lecture Notes in Statistics, **111**. Springer, New York.

Willenborg, L., and Hunderpool, A. (1998). ARGUS for statistical disclosure control. In *Statistical Data Protection (SDP'98) Proceedings*. IOS Press. (To appear)

Wills, G. J. (1997). NicheWorks – interactive visualization of very large graphs. *Graph Drawing '97 Conf. Proc. Lecture Notes in Computer Science*. Springer, New York.