# Report on DIMACS/PORTIA* Workshop and Working Group on Privacy-Preserving Data Mining

Workshop March 15 – 16, 2004
Working group March 17, 2004

Organizers

Cynthia Dwork, Microsoft
Benny Pinkas, HP Labs
Rebecca Wright, Stevens Institute of Technology

Report Authors

Geetha Jagannathan, Department of Computer Science
Stevens Institute of Technology
Hong Jiang, Department of Computer Science
Yale University

January 18, 2005

*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. PORTIA is supported by the National Science Foundation through its Information Technology Research program.

# 1 Introduction

The DIMACS/PORTIA workshop and working group on privacy-preserving data mining brought together researchers and practitioners in cryptography, data mining, and other areas to discuss privacy-preserving data mining. The workshop sessions on March 15 and 16, 2004 consisted of invited talks and discussions. March 17, 2004 was a working group meeting of invited participants to identify and explore approaches that could serve as the basis for more sophisticated algorithms and implementations than presently exist, and to discuss directions for further research and collaboration. The workshop was attended by 88 people from 3 countries, including 32 students. The working group meeting was attended by 52 people from 3 countries, including 10 students.

Both the workshop and the working group meeting investigated the construction and exploitation of "private" databases, e.g.

- Merging information from multiple data sets in a consistent, secure, efficient and privacy-preserving manner;
- Sanitizing databases to permit privacy-preserving public study.

In a wide variety of applications it is useful to be able to gather information from several different data sets. The owners of these data sets may not be willing, or legally able, to share their complete data with each other. The ability to collaborate without revealing information could be instrumental in fostering inter-agency collaboration.

Specific topics discussed in the workshop and working group include:

- Secure multi-party computation: This is a very general and well-studied paradigm that unfortunately has not been used in practice so far.
- Statistical techniques such as data swapping, post-randomization, and perturbation.
- Articulation of different notions and aspects of privacy.
- Tradeoffs between privacy, accuracy and efficiency.
- Architectures that facilitate private queries by a (semi-trusted) third party.

- Methods for handling different or incompatible formats, and erroneous data.
- Additional issues such as ensuring the accuracy and reliability of responses, query authentication, logging, auditing, access control and authorization policies.

We summarize the presentations of the workshop in Sections 2 and 3. The discussions of the working group are summarized in Section 4.

# 2 Workshop (March 15)

## 2.1 From Idiosyncratic to Stereotypical Toward Privacy in Public Databases

Speaker: Shuchi Chawla, CMU

This talk, which described ongoing work with Cynthia Dwork, Frank McSherry, Adam Smith, Larry Stockmeyer and Hoeteck Wee, addressed the problem of modifying/sanitizing high-dimensional data, such as census data, in such a way that the sanitized data can be published without violating the privacy of the original data, and yet retain its utility. Their privacy definition is based on the intuition that the privacy of an individual is preserved to the extent that the individual blends in with the crowd. One well-known statistical approach to privacy preservation is to alter the frequency of particular features, while preserving means. Additionally, one could erase values that reveal too much. Query-based approaches often involve query monitoring to dissallow queries that breach privacy, or use perturbation to add noise to the query output.

In the geometric view assumed by this work, the real database is assumed to be a set of $n$ points in some high-dimensional space. The distance between these points is considered crucial. The idea is to transform this real database into a sanitized database, which is a set $n'$ of new points, perhaps in an entirely new space. In such a setting, an adversary is considered to be successful in isolating a point $x$ if it is able to output a point $q$ that is much closer to $x$ than $x$'s neighbors. This is considered a breach in privacy.

The key idea behind the sanitization method is to randomly perturb each point $x$ in proportion to its $T$-radius. (The $T$-radius of a point is the distance from $x$ to its $T$-nearest neighbor.)

The talk was concluded with some future directions in this area, including:

1. extending the privacy argument to other nice distributions,

2. characterizing acceptable auxiliary information,

3. finding if these results can be extended to low-dimensional spaces, and

4. extending the utility argument to other interesting macroscopic properties, such as correlation.

## 2.2 Privacy-Preserving Data Mining on Vertically Partitioned Databases

Speaker: Kobbi Nissim, Microsoft Research

The talk began with Dr. Nissim's definition of privacy. The presentation addressed privacy issues relating to the user of the perturbation approach to answering queries in statistical databases.

In previous work by Dinur and Nissim, the authors had considered a statistical database in which a trusted database administrator monitors queries and introduces noise to the responses with the goal of maintaining data privacy. They proved that noise of magnitude at least $\sqrt{n}$ is required to avoid privacy breaches by a polynomial-time adversary making, roughly, a linear number of queries. As databases grow increasingly large, the possibility of being able to issue only a sub-linear number of queries becomes realistic. In such cases it might be possible to exploit a sub-linear bound on the adversary. This was first explored by Dinur, Dwork, and Nissim who demonstrated a database mechanism that guarantees privacy with noise magnitude less than $\sqrt{n}$.

This prior work can be generalized in two directions:

1. Defining and analyzing privacy for multi-attribute databases. With respect to this aspect, their privacy requirements dealt not only with single attributes but also with functions of attributes.

2. The usability of their database mechanism for data mining of vertically partitioned databases.

Dr. Nissim concluded his talk by briefly indicating a few open problems, including:

- Improving the privacy definition to cover everything a realistic adversary will do.

- Improving usability and efficiency, such as finding an alternate way to perturb and use the data that would result in more efficient and accurate data mining and for data mining published statistics.
- Data mining 3-ary Boolean functions from single attribute sublinear query databases (SuLQ DBs).
- Obtaining strong privacy definition and rigorous privacy proof in SuLQ DB.
- Extending the Dinur, Dwork and Nissims observation that privacy may be preserved in large databases.
- Obtaining usability for the data miner for both single and vertically split databases.

### 2.3 Confidentiality in Tables Viewed from an Algebraic Perspective

Speaker: Lawrence Cox, CDC

This talk was about the methods to preserve the confidentiality of individually identifiable data collected from individual respondents that are released by organizations like the National Statistical Offices. It addressed the issue of confidentiality protection when viewed from an algebraic perspective. This relates to the algebraic problem of determining if a given element of a ring is an ideal, given the ring $R$ and an ideal $I$ of $R$. Traditional methods for confidentiality protection in statistical tabulations (tables) of nonnegative counts include:

- rounding—replacing a value in the table with its nearest multiple of the rounding base,
- cell suppression—hiding those values in the table than can be inferred from other values, and
- perturbation—changing the value of an entry using randomization.

Each of these conventional methods either leaks information, or introduces undesirable artifacts in the data. However, variants of these methods can be more effective in preserving privacy. Controlled rounding, which rounds to the adjacent multiple of the base (instead of rounding to the nearest multiple

of the base) introduces fewer distortions of the data. Controlled perturbation is similar to controlled rounding. Complementary cell suppression hides additional data cells (beyond the ones that would be disclosed). However, linear analysis can still reveal exact bounds of some cells and a careful approach is needed to prevent disclosure of cell values. Complementary cell suppression is an NP-hard problem, and the latter three methods are similarly complicated by the need to preserve tabular structure (additivity of item detail to lowest-level totals, of lowest totals to next-level totals, etc.), resulting in challenging mathematical and computational problems. All of these methods relate to the problem of computing moves between neighboring solutions to a corresponding integer linear programming problem.

### 2.4 Privacy Preserving Computation of the $k^{\text{th}}$ Ranked Element

Speaker: Gagan Aggarwal, Stanford University

This talk was concerned with the problem of computing the $k^{\text{th}}$ ranked element of a large confidential dataset shared by two or more parties. It was based on joint work with Nina Mishra and Benny Pinkas. When privacy needs to be preserved in a computation where data is distributed among multiple participants, there should ideally be a trusted third party who can compute and send the results back to the participants, without revealing any additional information. However, since such trusted third parties are rare, there is the need for secure protocols. Such protocols should guard against one of two kinds of adversaries: semi-honest participants, and malicious participants. Ms. Aggarwal presented a protocol for the secure computation of the $k^{\text{th}}$ ranked element. The problem can be described as follows: Assume that each of two parties $A$ and $B$ has a data set ($S_A$ and $S_B$, respectively). The protocol should output an $x \in S_A \cup S_B$ such that $x$ has $k-1$ elements smaller than it. The protocol runs in $O(\log k)$ rounds, where each round involves a communication cost of $O(\log M)$ (the number of bits needed to describe each element of the input data). When $k = n/2$, the problem amounts to finding the median of $S_A \cup S_B$. The protocol was motivated by presenting an insecure two-party protocol for computing the median. A key step in this insecure protocol involves a comparison of two values, one from each of the parties. A secure protocol for computing the median is achieved by simply replacing the (insecure) comparison with a secure comparison (which can be performed efficiently because it involves a small circuit). This protocol is secure against semi-honest adversaries,

but not against malicious adversaries. To ensure that the protocol is also secure against malicious adversaries, the parties are required to maintain additional state information so as to assure that inputs to comparisons are consistent across rounds. For the multi-party case (more than two parties), the authors assume that all values held by all parties lie in some range $[a, b]$ which is known in advance. In this situation, the basic idea is to perform a secure binary search on $[a, b]$. The resulting protocol can also be made secure against malicious adversaries using consistency checking.

### 2.5 Efficient Private Matching and Set Intersection

Speaker: Michael Freedman, NYU

This talk was based on joint work with Kobi Nissim and Benny Pinkas. The problem here is the computation of the intersection of private datasets held by two parties, where the datasets contain lists of elements taken from a large domain. There are a large number of potential applications for such a list intersection protocol. The problem can be solved using a number of general purpose methods, including Yao's protocol, and by using $O(n^2)$ private equality tests. However, these existing techniques are inefficient. The new protocol presented is efficient, requiring $\mathrm{O}(n \log\log n)$ communication. It is based on on the use of homomorphic encryption. In such an encryption scheme it is easy to compute the encryption of $M_1 + M_2$ given the encryption of $M_1$ and the encryption of $M_2$ (without knowing the private key). The specific homomorphic encryption used is a modified form of the El-Gamal encryption scheme. The protocol assumes a client-server environment where the client learns the intersection of the sets held by itself and by the server, while the server learns nothing at all. The client computes the coefficients of a polynomial based on its list, and sends the encryption of these coefficients to the server. The server "evaluates" this polynomial on each of its values, perturbs the results by multiplying them with random values, and then sends them to the client. The client can then infer from these results which of its elements belong in the intersection. The protocol can be sped up using Horner's rule (if the list values are from a small domain), and by using hashing. The protocol for the semi-honest environment is secure in the standard model, while for the malicious environment it is secure in the random oracle model.

The related problem of approximating the size of the intersection has a linear lower-bound for the communication overhead. The speaker presented a protocol for the problem that matches this lower bound. This protocol is

based on sampling.

Mr. Freedman concluded his talk by presenting some open problems such as finding a more computationally efficient protocol, and, in fuzzy matching, obtaining protocols which are secure against malicious adversaries.

### 2.6 An Experimental Study of Association Rule Hiding Techniques

Speaker: Emmanuel Pontikakis, University of Patras

In this presentation, which was based on joint work with Vassilios Verykios, Mr. Pontikakis presented and compared two algorithms for hiding association rules. Association rules are defined on databases of transactions, where each transaction is a subset of a universal set of items. Each association rule in as implication of the form $X \Rightarrow Y$, where $X$ and $Y$ are subsets of items. A rule of of the form $X \Rightarrow Y$ is said to hold with confidence $c$ if $|X \cup Y|/|X| > c$, where $|Z|$ denotes the number of transactions that contain the set $Z$. Also, $X \Rightarrow Y$ has support $s$ if $|X \cup Y|/N > s$, where $N$ is the total number of transactions in the database. Some association rules might be considered "sensitive", and the publisher of a database might not want to reveal them. The techniques described in this talk seek to modify databases so as to hide such sensitive association rules. The first general technique is called the distortion-based technique. The idea here to change specific values in a database so that mining the (modified) database no longer yields sensitive rules. However, this technique could lead to undesirable side-effects, such as rule elimination, ghost rules, and itemset elimination. The challenge is to create techniques so as to minimize the side effects, minimize the number of changes to be made, do it all in time linear in the size of the database.

The speaker presented an algorithm called the weight-based sorting distortion algorithm (WSDA). The idea here is to prioritize transactions that support a given rule, based on how much these transactions support other strong rules. The transactions with the lowest priorities are modified (by removing items from them) to hide the given association rule. Experimental results show that fewer rules are changed by the WSDA algorithm in comparison with the other well-known distortion algorithm called 1.$b$.

Since it could be dangerous to delete items in certain transactions (such as in medical databases), another way of hiding association rules is by adding uncertainty without distorting the databases. Blocking-based techniques for association rule hiding work by indicating the absence of information regarding certains items in certain transactions. Mr. Pontikakis presented a high

level description of their blocking algorithm called CRA and experimentally compared it with the straightfoward blocking algorithm called BA.

The talk concluded with the presentation of various open problems including:

- What techniques must be used in order to reduce the privacy breaches?
- In what other ways can we prevent an adversary from inferring the association rules in the database?

He also suggested that applying a $\chi$-square test to the final database may reveal some correlations between the items.

### 2.7 Public-Key Encryption with Keyword Search

Speaker: Giovanni DiCrescenzo, Telcordia Technologies

In this talk, which was based on a paper written in collaboration with Dan Boneh, Rafail Ostrovsky, and Giuseppe Persiano, Dr. DiCrescenzo spoke about the problem of searching on data that is encrypted using a public-key system. The problem can be motivated using the following scenario: Consider user Bob who sends email to user Alice encrypted under Alice's public key. An email gateway wants to test whether the email contains the keyword "urgent" so that it could route the email accordingly. Alice, on the other hand does not wish to give the gateway the ability to decrypt all her messages. The task is to define and construct a mechanism that enables Alice to provide a key to the gateway that enables the gateway to test whether the word urgent is a keyword in the email without learning anything else about the email.

Dr. DiCrescenzo defined the concept of a secure Searchable Public Key Encryption (SPKE) scheme and spoke about its relation to Identity Based Encryption (IBE). Three constructions for public-key searchable encryption were presented, namely:

1. an efficient system based on a variant of the Decision Diffie-Hellman assumption.
2. a somewhat less efficient system using elements modulo a composite.
3. a system based on general trapdoor permutations.

He concluded the talk by indicating that searchable public key encryption can be used for processing encrypted email and other public-key encrypted data.

## 2.8 Privacy-Enhanced Searches Using Encrypted Bloom Filters

Speaker: Steve Bellovin, AT & T Research

Dr. Bellovin's talk, based on joint work with William Cheswick, was about how to authorize searches when one organization wants to search for documents in another organization. The security requirements for sharing data between two parties (one called the queries, the other the provider) who do not trust each other include:

1. The querier gains no knowledge of the contents of the provider's database, except for documents that are matched by valid queries.
2. The provider gains no knowledge of the contents of the queries; if possible, that should include inferences based on the documents retrieved.
3. An independent party may restrict the set of legitimate queries.
4. No third parties may gain any knowledge of the queries or the documents.

Dr. Bellovin then spoke about their proposed search scheme based on Bloom filters and Pohlig-Hellman encryption. A semi-trusted third party can transform one party's search queries to a form suitable for querying the other party's database in such a way that neither the third party nor the database owner can see the original query. Furthermore, the encryption keys used to construct the Bloom filters are not shared with this third party. Provision can be made for third-party warrant servers, as well as censorship sets that limit the data to be shared.

## 2.9 Secure Indexes

Speaker: Eu-Jin Goh, Stanford University

Mr. Goh's talk dealt with a data structure for searching on encrypted data. He defined secure indexes as data structures that index the words in a document such that a person with a trapdoor for a word $w$ would be able to search for $w$ in constant time. Without the trapdoor, all contents of the document are hidden. In addition to the application of searching on encrypted data, secure indexes can be used to create searchable encrypted log files, and provide a mechanism for private database queries and private set membership tests.

He listed two possible security models under which secure indexes function. They are IND-CKA (chosen keyword attack on indexes, assuming documents are of equal size) and IND2-CKA (chosen keyword attack on indexes, without assuming documents are of equal size). The former is weaker than the latter, but is almost always sufficient. The IND-CKA model requires that the adversary not be able to deduce the contents of a document from its index (semantic security). Specifically, if the adversary is given the index of one of two equal-sized documents, she should not be able find the document to which the index corresponds. The system needs to be resilient to chosen keyword attack. The IND2-CKA model differs from IND-CKA by requiring the system to be secure even if the documents are of different sizes.

Mr. Goh's secure index scheme called Z-IDX consists of 4 algorithms: Keygen, Trapdoor, BuildIndex, and SearchIndex. These algorithms are based on Bloom filters for efficient set membership tests, and pseudo-random functions. After describing each of the four algorithms, he presented the following advantages of Z-IDX: (1) It can handle arbitrary updates, (2) it produces compressible indexes, (3) it is space efficient for small and medium size documents, (4) the trapdoors are short, (5) it is computationally very efficient, (6) it provides for occurrence search, (7) it provides for efficient Boolean queries and has limited support for regular expression, and (8) it has simple key management. Finally, he provided a contrast between Z-IDX and a recent data structure provided by Chang and Mitzenmacher (see Section 2.10).

## 2.10 Privacy Preserving Keyword Searches on Remote Encrypted Data

Speaker: Yan-Cheng Chang, Harward University

Mr. Chang based his presentation on work jointly done with Michael Mitzenmacher. The talk addressed the following problem: A user $U$ wants to store his files in an encrypted form on a remote file server $S$. Later the user $U$ wants to efficiently retrieve some of the encrypted files containing (or indexed by) specific keywords, keeping the keywords secret and not jeopardizing the security of the remotely stored files. For example, a user may want to store old e-mail messages encrypted on a server managed by Yahoo or another large vendor, and later retrieve certain messages while traveling with a mobile device.

Mr. Chang briefly spoke about two simple solutions and also discussed

the drawbacks in those solutions. The solutions for the problem work under the security requirement that $S$ can learn nothing more than a "keyword is shared by the sent encrypted files."

The key idea is as follows: $U$ uses pseudo-random bits to mask a dictionary-based keyword index for each file and sends it to $S$ in such a way that later $U$ can use the short seeds to help $S$ recover selective parts of the index, while keeping the remaining parts pseudo-random. These schemes are efficient since they do not involve public-key cryptosystems; only heuristic pseudo-random functions are used, which can be implemented efficiently. The approach is independent of the encryption method chosen for the remote files, as long as a keyword index on the corresponding content can be built a priori. They are also incremental, in the sense that $U$ can submit new files which are totally secure against previous queries but still searchable against future queries.

Mr. Chang finally concluded his talk by mentioning the possible directions of research such as ensuring file integrity, preventing file omission, Boolean searches, pattern matching and new applications like P2P.

## 2.11 Completeness in Two-Party Secure Computation—A Computational View

Speaker: Moni Naor, Weizmann Institute

Dr. Naor presented work done jointly with Danny Harnik, Omer Reingold and Alon Rosen. The talk began with the definition of Secure Function Evaluation (SFE). An SFE of a two-variable function $f(x, y)$ is a protocol that allows two parties with inputs $x$ and $y$ to evaluate $f(x, y)$ in a manner where neither party learns more than is necessary. The asymmetric version of SFE is one where only one party gets the output. A function $f$ is *complete for SFE* if a protocol for securely evaluating $f$ allows the secure evaluation of all (efficiently computable) functions. The various important questions in this area include:

- which functions are complete for SFE?,
- which functions have SFE protocols unconditionally?, and
- are there functions which are neither complete nor have efficient SFE protocols?

The previous study of these questions was mainly conducted from an information theoretic point of view and provided strong answers in the form

of combinatorial properties. However, there are major differences between the information theoretic and computational settings. It is possible to show that the functions that are considered as having SFE unconditionally by the combinatorial criteria are actually complete in the computational setting.

Dr. Naor then spoke about his group's work on an almost full characterization of complete functions. A function $f$ is called *computational row transitive* if one can efficiently compute $f(x_1, y)$ when given $x_0, x_1$ and $f(x_0, y)$ for every $x_0, x_1$ and $y$. A function $f$ is said to be *computational row non-transitive* if for some $x_0, x_1$ it is (somewhat) hard to compute $f(x_1, y)$ from $x_0, x_1$ and $f(x_0, y)$ for a random (unknown) $y$. The key results are the following:

1. If $f$ is row transitive then it has an efficient SFE protocol.
2. If $f$ is row non-transitive then it is complete.

Dr. Naor concluded his talk by suggesting open problems, including the following:

1. extending the results to the symmetric SFE model (for non-Boolean functions), and
2. finding computational results for probabilistic functionalities.

## 2.12 Data Mining and Information Privacy—New Problems and the Search For Solutions

Speaker: Tal Zarsky, Yale University

Dr. Zarsky's talk addressed those publicly debated privacy problems that he deemed as critical, and then proposed solutions to them. The talk also examined which forms of privacy policy are adequate.

The focus of this work is restricted to the Internet because it has the advantage of widespread availability, and because both quantity and quality of data collection can be high. Also, the digital environment makes it easy to analyze the data. The Internet is a very interesting test case and an opportunity to learn about policy implications in a wider setting as well. He spoke about recent technological and social changes that have lead to the emergence of information privacy concerns and indicated that data mining, which is a powerful form of data analysis, plays several roles in the information privacy context.

There are three major problems stemming from the collection of personal data, namely, (1) fear that the data will be passed on to the government, (2) fear of the collection of personal data, and (3) fear of specific detriments from the use of personal data. For example, it is possible to use personal information to discriminate among individuals, harm their autonomy, and can lead to unfair results in view of erroneous data and processes.

Dr. Zarsky also discussed how the use of data mining can affect policy decisions which dictate what forms of solutions are needed for solving privacy concerns. He addressed ways in which data mining might undermine the effectiveness of specific solutions, and described how privacy enhancing solutions might undermine the benefits of data mining.

### 2.13 On the Difficulty of Defining Ideal Functionalities for Privacy-Preserving Data Mining: Why naive Secure Multi-party Computation Fails?

Speaker: Yehuda Lindell, IBM Research

Dr. Lindell's talk was based on joint work with Rosario Gennaro, Tal Rabin and Tal Zarsky. The talk addressed the importance of secure multi-party computation and about applying them for privacy-preserving data mining. He discussed the security requirements for privacy-preserving data mining.

Dr. Lindell presented the ideal/real model of defining security and showed how it can be used to model problems of secure distributed computing, and in particular, the problem of privacy-preserving data mining. He then presented two specific data-mining applications: (1) data mining for medical research, and (2) data mining for obtaining personalized newspapers.

Dr. Lindell concluded the talk by stating that a deep understanding of the dangers of non-private data mining must be obtained before attempting to solve the cryptographic privacy-preserving data mining problem.

## 3 Workshop (March 16)

### 3.1 Extending Oblivious Transfer Efficiently

Speaker: Yuval Ishai, Technion

As a common building block in secure computation protocols, oblivious transfer (OT) often is the efficiency bottleneck. Dr. Ishai's talk addressed the problem of *extending* oblivious transfers: Given a small number of OTs

"for free", can one implement a large number of OTs? Although Beaver has shown that a one-way function is sufficient to extend OTs, his extension method is inefficient in practice because of non-black-box use of the underlying one-way function.

Dr. Ishai presented efficient protocols for extending OTs in the random oracle model and a new cryptographic primitive that can be used as a black box to instantiate the random oracle in these protocols. His methods aim at reducing the amortized cost of OT and allowing each additional OT to be generated using a few applications of a cheap *symmetric* primitive. He proposed several questions for further research extending OT using a one-way function as a black-box, improving efficiency in the malicious case, and obtaining similar results for primitives that do not efficiently reduce to OT.

### 3.2 Amortized PIR via Batch Codes

Speaker: Eyal Kushilevitz, Technion

In his talk, Dr. Kushilevitz reviewed the concept and solutions of private information retrieval (PIR) and some known bounds on time and communication. He introduced two approaches to achieve efficient PIR amortization via hashing and amortization via batch codes. Some examples and analysis of both approaches were given. He noted that amortizing PIR from batch codes has the virtue that it is independent of the underlying PIR protocol. Amortization in the multi-user setting was left as an open problem.

### 3.3 Private Inference Control

Speaker: David Woodruff, MIT

Dr. Woodruff's talk discussed private inference control for databases, which includes protocols to prevent the user from inferring private information by sending the database multiple queries which are innocent individually. Private inference contol also seeks to protect the querier from having to reveal its queries to the database.

### 3.4 Privacy as Contextual Integrity

Speaker: Helen Nissenbaum, NYU

In her talk Dr. Nissenbaum reviewed prior philosophical definitions of privacy, and presented her definition of privacy in terms of contextual integrity.

Norms of *appropriateness* determine what types of information are/are not appropriate for a given context. Norms of *distribution* determine the principles governing distribution of information from one party to another. Contextual Integrity is respected when norms of appropriateness and distribution are respected, and it is violated when any of the norms are infringed.

She proposed two questions:

1. Can we develop systematic ways to inform the technical mission of privacy-preserving data transactions with contextual norms?

2. How do we establish meaningful, ongoing conversation across the disciplines - despite vast differences in knowledge bases and methodologies?

### 3.5 Generating Strong Keys from Biometric Data

Speaker: Adam Smith, MIT

Mr. Smith's talk addressed the problem that secure cryptographic keys are long and random strings, which are hard to remember, and the popular biometric solutions are somewhat ad hoc. He talked about how to build a formal framework and general tools for handling personal/biometric key materials. The tasks that need to be addressed are secure password storage and key encapsulation.

He talked about the challenges in using biometric data for cryptographic purposes. For instance, noise and human error in biometric data make it hard to utilize, because cryptography needs precision. Generated keys are not uniformly random, and actually the distribution is unknown. Keys cannot be changed many times, because biometric sources like fingerprints are limited in quantity for each individual. How can we use these data as passwords? He used authentication as an example. A simple abstraction of his approach is *Fuzzy Sketch*. He also addressed the problem of measuring the security of biometrically generated keys. Other topics mentioned in his lecture include code-offset construction, using sketches for authentication, Reed-Solomon-based sketches, and biometric embeddings.

### 3.6 When Can Randomization Fail to Protect Privacy?

Speaker: Wenliang Du, Syracuse University

Dr. Du discussed the question of when randomization fails to protect privacy.

He pointed out that most security analysis methods based on randomization treat each attribute separately, but that is not enough, because the relationship among them may affect privacy. He used the example of perturbing the same number multiple times to show that it is not safe to use simple perturbation on highly correlated data attributes. Then he presented an algorithm that reveals information about the original data based on the available information (the perturbed data and the correlation of the original data).

Problems posed by his presentation are:

- How to improve randomization to reduce the information disclosure?(Making the noises correlated is a possible solution.)
- How to combine principle component analysis with univariate data reconstruction?

### 3.7 Computing Sketches of Matrices Efficiently and Applications to Privacy-preserving Data Mining

Speaker: Petros Drineas, Rensselaer Polytechnic Institute

Dr. Drineas presented methods to compute sketches of matrices efficiently in privacy-preserving data mining. Many applications have to deal with matrices that are too large to store in RAM. He proposed that one can make a few sequential reads through the matrices and create a small *sketch* of the matrices in the RAM. This process should be very fast. The sketch may be used to achieve privacy preservation because it is an approximation to the original matrix. His goal was to formulate a technical definition of privacy that might be achievable by such *sketching* algorithms and provide meaningful and quantifiable protection.

He then described their approach to approximate a large matrix with a sketch, including an algorithm, error bounds, and tightness of the results. He also gave a short description of an alternative approach by Achlioptas and McSherry, and compared the two approaches.

Given the small sketch of a matrix $A$, a *friendly user* can reconstruct a provably accurate approximation $A'$ to the original matrix $A$ and employ any algorithms that he would use to process the original matrix on $A'$ and use the Frobenius and spectral norm bounds for $A \cdot A'$ to argue about the approximation error of his algorithms.

He concluded the talk with two questions:

- How do we ensure privacy for the object-vectors (rows) of $A$ that are revealed as part of the result?
- Do such sketches offer some privacy-preserving guarantees, under some (relaxed) definition of privacy?

### 3.8 Information Leakage and Privacy in Data Mining

Speaker: Poorvi Vora, GWU

In her talk Dr. Vora described a security framework in which the world is divided into *trusted parties* and *untrusted parties*. The trusted parties can be provided with complete data revelation, while information must be completely protected from the untrusted parties. She argued that even secure multi-party computation techniques may leak *private* information, therefore typical security assurance is not sufficient. When talking about the proper definition of privacy, she said the tensions are between access to aggregated information for community and individual control, and between reputation and prejudice. Then she described a model to which she introduced uncertainty to describe players' behavior toward privacy in practice.

Dr. Vora stated that a protocol is a mathematical game between Alice and Bob, and the optimal situation is not when no information is revealed but Alice can get the maximum benefit for her revealed information. She then presented several analogies between cryptographic protocols, communication theory and coding theory. For instance, a protocol can be viewed as a communication channel. Lies are noise and error in the channel, and attacks are codes. As a summary, she described the properties of a desirable model of privacy:

- Provide choices not currently typically available to users.
- Extend the security framework to include problems like those in statistical databases.
- Provide a means of measuring uncertainty in situations where there is some.
- Include other leakage from security-related protocols such as anonymous delivery and ciphers.
- Be useful for measuring the economic value of information.

### 3.9 Random Encodings, Privacy Loss, and Some Possible Solutions - A Coding Theory Perspective

Speaker: Hillol Kargupta, University of Maryland

Dr. Kargupta's talk was about the perspective of privacy-preserving data mining using randomized perturbations in the light of coding theory. He discussed the interplay of additive noise and randomized linear encodings and raised the following question: When the noise is additive, can linear encodings be used for estimating the original data from the perturbed version (i.e. breaching the privacy) with arbitrarily high accuracy by increasing the dimension of the encoded space? This issue appeared to be related to the idea of increasing the accuracy of the data transmission quality through a noisy channel by slowing down the data transmission rate, which is a direct outcome of Shannon's Second Theorem. The talk also included a discussion of the possibility of a new approach toward designing privacy-preserving data mining algorithms by constructing two separate *transmission channels*—one for the original data and another for the type of patterns in the data that we intend to preserve. The idea is to introduce perturbations or transformations that reduce the *transmission capacity* of the data channel for minimizing the privacy loss for the data but increase the *transmission capacity* of the pattern channel for enhancing the quality of the patterns detected by the data mining algorithm.

### 3.10 Secure Regression on Distributed Databases

Speaker: Alan Karr, National Institute of Statistical Sciences

In this talk, Dr. Karr presented methods for performing linear regression on the union of distributed databases that preserve, to varying degrees, confidentiality of those databases. Such methods can be used by federal or state statistical agencies to share information from their individual databases, or to make such information available to others. He also discussed *secure data integration* that provides the lowest level of protection and actually integrates the databases in such a manner that no database owner can determine the origin of any records other than its own, and *regression*, which is associated diagnostics or any other analysis that can be performed on the integrated data. In secure multi-party computation based on shared local statistics, it is necessary to compute least squares estimators of regression coefficients and error variances by means of analogous local computations

that are combined additively using the secure summation protocol. Two approaches were described to model diagnostics in this setting, one using shared residual statistics and the other using secure integration of synthetic residuals.

### 3.11 Tabular Data Release of Conditional and Marginals

Speaker: Aleksandra Slavkovic, Carnegie Mellon University

Dr. Slavkovic introduced prior work to establish bounds and distributions for the cell entries in contingency tables. She presented a framework for finding the bounds and distribution when given an arbitrary collection of marginals and conditionals. This result extends the work of Arnold et al. on the uniqueness of discrete distributions. She described new results on bounds for cell entries in $k$-way tables estimated via optimization methods such as linear and integer programming and gave a complete characterization of the two-way table problem and discuss extensions to multi-way tables including relationships to directed acyclic graphical models. They used tools from algebraic geometry to represent the tables of counts and describe the locus $T$ of all possible tables under the given constraints. Markov bases needed to construct a connected Markov chain over $T$ were described. These bases can be used to induce probability distributions over the space of possible tables via Markov Chain Monte Carlo sampling. This research presents new theoretical links between disclosure limitation, statistical theory and computational algebraic geometry and practical implications for confidentiality and statistical disclosure limitation.

### 3.12 Private Data Mining Based on Randomized Linear Projections

Speaker: Martin Strauss, AT&T Research

Dr. Strauss's talk addressed the main issues in multi-party computation on merged datasets privacy, efficiency (of both time and communication), and (approximate) correctness. Dr. Strauss presented a particular approach that is different from some previous work in its notion of privacy and gave a number of examples. Specifically, the requirement is that neither the messages nor the output of a suitable protocol give more information than the ideal exact answer. By contrast, traditional cryptographic techniques applied to an approximation protocol will preserve time efficiency and hide informa-

tion leakage in the *messages*; unfortunately, the communication complexity may be very high and the approximate *output* may leak private information, yielding a result that is neither efficient nor private. Another class of techniques calls for adding noise to the inputs and/or outputs of a computation; generally, the privacy guarantee is weak unless so much noise is added that the useful information is also obliterated. He showed that several approximations to the vector sum fit our notion of privacy, after straightforward modification. These approximations, including histograms, Fourier decompositions and quantile summaries, were developed without privacy in mind, and are based on random linear projections. He stated that an approximation of size $B$ (e.g., a $B$-bucket histogram) to the vector sum of vectors of length $N$, which has sum square error at most $(1+\epsilon)$ times optimal, can be constructed in time $\mathbf{poly}(\frac{BN}{\epsilon})$ and communication $\mathbf{poly}(\frac{B \cdot log(N)}{\epsilon})$, which is similar to the non-private setting. Nothing is learned by any party that could not be computed from the true exact value of the vector sum.

# 4 Working Group Meeting (March 17)

In the working group, talks were less formal and more speculative, often describing works in progress.

## 4.1 Some Successes and Some Open Questions in Privacy-preserving Data-mining

Speaker: Rafi Ostrovsky, UCLA

Dr. Ostrovsky began his talk by presenting various examples such as PIR, searching encrypted data, and searching on small-CC as computational tasks that preserve privacy. The secure 2-party protocol solves almost all problems but it is very inefficient. He spoke about three measures of computation namely, computation complexity, communication complexity and the number of rounds in the protocol. In his talk he focused on the number of rounds in the general setting.

Dr. Ostrovsky discussed the lower and the upper bounds in terms of the number of rounds for the problem of secure polynomial evaluation in a two party case. He proved that the protocol takes only five rounds and is secure against one (dynamic) Byzantine fault. Though this is not efficient in practice it solves the problem for black-box round complexity in the two-party case.

Finally, Dr. Ostrovsky spoke about the questions that remain to be answered such as the non-black-box approximation of the functions. He also indicated that certain specific functions can be computed more efficiently.

### 4.2 Privacy-preserving Data Sharing in Peer-to-peer Networks

Speaker: Hong Jiang, Yale University

Mr. Jiang presented a joint research agenda with Michael Fischer on privacy issues in peer-to-peer systems. Some prior results (such as the dining cryptographers problem and some practical efforts to implement anonymous communication in decentralized systems) were reviewed. He said that these results, while interesting and inspiring, are not applicable in practical large-scale peer-to-peer networks, because of the mismatch between assumptions and reality in theoretical results and the lack of rigor in practical projects. Jiang and Fischer advocate an approach that carefully defines and studies problems of privacy in peer-to-peer systems in a rigorous model that closely reflects practical systems.

### 4.3 Database Privacy Research at Stanford

Speaker: Krishnaram Kenthapadi, Stanford University

Mr. Kenthapadi gave a brief overview of the database privacy research at Stanford. He listed the various areas of research and briefly discussed each of them.

- Individual centric privacy: The goal here is the ability to manage personal information so that the individual continues to retain control over his/her information even after it is released to an organization. To achieve this it is necessary to design models and mechanisms for the release, acquisition, use and update of personal information (the P4P framework).

- Search over access-controlled data: Here the problem is to design techniques for searching data when the searc engine is not trusted by the content providers. These providers do not want to send documents and provide access-lists to the search engine.

- Aggregates on vertically-partitioned databases: Mr. Kenthapadi discussed the privacy concerns involved in vertically-partitioned databases

where the databases cannot be released as is. The databases can be released after the data is perturbed. The goal is to return high precision aggregate answers.

- Approximations for $k$-anonymity: When databases are released for public use it is a usual practice to delete some attributes. Mr. Kenthapadi spoke about the recent discovery (by other authors) of finding an individual's record from several such databases. To overcome this problem the $k$-anonymity technique suggests generalizing some attributes. He briefly indicated some of the recent work in this direction.

### 4.4 The PORTIA Project

Speaker: Rebecca Wright, Stevens Institute of Technology

Dr. Wright, one of the PIs of the NSF-funded PORTIA (Privacy, Obligations, and Rights in Technologies of Information Assessment) project, co-sponsoring this workshop and the working group, introduced the project. She talked about how changes in technology are making privacy harder and possible threats posed by abuses of sensitive data. She argued that the old models and modes of thought no longer apply to situations arising from new technology, and we should use technology, policy, and education to provide new social mechanisms and new models for better understanding. She used the Metro cards in Washington and New York as an example to illustrate that product decisions by large companies or public organizations become de facto policy decisions but are usually made without conscious thought to privacy impacts, and without public discussion. She also introduced the goals that the PORTIA project is trying to achieve, as well as the institutes, investigators and students actively participating in the project. Finally, she discussed how different techniques like cryptography, secure computation, perturbation and aggregation may be integrated for privacy protection in data mining, and how contextual integrity may help to achieve the goals.

### 4.5 When Do Data Mining Results Violate Privacy?

Speaker: Chris Clifton, Purdue University

Dr. Clifton discussed the question of when data mining results violate privacy. He gave several examples in which private information can or cannot be learned from the given data. To constrain the results of data mining, it is

not enough to only have rules specifying what is acceptable and what is not, but we need constraint-based data mining to enforce the bounds on what we can/cannot learn. To achieve this goal, we need a metric to quantify privacy. He proposed such a metric and a formal definition of anonymity, and he also proposed to study other metrics for privacy and the definition of *privacy-preserving*.

## 4.6 Open Questions and Research Areas in Privacy-preserving Data Mining

Speaker: Yehuda Lindell, IBM Research

Dr. Lindell talked about some open questions and future research areas in privacy-preserving data mining. Specifically, he proposed to study alternative or extended models, including different assumptions about the power of the adversary, assumptions about trust in practice, and security under concurrent general composition instead of a stand-alone model.

Finally, Dr. Lindell spoke about the challenges of applying secure multiparty computation in practical scenarios, and proposed to build a prototype for a realistic scenario. He also insisted that implementation is essential for determining usability since many real problems may only be revealed upon implementation.

## 4.7 Usability and Protection of Privacy

Speaker: Tomas Sander, HP Labs

Dr. Sander discussed the trade-off between usability and protection of privacy. He said that too fine-grained privacy protection is some times not necessary. Building complicated solutions is the wrong idea. The simpler the solution is, the better. If privacy technology is to be used, there must be somebody who implements it, somebody who uses it, and somebody who pays for it. It is a problem that privacy mechanisms are not implemented by those who need them, and companies need utility to adopt them.

Then he talked about the idea of applying digital rights management solutions to privacy problems, and generalizing policy management to personal privacy. With trusted systems, many problems would become trivial since one could send private data to a trusted "black-box" system, do the computation, and then the system would delete the private data. How to achieve this is an interesting research question.

Two questions are posed for consideration:

- Discussions about privacy policy occur at a different level than the technical level. How can the technical community influence policy?
- How can we make these ideas practical?

### 4.8 Research in Different Communities

Speaker: Lawrence Cox, CDC

Dr. Cox stated his view points about the similarity and differences between the cryptography community and the statistical analysis and database community. The cryptography community uses algebraic tools based on theory and looks for stronger protections. In contrast, the results from the latter is data-driven, quick and dirty.

### 4.9 Secure Function Evaluation - the State of the Art

Speaker: Moni Naor, Weizmann Institute

Dr. Naor introduced the definition of secure function evaluation and its security properties. He also reviewed the known general and semi-general results such as the completeness theorem for secure function evaluation (see Section 2.11), as well as some open problems. Then he discussed the technical difficulties, social, legal and commercial implications of verifiable electronic voting.

### 4.10 Friends Trouble Shooting Networks

Speaker: Cynthia Dwork[1], Microsoft Research

Dr. Dwork talked about the *friends troubleshooting network*. The observation is that a considerable proportion of the total cost of owning a desktop PC goes to troubleshooting, and the typical cause of application failure is misconfiguration. Instead of using a centralized approach that could potentially enable aggregation of private data, we could use a peer-to-peer approach. We can use techniques like random walks to provide anonymity in peer-to-peer applications. Dr. Dwork's comments were that this is a great problem to study, and should be generalized. A more rigorous definition of

[1] Dr. Dwork spoke on behalf of Helen J. Wang, Microsoft Research.

the goal of such a system is also needed. Other problems concerning practicability also need further study.

### 4.11 Problems in Privacy Protection

Speaker: Tal Zarsky, Yale University

Dr. Zarsky talked about some legislation problems concerning privacy protection. Although there are constitutional guarantees against government agencies going on "fishing expeditions," it is not clear if certain actions are legal or not. He also discussed privacy and disclosure, and the tension between having a lot more data and preserving privacy. At last he discussed the relation between information equality and the fuzzy notion of full disclosure. Even if everyone has access to all data, not all people have the ability to utilize those data (such as practicing data mining). Therefore full disclosure doesn't necessarily provide equality.

### 4.12 Regulating Sensitive Data Usage

Speaker: Joan Feigenbaum, Yale University

Dr. Feigenbaum talked about challenges in enforcing the regulation of sensitive data use. There are many proposed mechanisms to regulate the use of sensitive data, such as contextual integrity, P3P, HIPAA. But how to enforce these mechanisms is a big problem. We need software that complies with agreed-upon rules and that cannot be hacked around in an undetectable fashion. She mentioned *trusted systems* like Microsoft's *trusted computing* as a possible solution. These systems provide hardware-supported cryptographically strong *attestations* that a remote machine is running an approved software stack. She also talked about problems of applying such technologies in practice, such as technical feasibility, generality, and desirability.

### 4.13 The Statistical Approach

Speaker: Aleksandra Slavkovic, Carnegie Mellon University

Dr. Slavkovic stated two points representing the viewpoint of the statistical community:

1. The guiding principle is to understand potential statistical uses of data in advance. The statistical approach can be related to the context of

the problem, which can lead to new statistical models.

2. We need more communication between statisticians, data miners, and cryptographers. The statistical community is developing statistical disclosure methods almost independently without knowing how they affect data mining tools.

### 4.14 Open Problems in Secure Computation

Speaker: Kobbi Nissim, Microsoft Research

Dr. Nissim reviewed some known general results on secure function evaluation and presented new problems. He talked about private approximation of functions with a concrete example, that secure function evaluation may possibly be applied to the approximation of vertex cover. Then he proposed some open problems for research, including more applications of secure function evaluation, private approximation for objective functions of natural NP-hard problems, private approximation for hard search problems, and whether small leakage in one context may help the adversary in other contexts.

### 4.15 Private Information Retrieval

Speaker: Yuval Ishai, Technion

Dr. Ishai reviewed the two approaches of private information retrieval (PIR): computational PIR and information-theoretic PIR. He also presented the known bounds for both approaches of PIR. Then he bought up some open problems concerning the time complexity and assumptions of PIR, and the problem of finding better upper/lower bounds for information-theoretic PIR.

### 4.16 Open Problems in Privacy Protection

Speaker: Poorvi Vora, GWU

Dr. Vora talked about some open questions in coding, information theory, and signal processing. She also talked about the relationship between privacy and data representation, and how to prevent queries of certain patterns which allow the adversary to infer private information. The problem of lying in a multi-party computation was also discussed.

# 5 Conclusions and Future Directions of Research

This three-day event consisting of a workshop and a working group meeting successfully brought together researchers from the computer science community, the statistics and applied math community, law schools, and industrial research departments. Many speakers insisted that more communication between statisticians, data miners, and cryptographers is necessary. They also suggested that a deeper understanding of the dangers of non-private data mining is necessary before attempting to solve the cryptographic privacy-preserving data mining problem. They also suggested that focusing on implementation for the existing privacy-preserving protocols is essential since many real problems will be revealed upon implementation. The issues discussed and problems identified cover many aspects of privacy-preserving data mining, which is a new broad research scope. A number of open problems and possible directions for future research were proposed in the workshop and in the working group meeting. Some of them are listed below.

1. How to improve the privacy definition to cover everything a realistic adversary will do.

2. How to improve usability and efficiency, such as to determine an alternate way to perturb and use the data that would result in more efficient and accurate data mining and for data mining published statistics.

3. In what ways can one prevent an adversary from inferring the association rules in the database?

4. How to extend the Oblivious Transfer protocols using a one-way function as a black-box and to improve efficiency in the malicious case.

5. Can we develop systematic ways to inform the technical mission of privacy-preserving data transactions with contextual norms?

6. How to establish meaningful, ongoing conversation across the disciplines, despite vast differences in knowledge bases and methodologies.

7. How to improve randomization to reduce the information disclosure.

The event was also an excellent opportunity for graduate students from various institutions to be informed of the state-of-the-art research in this discipline.

In the discussion following the working group meeting, many researchers indicated that they found this event very successful and hoped that such meetings could be held again sometime in the future.

## Acknowledgments

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

We are grateful to all the speakers, especially those who aided us in writing this report by sending us their slides used in the workshop and the working group meeting. Hong Jiang wants to thank Professor Michael Fischer and Professor Chris Clifton for sharing with him their notes for the working group. Geetha Jagannathan wants to thank Rebecca Wright and Brenda Latka for their useful comments on the earlier drafts.

We realize that because of the limitation of the authors' knowledge and the availability of materials, the report may not be accurate or complete. We apologize for those whose work has not been fully and accurately described in this report.