Journal of Privacy and Confidentiality (2010) **2**, Number 1, pp. 57–71

The Science and Technology of Privacy Protection: Appendix L of "Protecting Individual Privacy in the Struggle Against Terrorists"

Committee on Technical and Privacy Dimensions of Information for Terrorism Prevention and Other National Goals, National Research Council*

To the extent that there is a tension between counterterrorism efforts and protection of citizens' privacy, it is useful to understand how it may be possible to design counterterrorism information systems to minimize their impact on privacy. This appendix considers privacy protection from two complementary perspectives—privacy protection that is built into the analytical techniques themselves and privacy protection that can be engineered into an operational system. The appendix concludes with a brief illustration of how government statistical agencies have approached confidential data collection and analysis over the years. A number of techniques described here have been proposed for use in protecting privacy; none would be a panacea, and several have important weaknesses that are not well understood and that are discussed and illustrated.

1 The Cybersecurity Dimension of Privacy

Respecting privacy interests necessarily means that parties that *should* not have access to personal information *do* not have such access. Security breaches are incompatible with protecting the privacy of personal information, and good cybersecurity for electronically stored personal information is a necessary (but not sufficient) condition for protecting privacy.

From a privacy standpoint, the most relevant cybersecurity technologies are encryption and access controls. Encryption obscures digitally stored information so that it cannot be read without having the key necessary to decrypt it. Access controls provide privileges of different sorts to specified users (for example, the system may grant John Doe the right to know that a file exists but not the right to view its contents, and it may give Jane Doe both rights). Access controls may also be associated with audit logs that record what files were accessed by a given user.

Because of the convergence of and similarities between communication and information technologies, the technologies face increasingly similar threats and vulnerabilities. Furthermore, addressing these threats and vulnerabilities entails similar countermeasures or protection solutions. A fundamental principle of security is that no digital resource that is in use can be absolutely secure; as long as information is accessible, it is vulnerable. Security can be increased, but the value of increased security must be

*National Research Council of the National Academies, Washington, D.C., www.nap.edu

weighed against the increase in cost and the decrease in accessibility.

Human error, accident, and acts of God are the dominant sources of loss and damage in information and communication systems, but the actions of hackers and criminals are also of substantial concern. Terrorists account for a small percentage of losses, financial and otherwise, but could easily exploit vulnerabilities in government and business to cause much more serious damage to the nation. Security analysts and specialists report a large growth in the number and diversity of cyberthreats[1] and vulnerabilities.[2] Despite a concurrent growth in countermeasures (that is, security technologies[3]) penetrations and losses are increasing. A data-breach chronology reports losses of 104 million records (for example, in lost laptop computers) containing personally identifiable information from January 2005 to February 2007.[4] The Department of Homeland Security National Cyber Security Division reports that over 25 new vulnerabilities were discovered each day in 2006.[5]

The state of government information security is unnecessarily weak. For example, the U.S. Government Accountability Office (GAO) noted in March 2008 that

> [m]ajor federal agencies continue to experience significant information security control deficiencies that limit the effectiveness of their efforts to protect the confidentiality, integrity, and availability of their information and information systems. Most agencies did not implement controls to sufficiently prevent, limit, or detect access to computer networks, systems, or information. In addition, agencies did not always effectively manage the configuration of network devices to prevent unauthorized access and ensure system integrity, patch key servers and workstations in a timely manner, assign duties to different individuals or groups so that one individual did not control all aspects of a process or transaction, and maintain complete continuity of operations plans for key information systems. An underlying cause for these weaknesses is that agencies have not fully or effectively implemented agencywide information security programs. As a result, federal systems and information are at increased risk of unauthorized access to and disclosure, modification, or destruction of sensitive information, as well as inadvertent or deliberate disruption of system operations and services. Such risks are illustrated, in part, by an increasing number of security incidents experienced

[1] A.T. Williams, A. Hallawell, R. Mogull, J. Pescatore, N. MacDonald, J. Girard, A. Litan, L. Orans, V. Wheatman, A. Allan, P. Firstbrook, G. Young, J. Heiser, and J. Feiman, *Hype Cycle for Cyberthreats*, Gartner, Inc., Stamford, Conn., September 13, 2006.

[2] National Vulnerability Database, National Institute of Standards and Technology Computer Security Division, sponsored by the U.S. Department of Homeland Security National Cyber Security Division/U.S. Computer Emergency Readiness Team (US-CERT), available at `http://nvd.nist.gov/`.

[3] A.T. Williams, A. Hallawell, R. Mogull, J. Pescatore, N. MacDonald, J. Girard, A. Litan, L. Orans, V. Wheatman, A. Allan, P. Firstbrook, G. Young, J. Heiser, and J. Feiman, *Hype Cycle for Cyberthreats*, Gartner, Inc., Stamford, Conn., September 13, 2006.

[4] A Chronology of Data Breaches, Privacy Rights Clearing House.

[5] National Vulnerability Database, National Institute of Standards and Technology Computer Security Division, sponsored by the U.S. Department of Homeland Security National Cyber Security Division/U.S. Computer Emergency Readiness Team (US-CERT), available at `http://nvd.nist.gov/`.

by federal agencies.[6]

Such performance is reflected in the public's lack of trust in government agencies' ability to protect personal information.[7] Security of government information systems is poor despite many relevant regulations and guidelines.[8] Most communication and information systems are unnecessarily vulnerable to attack because of poor security practices, and the framework outlined in Chapter 2 identifies data stewardship as a critical evaluation criterion.[9]

Although cybersecurity and privacy are conceptually different, they are often conflated —with good reason—in the public's mind. Cybersecurity breaches—which occur, for example, when a hacker breaks into a government information system that contains personally identifiable information (addresses, Social Security numbers, and so on)—are naturally worrisome to the citizens who may be affected. They do not particularly care about the subtle differences between a cybersecurity breach and a loss of privacy through other means; they know only that their privacy has been (potentially) invaded and that their loss of privacy may have deleterious consequences for them. That reaction has policy significance: the government agency responsible (perhaps even the entire government) is viewed as being incapable of protecting privacy, and public confidence is undermined when it asserts that it will be a responsible steward of the personal information it collects in its counterterrorism mission.

2 Privacy-Preserving Data Analysis

2.1 Basic Concepts

It is intuitive that the goal of privacy-preserving data analysis is to allow the learning of particular facts or kinds of facts about individuals (units) in a data set while keeping other facts secret. The term *data set* is used loosely; it may refer to a single database or to a collection of data sources. Under various names, privacy-preserving data analysis has been addressed in various disciplines.

[6]Statement of Gregory C. Wilshusen, GAO Director for Information Security Issues, "Information Security: Progress Reported, but Weaknesses at Federal Agencies Persist," Testimony Before the Subcommittee on Federal Financial Management, Government Information, Federal Services, and International Security, Committee on Homeland Security and Governmental Affairs, U.S. Senate, GAO-08-571T, March 12, 2008. Available at `http://www.gao.gov/new.items/d08571t.pdf`.

[7]L. Ponemon, *Privacy Trust Study of United States Government*, The Ponemon Institute, Traverse City, Mich., February 15, 2007.

[8]Appendix III, OMB Circular A-130, "Security of Federal Automated Information Resources," (Office of Management and Budget, Washington, D.C.) revises procedures formerly contained in Appendix III, OMB Circular No. A-130 (50 FR 52730; December 24, 1985), and incorporates requirements of the Computer Security Act of 1987 (P.L. 100-235) and responsibilities assigned in applicable national security directives. See also Federal Information Security Management Act of 2002 (FISMA), 44 U.S.C. S 3541, et seq., Title III of the E-Government Act of 2002, Public Law 107-347, 116 Stat. 2899, available at `http://csrc.nist.gov/drivers/documents/FISMA-final.pdf`.

[9]Data stewardship is accountability for program resources being used and protected appropriately according to the defined and authorized purpose.

A statistic is a quantity computed on the basis of a sample. A major goal of official statistics is to learn broad trends about a population by studying a relatively small sample of members of the population. In many cases, such as in the case of U.S. census data and data collected by the Internal Revenue Service (IRS), privacy is legally mandated. Thus, the goal is to identify and report trends while protecting the privacy of individuals. That sort of challenge is central to medical studies: the analyst wishes to learn and report facts of life, such as "smoking causes cancer," while preserving the privacy of individual cancer patients. The analyst must be certain that the privacy of individuals is not even inadvertently compromised.

Providing such protection is a difficult task, and a number of seemingly obvious approaches do not work even in the best of circumstances, for example, when a trusted party holds all the confidential data in one place and can prepare a "sanitized" version of the data for release to the analyst or can monitor questions and refuse to answer when privacy might be at risk. (This point is discussed further in Section 2.2 below.)

In the context of counterterrorism, privacy-preserving data analysis is excellent for teaching the data analyst about "normal" behavior while preserving the privacy of individuals. The task of the counterterrorism analyst is to identify "atypical" behavior, which can be defined only in contrast with what is typical. It is immediately obvious that the data on any single specific individual should have little effect on the determination of what is normal, and in fact this point precisely captures the source of the intuition that broad statistical trends do not violate individual privacy. Assuming a good knowledge of what is "normal," technology is necessary for counterterrorism that will scrutinize data in an automated or semiautomated fashion and flag any person whose data are abnormal, i.e., that satisfy a putatively "problematic" profile. In other words, the outcome of data analysis in this context must necessarily vary widely ("yes, it satisfies the profile" or "no, it does not satisfy the profile"), depending on the specific person whose data is being scrutinized. Whether the profile is genuinely "problematic" is a separate matter.

In summary, privacy-preserving data analysis may permit the analyst to learn the definition of *normal* in a privacy-preserving way, but it does not directly address the counterterrorism goal: privacy-preserving data analysis "masks" *all* individuals, whereas counterterrorism requires the exposure of selected individuals. There is no such thing as privacy-preserving examination of an individual's records or privacy-preserving examination of a database to pinpoint problematic individuals.

The question, therefore, is whether the counterterrorism goal can be satisfied while protecting the privacy of "typical" people. More precisely, suppose the existence of a perfect profile of a terrorist: the false-positive and false-negative rates are very low. (The existence of such a perfect profile is magical thinking and contrary to fact, but suppose it anyway.) Would it be possible to analyze data, probably from diverse sources and in diverse formats, in such a way that the analyst learns only information about people who satisfy the profile? As far as we know, the answer to that question is no. However, it might be possible to limit the amount of information revealed about those who do not satisfy the profile, perhaps by controlling the information and sources used or by

editing them after they are acquired. That would require major efforts and attention to the quality and utility of information in integrated databases.

2.2 Some Simple Ideas That Do Not Work in Practice

There are many ideas for protecting privacy, and what may seem like sensible ideas often fail. Understanding how to approach privacy protection requires rigor in two senses: spelling out what "privacy protection" means and explaining the extent to which a particular technique succeeds in providing protection.

For example, assume that all the data are held by a trustworthy curator, who answers queries about them while attempting to ensure privacy. Clearly, queries about the data on any specific person cannot be answered, for example, What is the sickle-cell status of Averill Harriman? It is therefore instructive to consider the common suggestion of insisting that queries be made only on large subsets of the complete database. A well-known differencing argument (the "set differencing" attack) demonstrates the inadequacy of the suggestion: If the database permits the user to learn exact answers, say, to the two questions, How many people in the database have the sickle-cell trait? and, How many people—not named X—in the database have the sickle-cell trait? then the user learns X's sickle-cell trait status. The example also shows that encrypting the data (another frequent suggestion) would be of no help. Encryption protects against an intruder, but in this instance the privacy compromise emerges even when the database is operated correctly, that is, in conformance with all stated security policies.

Another suggestion is to monitor query sequences to rule out attacks of the nature just described. Such a suggestion is problematic for two reasons: it may be computationally infeasible to determine whether a query sequence compromises privacy,[10] and, more surprising, the *refusal* to answer a query may itself reveal information.[11]

A different approach to preventing the set differencing attack is to add random noise to the true answer to a query; for example, the response to a query about the average income of a set of individuals is the sum of the true answer and some random noise. That approach has merit, but it must be used with care. Otherwise, the same query may be issued over and over and each time produce a different perturbation of the truth. With enough queries, the noise may cancel out and reveal the true answer. Insisting that a given query always results the same answer is problematic in that it may be impossible to decide whether two syntactically different are semantically equivalent. Related lower bounds on noise (the degree of distortion) can be given as a function of the number of queries.[12]

[10] J. Kleinberg, C. Papadimitriou, and P. Raghavan, "Auditing boolean attributes," pp. 86- 91 in *Proceedings of 19th ACM Symposium on Principles of Database Systems*, Association for Computing Machinery, New York, N.Y., 2000.

[11] K. Kenthapadi, N. Mishra, and K. Nissim, "Simulatable auditing," pp. 118-127 in *Proceedings of the 24th ACM Symposium on Principles of Database Systems*, Association for Computing Machinery, New York, N.Y., 2005.

[12] I. Dinur and K. Nissim, "Revealing information while preserving privacy," pp. 202-210 in *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*,

2.3 Private Computation

The cryptographic literature on private computation addresses a distinctly different goal known as secure function evaluation.[13] In this work, the term private has a specific technical meaning that is not intuitive and is described below. To motivate the description, recall the original description of privacy-preserving data analysis as permitting the learning of some facts in a data set while keeping other facts secret. If privacy is to be completely protected, some things simply cannot be learned. For example, suppose that the database has scholastic records of students in Middletown High School and that the Middletown school district releases the fact that no student at the school has a perfect 5.0 average. That statement compromises the privacy of every student known to be enrolled at the school—it is now known, for example, that neither Sergey nor Laticia has a 5.0 average. Arguably, that is no one else's business. (Some might try to argue that no harm comes from the release of such information, but this is defeating the example without refuting the principle that it illustrates.) Similarly, publishing the average net worth of a small set of people may reveal that at least one person has a very high net worth; a little extra information may allow that person's identity to be disclosed despite her modest lifestyle.

Private computation does not address those difficulties, and the question of which information is safe to release is not the subject of study at all.[14] Rather, it is assumed that some facts are, by fiat, going to be released, for example, a histogram of students' grade point averages or average income by block. The "privacy" requirement is that no information *that cannot be inferred from those quantities* will be leaked. The typical setting is that each person (say, each student in Middletown High School) participates in a cryptographic protocol whose goal is the cooperative computing of the quantity of interest (the histogram of grade point averages) and that the cryptographic protocol will not cause any information to be leaked that a student cannot infer from the histogram and his or her own data (that is, from the grade point histogram and his or her own grade point average).

2.4 The Need for Rigor

Privacy-preservation techniques typically involve altering raw data or the answers to queries. Those general actions are referred to as input perturbation and output perturbation,[15] depending on whether the alterations are made before the queries or in

Association for Computing Machinery, New York, N.Y., 2003; C. Dwork, F. McSherry, and K. Talwar, "The price of privacy and the limits of LP decoding," pp. 85-94 in *Proceedings of the 39th Annual ACM SIGACT Symposium on Theory of Computing*, Association for Computing Machinery, New York, N.Y., 2007. See also the related work on compressed sensing cited in the latter.

[13] O. Goldreich, S. Micali, and A. Wigderson, "How to solve any protocol problem," pp. 218-229 in *Proceedings of the 19th ACM SIGACT Symposium on Computing*, Association for Computing Machinery, New York, N.Y., 1987.

[14] O. Goldreich, S. Micali, and A. Wigderson, "How to solve any protocol problem," pp. 218-229 in *Proceedings of the 19th ACM SIGACT Symposium on Computing*, Association for Computing Machinery, New York, N.Y., 1987.

response to them.

Various methods are used for input and output perturbation. Some involve redaction of information (for example, removing "real" identifiers, the use of indirect identifiers, selective reporting, or forms of aggregation) or alteration of data elements by adding noise, swapping, recoding (for example, collapsing categories), and data simulation.[16] But no matter what the technique or approach, there are two basic questions: What does it mean to protect the data? How much alteration is required to achieve that goal?

The need for a rigorous treatment of both questions cannot be overstated, inasmuch as "partially protecting privacy" is an oxymoron. An extremely important and often overlooked factor in ensuring privacy is the need to protect against the availability of arbitrary context information, including other databases, books, newspapers, blogs, and so on.

Consider the anonymization of a social-network graph. In a social network, nodes correspond to people or other social entities, such as organizations or Web sites, and edges correspond to social links between them, such as e-mail contact or instant-messaging. In an effort to preserve privacy, the practice of anonymization replaces names with meaningless unique identifiers. The motivation is roughly as follows: the social network labeled with actual names is sensitive and cannot be released, but there may be considerable value in enabling the study of its structure. Anonymization is intended to preserve the pure unannotated structure of the graph while suppressing the information about precisely who has contact with whom. The difficulty is that anony-

[15] A relevant survey article is N. Adam and J. Wortmann, "Security-control methods for statistical databases: A comparative study," *ACM Computing Surveys* 21(4):515-556, 1989. Some approaches post-dating the survey are given in L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10(5):557-570, 2002; A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," pp. 211-222 in *Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, Association for Computing Machinery, New York, N.Y., 2003; and C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity of functions in private data analysis," pp. 265-284 in *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*, Association for Computing Machinery, New York, N.Y., 2006, and references therein.

[16] Many of these methods are described in the following papers: S.E. Fienberg, "Conflicts between the needs for access to statistical information and demands for confidentiality," *Journal of Official Statistics* 10(2):115-132, 1994; Federal Committee on Statistical Methodology, Office of Management and Budget (OMB), "Statistical Policy Working Paper 2. Report on Statistical Disclosure and Disclosure-Avoidance Techniques," OMB, Washington, D.C., 1978, available at `http://www.fcsm.gov/working-papers/sw2.html`; Federal Committee on Statistical Methodology, OMB, "Statistical Policy Working Paper 22 (Second version, 2005), Report on Statistical Disclosure Limitation Methodology," originally prepared by Subcommittee on Disclosure Limitation Methodology, OMB, Washington, D.C., 1994, and revised by the Confidentiality and Data Access Committee, 2005, available at `http://www.fcsm.gov/working-papers/spwp22.html`. Many of these techniques are characterized as belonging to the family of matrix masking methods in G.T. Duncan and R.W. Pearson, "Enhancing access to microdata while protecting confidentiality: prospects for the future (with discussion)," *Statistical Science* 6:219-239, 1991. The use of these techniques in a public-policy context is set by the following publications: National Research Council (NRC), *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*, G.T. Duncan, T.B. Jabine, and V.A. de Wolf, eds., National Academy Press, Washington, D.C., 1993; NRC, *Expanding Access to Research Data: Reconciling Risks and Opportunities*, The National Academies Press, Washington, D.C., 2005.

mous social-network data almost never exist in the absence of outside context, and an adversary can potentially combine this knowledge with the observed structure to begin compromising privacy, deanonymizing nodes and even learning the edge relations between explicitly named (deanonymized) individuals in the system.[17]

A more traditional example of the difficulties posed by context begins with the publication of redacted confidential data. The Census Bureau receives confidential information from enterprises as part of the economic census and publishes a redacted version in which identifying information on companies is suppressed. At the same time, a company may release information in its annual reports about the number of shares held by particular holders of very large numbers of shares. Although the redaction may be privacy-protective, by using very simple *linkage tools* on the redacted data and the public information, an adversary will be able to add back some of the identifying tags to the redacted confidential data. Roughly speaking, those tools allow the merging of data sets that contain, for example, different types of information about the same set of entities. The key point is that entities need not be directly identifiable by name to be identified. Companies can be identified by industrial code, size, region of the country, and so on. Any public company can be identified by using a small number of such variables, which may well be deduced from the company's public information and thus provide a means of matching against the confidential data.

Similarly, individuals need not be identified only by their names, addresses, or Social Security numbers. The linkage software may use any collection of data fields, or variables, to determine that records in two distinct data sets correspond to the same person. And if the "privacy-protected" or deidentified records include values for additional variables that are not yet public, simple record-linkage tools might let an intruder identify a person (that is, match files) with high probability and thus leak this additional information in the deidentified files. For example, an adversary may use publicly available data, including newspaper accounts from New Orleans on the effects of hurricane Katrina and who was rescued in what efforts, to identify people with unusual names in a confidential epidemiologic data set on rare genetic diseases gathered by the Centers for Disease Control and Prevention and thus learn all the medical and genetic information about the individuals that redaction was supposed to protect.

For a final, small-scale, example, consider records of hospital emergency room admissions, which contain such fields as name, year of birth, ZIP code, ethinicity, and medical complaint. The combinations of fields are known to identify many people uniquely. Such a collection of attributes is called a quasi-identifier. In microaggregation, or what is known as k-anonymization, released data are "coarsened"; for example, ZIP codes with the same first four digits are lumped together, so for every possible value of quasi-identifier, the data set contains at least k records. However, if someone sees an ambulance at his or her neighbor's house during the night and consults the published hospital emergency room records the following day, he or she can learn a small set of

[17] L. Backstrom, C. Dwork, and J. Kleinberg, "Wherefore art thou R3579X? Anonymized social networks, hidden patterns, and structural steganography," pp. 181-190 in *Proceedings of the 16th International Conference on World Wide Web*, 2007, available at `http://www2007.org/proceedings.html`.

complaints that contains the medical complaint of the neighbor. Additional information known to that person may allow the neighbor's precise complaint to be pinpointed.

Context also comes into play in how different privacy-preserving techniques interact when they are applied to different databases. For example, the work of Dwork et al. rigorously controlled the amount of information leaked about a single record.[18] If several databases, all containing the same record, use the same technique, and if the analyst has access to all these databases, the cumulative erosion of privacy of the given record may be as great as the sum of the leakages suffered in the separate databases that contain it.

And that is a good case! The many methods in fields spanning computer science, operations research, economics, and statistics deal with data of different types recorded in many forms. For a targeted set of methods and specific kinds of data, although there may be results that can "guarantee" privacy in a released data file or a system responding to a series of queries, many well-known approaches fail to offer such guarantees or even weaker assurances. For example, some literature on data imputation for privacy protection never defines *privacy* at all;[19] thus, it is difficult to assess the extent to which the methods, although heuristically reasonable, actually guarantee privacy.

2.5 The Effect of Data Errors on Privacy

In the real world, data records are imperfect. For example,

- Honest people make errors when providing information.
- Clerical errors yield flawed recording of correct data.
- Many data values may be measurements of quantities that regularly fluctuate or that for various other reasons are subject to measurement error.

Because of imperfections in the data, a person may be mischaracterized as problematic. That is, the profile may be perfect, but the system may be operating with bad data. That appears to be an accuracy problem, but for several reasons it also constitutes a privacy problem.

[18] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity of functions in private data analysis," pp. 265-284 in *Proceedings of the 3rd Theory of Cryptography Conference*, Association for Computing Machinery, New York, N.Y., 2006.

[19] D.B. Rubin, "Discussion: Statistical disclosure limitation," *Journal of Official Statistics* 9(2):461-468, 1993; T.E. Raghunathan, J.P. Reiter, and D.B. Rubin, "Multiple imputation for statistical disclosure limitation," *Journal of Official Statistics* 19(2003):1-19, 2003. However, there is also a substantial literature that does provide an operational assessment of privacy and privacy protection. For example, see G.T. Duncan and D. Lambert, "The risk of disclosure for microdata," *Journal of Business and Economic Statistics* 7:207-217, 1989; E. Fienberg, U.E. Makov, and A.P. Sanil, "A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data," *Journal of Official Statistics* 13:75-89, 1997; and J.P. Reiter, "Estimating risks of identification disclosure for microdata," *Journal of the American Statistical Association* 100(2005):1103-1113, 2005.

Although we have not discussed a definition of *privacy*, the recent literature studies the appropriate technical definition at length. The approach favored in the cryptography community, modified for the present context, says that for anyone whose true data do not fit the profile, there is (in a quantifiable sense) almost no difference between the behavior of a system that contains the person's data and the behavior of a system that does not. That is, the behavior of the system in the two cases should be *indistinguishable*; it follows that the *increase* in the risk of adverse effects of participating in a data set is small. That approach allows us to avoid subjective decisions about which type of information leakage constitutes a privacy violation. Clearly, indistinguishability can fail to hold in the case of a nonterrorist whose data are incorrectly recorded. The harm to a person of *appearing* to satisfy the perfect profile may be severe: the person may be denied credit and the freedom to travel, be prevented from being hired for some jobs, or even be prosecuted. Finally, at the very least, such a misidentification will result in further scrutiny and consequent loss of privacy. (See Gavison on protection from being brought to the attention of others.[20])

The problem of errors is magnified by linkage practices because errors tend to propagate. Consider a database, such as the one assembled by ChoicePoint by linking multiple databases. Consider, say, three separate databases created by organizations A, B, and C. If A and B are extremely scrupulous about preventing data errors but C is not, the integrated database will contain inaccuracies. The accuracy of the integrated database is only as good as the accuracy of the worst input database. Furthermore, if each database contains errors, they may well compound to create a far greater percentage of files with errors in the integrated database. Finally, there are the errors of matching themselves, which are inherent in record linkage; if these are as substantial as the literature on record linkage suggests,[21] the level of error in the merged database is magnified, and this poses greater risks of misidentification.

All the above difficulties are manifested even when a perfect profile is developed for problematic people. But imperfect profiles combined with erroneous data will lead to higher levels of false positives than either alone. Moreover, if we believe that data are of higher quality and that profiles are more accurate than they actually are, the rate of false negatives—people who are potential terrorists but go undetected—will also grow, and this endangers all of us.

Record linkage also lies at the heart of data-fusion methods and has major implications for privacy protection and harm to people. The literature on record linkage[22] makes it clear that to achieve low rates of error (high accuracy) one needs both "good" variables for linkage (such as names) and ways to organize the data by "blocks," such as city blocks in a census context or well-defined subsets of individuals characterized by variables that contain little or no measurement error. As measurement error grows,

[20] R. Gavison, "Privacy and the limits of the law," pp. 332-351 in *Computers, Ethics, and Social Values*, D.G. Johnson and H. Nissenbaum, eds., Prentice Hall, Upper Saddle River, N.J., 1995.

[21] W.E. Winkler, *Overview of Record Linkage and Current Research Directions*, Statistical Research Report Series, No. RRS2006/02, U.S. Bureau of the Census, Statistical Research Division, Washington, D.C., 2006, and W.E. Winkler, "The quality of very large databases," *Proceedings of Quality in Official Statistics*, 2001, CD-ROM (available at `http://www.census.gov/srd/www/byyear.htmlasRR01/04`).

the quality of matches deteriorates rapidly in techniques based on the Fellegi-Sunter method. Similarly, as the size of blocks used for sorting data for matching purposes grows, so too do both the computational demands for comparing records in pairs and the probabilities of correct matches.

Low-quality record-linkage results will almost certainly increase the rates of both false positives and false negatives when merged databases are used to attempt to identify terrorists or potential terrorists. False negatives correspond to the failure of systems to detect terrorists when they are present and represent a systemic failure. False positives impinge on individual privacy. Government uses of such methods, either directly or indirectly, through the acquisition of commercial databases constructed with fusion technologies need to be based on adequate information on data quality especially as related to record-linkage methods.

3 Enhancing Privacy Through Information-System Design

Some aspects of information-system design are related to the ability to protect privacy while maintaining effectiveness, and there are many designs (and tradeoffs among those designs) for potential public policies regarding data privacy for information systems. Moreover, times and technology have changed, and a new set of policies regarding privacy and information use may be needed. To be rational in debating and choosing the policies and regulations that will provide the most appropriate combination of utility (such as security) and privacy, it is helpful to consider the generic factors that influence both. This section lists the primary components of information-system design that are related to privacy and indicates the issues that are raised in considering various options.

3.1 Data and Privacy

A number of factors substantially influence the effects of a deployed information system on privacy. Debates and regulations can benefit from differentiating systems and applications on the basis of the following:

- *Which data features are collected.* In wiretapping, recording the fact that person A telephoned person B might be less invasive than recording the conversation itself.

- *Covertness of collection.* Data may be collected covertly or with the awareness of those being monitored. For example, images of airport passengers might be collected covertly throughout the airport or with passenger awareness at the security check-in.

[22]See, for example, T.N. Herzog, F.J. Scheuren, and W.E. Winkler, *Data Quality and Record Linkage Techniques*, Springer Science and Business Media, New York, N.Y., 2007.

- *Dissemination.* Data might be collected and used only for a local application (for example, at a security checkpoint) or might be disseminated widely in a nationwide data storage facility accessible to many agencies.

- *Retention periods.* Data might be destroyed within a specified period or kept forever.

- *Use.* Data might be restricted to a particular use by policy (for example, anatomically revealing images of airport passengers might be available for the sole purpose of checking for hidden objects) or unrestricted for arbitrary future use. One policy choice of particular importance is whether the data are subject to court subpoena for arbitrary purposes or the ability to subpoena is restricted to specified purposes.

- *Audit trail.* An audit trail (showing who accessed the data and when) should be kept.

- *Control of permissions.* If data are retained, policy might specify who can grant permission for dissemination and use (for example, the collector of the data, a court, or the subject of the data).

- *Trust.* The perception of privacy violations depends heavily on the trust of the subject that the government and everyone who has access to the data will abide by the stated policy on data collection and use.

- *Analytical methods involved.* Analysis of data collected or the presentation of analytical results might be restricted by policy. For example, in searching for a weapon at a checkpoint, a scanner might generate anatomically correct images of a person's body in graphic detail. What is of interest is not those images but rather the image of a weapon, so analytical techniques that detected the presence or absence of a weapon in a particular scan could be used, and that fact (presence or absence) could be reported rather than the image itself.

3.2 Information Systems and Privacy

Chapter 2 describes a framework for assessing information-based programs. But the specifics of program's implementation make a huge difference in the extent to which it protects (or can protect) privacy. The following are some of the implementation issues that arise.

- *Does the application require access to data that explicitly identify individuals?* Applications such as searching a database for all information about a particular person clearly require access to data that are associated with individual names. Other applications, such as discovering the pattern of patient symptoms that are predictive of a particular disease, need not necessarily require that individual names be present.

- *Does the application require that individually identified data be reported to its human user, and, if so, under what conditions?* Some computer applications may require personally identified data but may not need to report personal identifications to their users. For example, a program to learn which over-the-counter drug purchases predict emergency room visits for influenza might need personally identified data of drug purchases so that it can merge them with personally identified emergency room records, but the patterns that it learns and reports to the user need not necessarily identify individuals or associate specific data with identifiable individuals. Other systems might examine many individually identified data records but report only records that match a criterion specified by a search warrant.

- *Is the search of the data driven by a particular starting point or person, or is it an indiscriminate search of the entire data set for a more general pattern?* Searches starting with a particular lead (for example, Find all people who have communicated with person A in the preceding week) differ from searches that consider all data equally (for example, Find all groups of people who have had e-mail exchanges regarding bombs). The justification for the former hinges on the justification for suspecting person A; the latter involves a different type of justification.

- *Can the data be analyzed with privacy-enhancing methods?* Technologies in existence and under development may in some cases enable discovery of general patterns and statistics from data while providing assurances that features of individual records are not divulged.

- *Does the data analysis involve integrating multiple data sources from which additional features can be inferred, and, if so, are these features inferred and reported to the user?* In some cases, it is possible to infer data features that are not explicit in the data set, especially when multiple data sets are merged. For example, it is possible in most cases to infer the names of people associated with individual medical records that contain only birthdates and ZIP codes if that data set is merged with a census database that contains names, ZIP codes, and birthdates.

4 Statistical Agency Data and Approaches

Government statistical agencies have been concerned with confidentiality protection since early in the 20th century and work very hard to "deidentify" information gathered from establishments and individuals. They have developed methods for protecting privacy. Their goals are to remove information that could be harmful to a respondent from released data and to protect the respondents from identification. As a consequence, released statistical data, even if they may be related to individuals, are highly unlikely to be linkable with any reasonable degree of precision to other databases that are of use in prevention of terrorism. That is, the nature of redaction of individually identifiable information seems to yield redacted data that are of little value for this purpose.

4.1 Confidentiality Protection and Public Data Release

Statistical agencies often promise confidentiality to their respondents regarding all data provided in connection with surveys and censuses, and, as noted above, these promises are often linked to legal statutes and provisions. But the same agencies have a mandate to report the results of their data-collection efforts to others either in summary form or in tables, reports, and public-use microdata sample (PUMS) files. PUMS files are computer-accessible files that contain records of a sample of housing units with information on the characteristics of each unit and the people in it. The data come in the form of a sample of a much larger population; as long as direct identifiers are removed and some subset of other variables "altered," there is broad agreement that sampling itself provides substantial protection. Roughly speaking, the probability of identifying an individual's record in the sample file is proportional to the probability of selection into the sample (given that it is not known whether a given individual is in the sample).[23] (In particular, if a person is not selected for the sample, the person's data are not collected and his or her privacy is protected.) It is also possible to provide privacy guarantees even in the worst case (that is, worst case over sampling).[24]

Nonetheless, many of the methods used by the agencies are ad hoc and may or may not "guarantee" privacy on their own, let alone when used with combining data from multiple databases. Nor would they satisfy the technical definitions of privacy described above. Rather, they represent an effort to balance data access with confidentiality protection—an approach that fits with technical statistical frameworks.[25] Such trade-offs may be considered informally, but there are various formal sets of tools for their quantification.[26]

Duncan and Stokes apply such an approach to the choice of "topcoding" for income, that is, truncating the income scale at some maximum value.[27] They illustrate trade-off choices for different values of topcoding in terms of risk (of reidentification through a specific form of record linkage) and utility (in terms of the inverse mean square error of

[23] See E.A.H. Elamir and C. Skinner, "Record level measures of disclosure risk for survey microdata," *Journal of Official Statistics* 22(3):525-539, 2006, and references therein.

[24] A. Evfimievski, J. Gehrke and R. Srikant, "Limiting Privacy Breaches in Privacy Preserving Data Mining," pp. 211-222 in *Proceedings of the Twenty-Second ACM SIGACT-SIGMODSIGART Symposium on Principles of Database Systems*, ACM, New York, N.Y., 2003; C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating Noise to Sensitivity of Functions in Private Data Analysis," pp. 265-284 in *3rd Theory of Cryptography Conference*, ACM, New York, N.Y., 2006.

[25] For a discussion of the approaches to trade-offs, see the various chapters in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, eds., North-Holland Publishing Company, Amsterdam, 2001.

[26] A framework is suggested in G.T. Duncan and D. Lambert, "Disclosure-limited data dissemination (with discussion)," *Journal of the American Statistical Association* 81:10-28, 1986. See additional discussion of the risk-utility trade-off by G.T. Duncan, S.E. Fienberg, R. Krishnan, R. Padman, and S.F. Roehrig, "Disclosure limitation methods and information loss for tabular data," pp. 135-166 in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, eds., North- Holland Publishing Company, Amsterdam, 2001. A full decision-theoretic framework is developed in M. Trottini and S.E. Fienberg, "Modelling user uncertainty for disclosure risk and data utility," *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems* 10(5):511-528, 2002; and M. Trottini, "A decision-theoretic approach to data disclosure problems," *Research in Official Statistics* 4(1):7-22, 2001.

estimation for the mean or a regression coefficient).

For some other approaches to agency confidentiality and data release in the European context, see Willenborg and de Waal.[28]

4.2 Record Linkage and Public Use Files

One activity that is highly developed in the context of statistical-agency data is record linkage. The original method that is still used in most approaches goes back to pioneering work by Fellegi and Sunter, who used formal probabilistic and statistical tools to decide on matches and nonmatches.[29] Inherent in the method is the need to assess accuracy of matching and error rates associated with decision rules.[30]

The same ideas are used, with refinements, by the Census Bureau to match persons in the Current Population Survey (sample size, about 60,000 households) with IRS returns. The Census Bureau and the IRS provide the data to a group that links the records to produce a set of files that contain information from both sources. The merged files are redacted, and noise is added until neither the Census Bureau nor the IRS can rematch the linked files with their original files.[31] The data are released as a form of PUMS file. Those who prepared the PUMS file have done sufficient testing to offer specific guarantees regarding the protection of individuals whose data went into the preparation of the file. This example illustrates not only the complexity of data protection associated with record linkage but the likely lack of utility of statistical-agency data for terrorism prevention, because linked files cannot be matched to individuals.

[27]G.T. Duncan and S.L. Stokes, "Disclosure risk vs. data utility: The R-U confidentiality map as applied to topcoding," Chance 3(3):16-20, 2004.

[28]L. Willenborg and T. de Waal, *Elements of Statistical Disclosure Control*, Springer-Verlag Inc., New York, N.Y., 2001.

[29]I. Fellegi and A. Sunter, "A theory for record linkage," *Journal of the American Statistical Association* 64:1183-1210, 1969.

[30]See, for example, W. Winkler, *The State of Record Linkage and Current Research Problems*, Statistical Research Report Series, No. RR99/04, U.S. Census Bureau, Washington, D.C., 1999; W.E. Winkler, "Re-identification methods for masked microdata," pp. 216-230 in *Privacy in Statistical Databases*, J. Domingo-Ferrer, ed., Springer, New York, N.Y., 2004; M. Bilenko, R. Mooney, W.W. Cohen, P. Ravikumar, and S.E. Fienberg, "Adaptive name-matching in information integration," *IEEE Intelligent Systems* 18(5):16-23, 2003.

[31]For more details, see J.J. Kim and W.E. Winkler, "Masking microdata files," pp. 114-119 in *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, Va., 1995; J.J. Kim and W.E. Winkler, *Masking Microdata Files*, Statistical Research Report Series, No. RR97-3, U.S. Bureau of the Census, Washington, D.C., 1997.