Report on DIMACS* Working Group on Privacy/Confidentiality of Health Data

Date of first working group meeting: December 10 – 12, 2003

Working Group Organizers:

Rakesh Agrawal, IBM Almaden
Larry Cox, CDC
Joe Fred Gonzalez, CDC, Chair
Harry Guess, University of North Carolina
Tomas Sander, HP Labs

Report Authors:
Hiran Subramaniam and Zhiqiang Yang
Department of Computer Science
Stevens Institute of Technology

Date of Report: December 20, 2003

*DIMACS was founded as a National Science Foundation Science and Technology Center. It is a joint project of Rutgers University, Princeton University, AT&T Labs-Research, Bell Labs, NEC Laboratories America, and Telcordia Technologies, with affiliated partners Avaya Labs, IBM Research, Microsoft Research, and HP Labs.

1 Working Group Focus

Privacy concerns are a major stumbling block to public health surveillance, in particular bioterrorism surveillance and epidemiological research. Moreover, the Health Insurance Portability and Accountability Act (HIPAA) of 2002 imposes very strict standards for rendering health information not individually identifiable. How to use large health care databases to detect medical or terrorist risks and improve health care quality while maintaining privacy and confidentiality of the data is a serious challenge. This working group explores computational techniques for ensuring that the identity of an individual contained in a released data set cannot be identified. The challenge is to produce anonymous data that is specific enough to be useful for research and analysis. It considers ways to remove direct identifiers (social security number, name, address, telephone number), and ways to aggregate, substitute, and remove information from data sets. Also of interest are questions having to do with using electronic data matching to link data elements from various sources/data sets in order to identify individuals, while maintaining privacy of others. The group investigates methods for privacy protection in field-structured data and ways to extend existing methods to large data sets, as well as systems to render textual data sufficiently anonymous. Finally, the group explores formal frameworks for disclosure control and formal protection models. Sixteen talks were presented in this working group meeting and a summary of those talks are given here.

2 Summary of Presentations

2.1 Overview of Statistical Disclosure Limitation

Speaker: Lawrence H. Cox, Office of Research and Methodology, National Center for Health Statistics, Centers for Disease Control and Prevention

Dr. Cox provided an overview of statistical disclosure limitation by defining statistical disclosure and providing methods for quantifying it. His talk showed the importance of preserving certain mathematical properties in two dimensional tables with non-negative integer entries while trying to limit disclosure. These properties rely on some statistical methods like stratified sampling, imputation, and fitting log-linear models to contingency tables.

He discussed various traditional techniques to limit disclosure such as rounding, perturbation, and cell suppression and also explained their deficiencies. While conventional rounding achieves disclosure limitation, it does not maintain the additive consistency of the one-dimensional margin totals. If the Controlled Rounding method, which is based on network optimization, was used in its place, the additivity within the table would be preserved. Also this method is optimal in the sense that the modified table is as close to the original table as possible.

Dr. Cox observed that traditional suppression patterns were inadequate since it was possible for an attacker to reconstruct one or more suppressions using row and/or column totals. He suggested that in lieu of that, complementary cell suppression be used, which would suppress additional nondisclosure cells to thwart reconstruction or narrow estimation of primary disclosure cells. He discussed several complementary cell suppression methods including heuristic complementary cell suppression and several variations of it. However, complementary cell suppression is an NP-hard problem. Also, this method resulted in tables with holes and hence thwarted statistical analysis.

The alternate method outlined by Dr. Cox was Controlled Tabular Adjustment. He illustrated with an example how this method produced fully analyzable tables, which were close to the original table both locally and globally. The method also preserved the important statistical properties of the table.

Dr. Cox also addressed the issues relevant to statistical disclosure of microdata. He discussed the different techniques available to curb disclosure in microdata. These techniques include restriction, sampling and sub-sampling the population file, data abbreviation by removing direct identifiers and salient records and by top-coding, data aggregation by collapsing data categories and micro-averaging responses as well as data fabrication by swapping / switching of data. He also discussed new approaches to address the above goal by super-sampling the data file, using contextual data, using statistical database query systems, and combining probabilistic methods of measuring disclosure risk with information based methods of measuring data utility.

Dr. Cox concluded his talk by pointing out the need for focussed research in the following areas of Statistical Disclosure Limitation: spatial models, statistical maps, statistical database query systems, and the possibility of releasing models instead of data.

2.2 Legal and Regulatory Framework in the United States and the European Union

Speaker: Oliver Johnson, Merck and Co., Inc.

In this presentation, Mr. Johnson provided an overview of the primary legal and regulatory privacy regimes impacting human-subject biomedical research. He began with a short discussion of the historical basis for privacy regulation in the United States and Europe and offered a comparison of those approaches. Mr. Johnson observed that the laws apply common principles but create significantly different administrative requirements.

In analyzing the impact of privacy principles on record based biomedical research, he provided a list of requirements under the European Union directive. These include

- Explicit consent of the data subject.
- Need to protect the vital interests of data subject when subject is not able to give consent.
- Data subject makes the data manifestly public.
- The data is required for preventive medicine, medical diagnosis, provision of care or treatment, or management of health care services, and the user is operating under rules of professional secrecy.

He also discussed the rules for international transfers of data and the exceptions to the no transfer of personal information from the European Economic Area (EEA) countries to non-EEA Countries rule.

In addressing the practical applications, he observed that governments maintained comprehensive medical databases and that most governments extracted data from these databases and made it available to researchers. He also mentioned that the data provided included dates, age, gender, race, geographic and medical information and that the governments considered them non-identifiable.

Mr. Johnson enumerated the HIPAA covered entities and defined Protected Health Information (PHI). He also discussed the research requirements on the uses and disclosures of PHIs. He identified the research exceptions allowed by HIPAA, which included limited data sets, research on decedents, and work preparatory to research. He concluded with the remark that HIPAA provides new rules but reasonably practical mechanisms for record based biomedical research.

2.3 The Health Insurance Portability and Accountability Act (HIPAA) and its Implications on Epidemiological Research Using Large Databases

Speaker: Dr. K. Arnold Chan, Harvard University

With the advance in information technology and the involvement of third party insurance in medical care delivery systems, large administrative databases in health care have been used around the world to address important public health questions. Unlike clinical trials and prospective observational studies, it is not feasible to obtain individual consent or authorization for studies in which these health care data are utilized. Under HIPAA regulations in the United States, investigators can access this information without individual authorization if the Institution Review Board or the Privacy Board grants a waiver of patient authorization. In order to obtain such waivers, investigators need to follow the "Minimal Necessary Principle" during data development, implement data transformation strategies to de-identify selected data elements, and have robust data systems to safeguard Protected Information.

In this talk, Dr.Chan talked about the use of large linked automated data for public health research. He showed what kind of data might be used for health research and what privacy and confidentiality problems might arise when those data are used. Then he gave the data development processes to ensure HIPAA-compliance. Several methods of data transformation were given, e.g. randomly generate study IDs to replace true IDs, roll-up or transform variables, and extract only the information relevant to the study. One example is given in Finkelstein et al., *Decreasing Antibiotic Use Among US Children: The Impact of Changing Diagnosis Patterns,* **Pediatrics**, 2003, 112: 620-627. This pediatric antibiotics use study was presented to illustrate how to transform data so as to get highly processed and de-identified data available for concatenation across study sites and complex analysis. Finally Dr. Chan gave some keys to protecting human subject information based on his own experience.

2.4 Health Care Databases under HIPAA: Statistical Approaches to De-identification of Protected Health Information

Speaker: Judith Beach, Quintiles Transnational

Dr. Beach addressed the evolution of de-identification standards, the HIPAA Privacy regulations, and the de-identification standards for health information in research. She discussed two specific methods for de-identification, namely, the Safe Harbor method and the Statistician Method. The Safe harbor method is intended to provide a simple and definitive method for de-identifying health information with protection from litigation. In this method, the covered entities must remove all of a list of 18 enumerated identifiers (direct and indirect) and have no actual knowledge that could be used alone or in combination to identify a subject. The Safe Harbor method tolerates some amount of disclosure such as geographic divisions no smaller than a state, age if it is less than 90, or geographic unit formed by combining all zip codes with the same initial three digits containing more than 20,000 people. Since this method was not developed for research but as an approved method of de-identification for any purpose by any entity, some researchers have complained that the Safe Harbor method renders data almost useless.

The Statistician Method retains some of the Safe Harbor's specified identifiers and demonstrates that the standard is met if a person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods, e.g., a biostatistician, documents that the risk of re-identification is very small. But in general the Statistician Method is considered too complicated and the Safe Harbor method is preferred. Dr. Beach also compared the two methods and enumerated the fields and factors to be considered by the statistician to render the statistical likelihood of re-identification comparable to the Safe Harbor method. Dr. Beach also discussed an alternate method called Limited Data Set. In this method, a limited data use agreement should be in place between the covered entity and the recipient of the limited data set. This method removes 16 direct identifiers but retains indirect identifiers such as 5 digit zip code, dates of service, dates of birth and death, and geographic subdivision. This method may be useful for records based research such as epidemiological research but not useful for patient recruitment. The talk also outlined privacy cases and controversies with de-identified health databases in the United Kingdom and the United States. Dr. Beach concluded by pointing out that the key to safeguarding protected health information is to encourage the use of federal standards for de-identification of health data for clinical research.

2.5 Protecting the Privacy of Healthcare Data While Preserving the Utility of Geographic Location Information for Epidemiologic Research

Speaker: Daniel Barth-Jones,Wayne State University, School of Medicine

Epidemiologic and healthcare systems research conducted with administrative healthcare data has demonstrated considerable utility and value for the healthcare system in the United States, which has resulted in a well-developed healthcare information industry utilizing such data. The recent implementation of the HIPAA Privacy Standards, however, has necessitated dramatic changes in the process of conducting research with administrative data. Under the privacy standards, conducting research with statistically de-identified administrative data is an attractive option because such data can be used without restrictions.

In this talk, Dr. Barth-Jones presented a framework for conducting disclosure risk analysis for administrative data that considers the real-world complications involved in data intrusion attempts through record linkage methods. Disclosure risk analysis was reported focusing on three variables commonly found in administrative data:

1. date of birth/age categorization,
2. gender,
3. geographic location detail.

Disclosure risks were examined as a result of population density and the cross-classification structure of the demographic variables. Finally, based on the results of his analysis Dr. Barth-Jones concluded that considerable disclosure control can be achieved with simple modifications of administrative data sets while preserving important geographic location detail.

2.6 Privacy Technologies and Challenges in their Deployment

Speaker: Tomas Sander, HP Labs

The research community has developed a variety of privacy-enhancing technologies over the last two decades. Unfortunately only a few of these technologies have been successfully deployed. This talk reviewed several of these technological approaches, what they accomplish, and difficulties in deploying them.

Dr. Sander first reviewed several approaches of privacy technology and he introduced the privacy-preserving ID3 tree algorithm for distributed data mining. He pointed out that so far there are still few companies that deploy those privacy technologies. The reason for this is that implementing those technologies involves high overhead and the commercial profit from deploying them is not attractive. He concluded that the biggest motivation for deploying them is legislation. If strong privacy legislation is enacted, industry will put more money into developing and deploying privacy technologies in their products.

2.7 Software Demonstration of the use of Hippocratic Database Technology in Supporting a Health Care Provider

Speaker: Tyrone Grandison, IBM

Dr. Grandison provided a broad overview of the Hippocratic Database project - an initiative taken by IBM to develop a new comprehensive privacy management solution that supports automatic enforcement of privacy policies. Hippocratic database systems take responsibility for the privacy of the data they manage while not impeding the flow of information.

The founding tenets of the Hippocratic Database are as follows:

- Information that is collected must be limited to the minimum necessary for accomplishing the specific purpose.
- Each piece of personal information in the database is associated with the purpose for which it was collected and the consent of the donor of the information.
- The database shall only run those queries that are consistent with the purposes for which the query has been collected. Also the personal information shall not be communicated outside the database.
- The information in the database should be accurate and should be retained only as long as it is required.
- The donor will be able to verify the information and can also check the conformance with the aforementioned rules at any time.

Dr. Grandison discussed the architecture of the Hippocratic Database that incorporated the modules implemented to enforce the above principles. The privacy policy of the company is coded into the meta data and the queries to the database are analyzed with respect to the policy. Based on the analysis, the query can be allowed to run as is or be blocked (if it violates the purpose) or be allowed to return a subset of records that reflect the individuals' opt-in or opt-out preferences.

He also demonstrated with an example how the Hippocratic Database enforced the policies of a sample health care provider. His talk concluded with the comparison of the performance of the Hippocratic Database with existing database technologies. Despite the fact that modern database technology has been tuned to optimal performance over several years, the new Hippocratic Database (with its added functionality) provide almost minimal overhead thus offering comparable performance.

Open questions that lie in this direction of research include how to incorporate legacy data into the Hippocratic Database? The question is more interesting when considering deceased donors and the opt-in and opt-out preferences in the Hippocratic Database.

2.8 Cryptographic Techniques for Confidentiality of Aggregate Statistics on Health Data

Speaker: Giovanni Di Crescenzo, Telecordia

In discussing relationships between cryptography and health care, Dr. Di Crescenzo argued that the latter area is finally approaching mature times for enhancements that use results from the former. Even more, he argued that cryptography has already produced secure systems that have quick applicability to health care. He exemplified this state of affairs by showing that his previous results on privacy for stock market operations, after minor further analysis and modified design, naturally applies to solving the following privacy problem in health care statistics: allow the collection and statistical analysis of data from medical records while keeping such records private both from other record holders and from the data collector itself. A new and efficient zero-knowledge protocol for proving sum-related statements about encrypted values was presented as one technology that can be applied to aggregate statistics on health data.

2.9 Tutorial on Data Mining

Speaker: David Madigan, Rutgers University

In his tutorial, Dr. Madigan provided a comprehensive overview of current research in data mining. He introduced the definitions of data mining and a data mining algorithm. Data mining is the process of finding interesting statistical patterns, predicative models, and interesting relationships. A data mining algorithm is a well-defined procedure that takes data as input and produces output in the form of patterns or models.

He discussed some sample data mining algorithms. In the back propagation algorithm, the input vector of values is multiplied by a weight matrix and the resultant vector is subject to a linear transformation. The result then is multiplied by another weight matrix. The back propagation algorithm was defined in terms of the weight parameters, score function, and the search criteria.

Dr. Madigan also discussed the classification of models into Predictions (including linear, non parametric, piecewise linear regressions), Probability distributions (such as mixtures of parametric models, graphical markov models), and structured data (including time series and spatial data). He pointed out how the score function must strike a balance between bias and variance. He also discussed the increase in the level of difficulty with the increase in the number of dimensions.

Dr. Madigan discussed algorithms for finding local and global patterns by permutation tests. With an example he illustrated how unusual patterns can be identified in a given data set with fixed and variable pattern length (window length). He also addressed variations of scan statistics such as spatial scan statistics that use circles instead of line segments for delimiting a search area and spatial-temporal statistics that use cylinders for the same purpose.

Dr. Madigan also addressed association rule mining, which is considered to be the origin of data mining. Association rule mining involves finding all the rules $Y \Rightarrow Z$ with confidence and support above given minimums, where support is the probability that a transaction contains both Y and Z and confidence is the conditional probability that a transaction containing Y also contains Z.

He discussed an a priori algorithm for generating frequent sets (sets having at least minimum support) and using the frequent sets to generate association rules. He concluded by pointing out potential areas of research in designing algorithms for finding infrequent data sets and for creating a knowledge base that would automatically catalog the association rules mined out.

2.10 Using Data Mining Techniques to Harvest Information from Clinical Trials

Speaker: Richard D. De Veaux, Williams College, Williamstown, MA.

Dr. De Veaux presented a data mining case study based on health data. The objective of this case was to determine early predictors of study dropout by mining the data from 692 depressed patients in two eight-week clinical trials. Patients were randomized to one of three after meeting entrance criteria and then treated for eight weeks. Clinical visits took place at baseline and days 7, 14, 21, 28, 42, and 56. Depression was measured via investigator-rated Hamilton Rating Scale for Depression (HAM-D). Other clinical measures included Hamilton Anxiety Scale (HAM-A), indicators of sexual dysfunction, and other adverse events. In those data, overall study dropout rate was 31%. Data available up to and including day 14 were mined in an attempt to determine early (within the first two weeks) predictors of eventual study dropout. A number of data mining techniques were applied to the data. The single greatest predictor of eventual dropout was the presence or absence of readings at day 14. Patient age also was relevant. Knowledge of such early warning signs could possibly improve patient retention and study quality. The conclusion of this case study was that signs of study dropout may be evident very early in clinical trials and every effort should be made to maintain enrollment of those patients who show early signs of eventual dropout.

2.11 Experimental Results on Privacy-Preserving Statistics Computation

Speaker: Rebecca Wright, Stevens Institute of Technology

In her presentation, Dr. Wright addressed the criticality of preserving privacy while allowing multiple parties to securely compute some function of their inputs without revealing them. She observed that this scenario was common in security related settings where two parties may wish to compute a common list of suspects or in a commercial setting where companies make extensive use of third party databases to compute some market statistics or querying for records in epidemiological research. She discussed the advantages of such private computations, which include protection of personal, proprietary, and sensitive information, enabling collaboration between different data owners, and compliance with legislation.

The concept of privacy preserving data mining enables analysis of data from diverse sources without requiring original data to be gathered in one place. Dr. Wright demonstrated how using cryptography can provide a construction that does not reveal any information other than the output of the computation. Her construction showed how statistics such as mean, median, counting frequencies, and variance could be computed in such a way that:

- The client learns the desired statistics and nothing else (including individual values or partial computations).
- The servers do not learn the fields that are queried or any other server's data.
- The computation and communication are very efficient.

Dr. Wright showed a perfectly secure construction to compute the sum of a subset of values held by the server. This construction makes use of a homomorphic encryption scheme, the Paillier encryption scheme. The client encrypts the indices of the elements in which it is interested and sends it to the server and the server performs a blind computation on the indices and its own values in such a way that it learns neither the indices nor the computed sum. The server then sends the encrypted sum to the client which decrypts it using its key.

This elegant protocol that conforms to the strongest cryptographic requirements of privacy suffers from the burden of encryption. Since the client has to completely blind the server, it should send encrypted indices for each element in the database. This places heavy processing burden on the client as evinced by the implementation results. However, if the encryptions were done ahead of time and stored in a database, the construction becomes lightweight and efficiently addresses the problem.

Dr. Wright concluded her talk by pointing out the demerits of the cryptographic approach in terms of the computational burden of encryption and the need to construct a specific solution for each problem. However, the cryptographic approach is attractive since it does not leak any information beyond what is required and can be easily extended to provide resistance to malicious adversaries or semi-honest participants.

2.12 Semantic Web Services for Privacy/confidentiality of Health Care Data

Speaker: Nabil Adam, Rutgers University

Information systems technology allows instant retrieval of medical information, widening access to a greater number of people. Computerization of medical records has also threatened patient privacy and, in particular, has increased the potential for misuse, especially in the form of non-consensual secondary use of personally identifiable records. The most fundamental principle of fair use of information is that no secondary use of medical information should take place unless authorized by the patient. This presents a challenge for ensuring privacy and confidentiality protection while providing authorized users with the convenience of e-Healthcare.

Dr. Adam presented the investigation of the markup of web services with a semantic policy language as an alternative to traditional authentication and access control methods. Authentication Semantic Web Services and Authorization Semantic Web Services were discussed but this talk focused on the latter. The presented approach is to investigate the markup of web services with a semantic policy language (alternative to traditional authentication and access control). Rule based access control is defined to deal with multi-user and multi-application and is properly viewed as a semantic construct around which access control policy is formulated. In Dr. Adam's research, rules are formulated by XML and that will reduce human error and policy conflicts and facilitate portability. Finally Dr. Adam concluded that these kind of services present a more efficient and flexible management capability for privacy and confidentiality related issues applicable to e-Healthcare and further provide support for complex problem solving, knowledge modelling, and reuse.

2.13 Privacy/Confidentiality Issues in Collecting Agricultural Data

Speaker: Gary Smith, University of Pennsylvania

Modern spatial models of infectious disease epidemics in domestic animals are becoming increasingly influential in informing policy decisions about disease control. Such models depend upon having accurate information concerning the location of farms, what species of animals are raised on each farm, and how many of each species are present. The exemplars for this kind of modelling are the foot and mouth disease models that were so influential during the 2001 foot and mouth disease epidemic in Britain. It seems unlikely that we shall ever be able to apply similar models in the United States. The reason for this is that farmers are often very reluctant to provide information that may eventually find its way into the hands of local, state, or federal government agencies and thus be rendered accessible to the public at large.

2.14 Private Analysis of Data Sets

Speaker: Benny Pinkas, HP Labs, NJ

Dr. Pinkas discussed methods of private analysis of data sets. There exist two parties, each holding a secret data set but desiring to compute a function F of the two data sets without revealing their inputs. Existing solutions to such secure multiparty computation offer polynomial overhead. Dr. Pinkas showed how to reduce this overhead to a linear or sublinear level and illustrated it with two constructions.

1. Computing intersection of two sets.
2. Computing the k^{th} ranked item of the union of two sets.

He also showed that these constructions were secure against both a semi-honest client and a malicious adversary.

His constructions made use of homomorphic encryption schemes that enable one to compute the encryption of a new plaintext from the knowledge of the encryptions of 2 different plaintexts. Mathematically, the properties of such encryption schemes can be stated as

$$E(m_1) \times E(m_2) = E(m_1 + m_2)$$
$$(E(m_1))^c = E(cm_1)$$

where m_1, m_2, c are all integers.

In his construction for computing the intersection of two sets, the client defines a polynomial of degree n (where n is the number of elements in the client's set) whose roots are the elements in the client's data set. The client then sends the encryptions of the polynomial coefficients to the server. The server uses random masking in such a way that only the values that are common to the client's and server's data sets will decrypt to meaningful values and the reset will decrypt to random values. This is sent to the client who can now decrypt and learn the intersection. The efficiency of the above scheme is polynomial and can be improved to sub-polynomial time also.

In the construction of the k^{th} element of the union of two sets, each party computes the median of the data set it holds and sends it to the other. If the median of party A is less than median of party B then party A eliminates all the elements less than the median from its data set. The parties then compute the medians of their modified data sets. The above procedure is repeated until the two sets are of size k. The median of the data sets then gives the k^{th} element of the union of the two sets.

Dr. Pinkas concluded by listing the open problems in this area of research, which include solving intersection problems with approximate matching and solving the median problem with clustering.

2.15 Overview of Masking Schemes for Microdata

Speaker: Jay J. Kim, Office of Research and Methodology, National Center for Health Statistics, Centers for Disease Control and Prevention

The U.S. Department of Health and Human Services (DHHS) has issued new national health information privacy standards. This is in response to the mandate of the Health Insurance Portability and Accountability Act (HIPAA) of 1996. The new standards provide protection for the privacy of certain individually identifiable health data.

This talk reviewed the existing procedures for masking microdata. Dr. Kim gave the masking schemes for discrete variables, including both those for dichotomous and polychotomous variables. Data swapping, coding approaches, and sampling without replacement are used for dichotomous variables. Data swapping, combination of categories, and coding approaches for polychotomous variables were also presented. Several different models were mentioned by Dr.Kim, e.g. the randomized response model, Warner's two-fold model, and Warner's contamination model. Masking schemes for the continuous variables were given including (1) additive noise, (2) multiplicative noise, (3) rounding, (4) micro aggregation, (5) interval data, (6) data swapping, and (7) suppression and generalization. The strengths and weaknesses of each scheme were discussed by Dr.Kim. The statistical properties of the masked data and recoverability of the original data were discussed, e.g. the mean, variance, and covariance.

2.16 Statistical Disclosure in Tabular Data and Related Mathematical and Computational Problems

Speaker: Lawrence H. Cox, Office of Research and Methodology, National Center for Health Statistics, Centers for Disease Control and Prevention

Dr. Cox addressed the new method for Statistical Disclosure Limitation in tabular data, namely, Controlled Tabular Adjustment (CTA). He illustrated with an example the step by step procedure of the CTA method. He clearly showed how the CTA method could be instructed to minimize the total values of adjustments, the percentage absolute adjustment, the number of cells changed, and logarithmic functions of absolute adjustments.

He also stated that the adjustments to non-sensitive cells could be restricted to be within measurement error. He explained that, using Linear Programming, the key statistics and linear models, in terms of mean, variance, correlation and regression slope, were preserved between the original and the adjusted data by using CTA.

He also pointed out that each principal SDL method and the associated problems were constrained by a linear system of the form TX = B, where T is the aggregation matrix, B is the vector of marginal totals, and X may be a vector of integer (count data). The above linear system is solved by the original data.

He showed that the above problem was simple for 2-dimensional tables but the solution could not be efficiently generalized to higher dimensions. He showed with an example how extending this model to 3 dimensions would lead to a non-integer polytope extreme point and other problems. He concluded by emphasizing the necessity of having efficient solutions extensible to higher dimensions, which is still an open problem.

3 Future Research Challenges and Future Plans

The working group developed a variety of ideas at this meeting that will lead to future investigations. A key set of challenges arises for teams involving cryptographers and epidemiologists. A meeting to explore these issues is currently being planned. A second major challenge falls in the area of data de-identification and the role of combinatorial optimization in this field. The working group plans a meeting at which statisticians, epidemiologists, and combinatorial optimizers all discuss the issues and lay out a research agenda. Additional challenges lie in identifying specific guidelines for statisticians in certifying HIPAA compliance. The working group will be organizing a tutorial on this topic.

Challenges at the interface between cryptography and epidemiology/health data analysis are given below. Future meetings will produce similar lists for the interface between data de-identification and combinatorial optimization and for the interface between HIPAA compliance and statistics.

1. Different Functionalities and Specific Challenges for Cryptography.
   (a) Does transferring data between a hospital and testing lab or other problems of transferring health data require any different cryptographic tools than we need for financial transactions?
   (b) We should distinguish between problems of transferring data and problems of computing with data, especially distributed data. See 2 for challenges in this direction.
   (c) How do we improve the performance of cryptographic schemes (secure multiparty computation) to make them affordable for practical applications?
   (d) How do we prove compliance, cryptographically, with a stated privacy policy?
2. Privacy-preserving Data Mining and Privacy-preserving Data Sharing.
   (a) Identify specific functionalities needed for health data applications.
   (b) Make secure multi-party computation more efficient for large databases (a generic challenge).
   (c) Extend secure multi-party computation to clustering. Since clustering is hard, we might have to settle for approximate solutions. More generally, can we extend secure multi-party approximation?
   (d) Is it possible to modify secure multi-party computation protocols so one doesn't have to access all data elements?
   (e) What are the issues involved in privacy-preserving data sharing in general and secure multiparty computation in particular if we want to take into consideration what the output itself might leak about the data?
3. Tracking Disclosed Information (a topic related to secure software and secure computing environments as well as cryptography)

   (a) Can we "send" with disclosed information some restrictions on its use, e.g., future disclosure?
   (b) Can we "send" with disclosed information restrictions on the length of time it can be saved/used?
   (c) Can we do this tracking if there are later changes in disclosure limitations?

4. Can we develop good auditing technologies?

   This question applies well beyond cryptography. In health data, it is concerned with distinguishing between a transaction (e.g., looking at a patient record) that is legitimate and one that is not. A well-known method involves tracking authorizations. However, are there smart methods to audit large data sets of transactions to find illegitimate transactions?

5. "Customizable" Privacy

   Software employed by different partners may differ in privacy protections/policy and processing. This presents cryptography with complex privacy management concerns and it would be important to develop privacy protocols that are readily "customizable" to different users. How do we achieve customized privacy that would satisfy/balance the privacy policies of all participants?

6. Dynamic Query Authorization and Forbidden Question Combinations

   (a) It is an old topic to change query authorization based on previous queries so as to make it impossible to make forbidden inferences. But how do we do this in the encryption situation and with widely distributed data sets?
   (b) A simpler challenge arises if we have specific questions and some combination of them that is forbidden in advance. Even here, there are cryptographic challenges if we hide the questions from the database owner.

7. Revealing Partial Information

   It may not be known in advance which information will and will not be sensitive. Traditionally, cryptography does not allow information leakage unless it is explicitly defined as part of the input. Dynamically-changing disclosure limitations pose challenges for cryptography, e.g., in secure multiparty computation.

8. Cleaning Data and Maintaining Privacy

   Data preparation and cleaning is a major part of real life statistics. Can this be done in a privacy enhanced way?

4 Acknowledgements

The author and the DIMACS Center acknowledge the support of the National Science Foundation under grant number CCR 03-14161 to Rutgers University.