DIMACS Center
Rutgers University

Special Focus on Communication Security and Information Privacy

Annual Report

April 2004

Participants

PI: Fred Roberts, DIMACS

Bill Aiello, AT&T Labs, Vice Chair of the Special Focus on Communication Security and Information Privacy organizing committee.

Ronitt Rubinfeld, NEC Research Institute, chair of the Special Focus on Communication Security and Information Privacy organizing committee.

Workshops to be held during this reporting period:

Workshop: Electronic Voting -- Theory and Practice
May 26 - 27, 2004
Organizers:
Markus Jakobsson, RSA Laboratories
Ari Juels, RSA Laboratories

Workshop: Security Analysis of Protocols
June 7 - 9, 2004
Organizers:
John Mitchell, Stanford
Ran Canetti, IBM Hawthorne

Workshop: Usable Privacy and Security Software
July 7 - 8, 2004
Organizers:
Lorrie Cranor, AT&T
Mark Ackerman, University of Michigan
Fabian Monrose, Johns Hopkins University
Andrew Patrick, NRC Canada
Norman Sadeh, Carnegie Mellon University

Working Group: Usable Privacy and Security Software
July 9, 2004
Organizers:
Lorrie Cranor, AT&T
Mark Ackerman, University of Michigan
Fabian Monrose, Johns Hopkins University
Andrew Patrick, NRC Canada
Norman Sadeh, Carnegie Mellon University

Other Collaborators

Mark Ackerman, University of Michigan, organizer of the Workshop and Working Group on Usable Privacy and Security Software.

Rakesh Agrawal, IBM Almaden, organizer of the Working Group on Privacy / Confidentiality of Health Data.

Bill Aiello, AT&T Labs, Vice Chair of the Special Focus on Communication Security and Information Privacy organizing committee.

Bill Arbaugh, University of Maryland, organizer of the Workshop on Mobile and Wireless Security

Steve Bellovin, AT&T Labs - Research, organizer of the Workshop on Large-scale Internet Attacks.

Ran Canetti, IBM Hawthorne, organizer of the Workshop on Security Analysis of Protocols.

Larry Cox, CDC, organizer of the Working Group on Privacy / Confidentiality of Health Data

Lorrie Cranor, Carnegie Mellon University, organizer of the Workshop and Working Group on Usable Privacy and Security Software.

Cynthia Dwork, Microsoft, organizer of the Workshop and Working Group on Privacy-Preserving Data Mining.

Joe Fred Gonzalez, CDC, organizer of the Working Group on Privacy / Confidentiality of Health Data.

Harry Guess, University of North Carolina, organizer of the Working Group on Privacy / Confidentiality of Health Data

Markus Jakobsson, RSA Laboratories, organizer of the Workshop on Electronic Voting -- Theory and Practice.

Ari Juels, RSA Laboratories, organizer of the Workshop on Electronic Voting -- Theory and Practice

Hugo Krawczyk, Technion, member of the Special Focus on Communication Security and Information Privacy organizing committee.

John Mitchell, Stanford, organizer of the Workshop on Security Analysis of Protocols.

Fabian Monrose, Johns Hopkins University, organizer of the Workshop and Working Group on Usable Privacy and Security Software.

Andrew Patrick, NRC Canada, organizer of the Workshop and Working Group on Usable Privacy and Security Software.

Vern Paxson, ICSI Center for Internet Research, organizer of the Workshop on Large-scale Internet Attacks.

Benny Pinkas, HP Labs, organizer of the Workshop and Working Group on Privacy-Preserving Data Mining.

Avi Rubin, AT&T, member of the Special Focus on Communication Security and Information Privacy organizing committee.

Ronitt Rubinfeld, NEC Research Institute, chair of the Special Focus on Communication Security and Information Privacy organizing committee.

Norman Sadeh, Carnegie Mellon University, organizer of the Workshop and Working Group on Usable Privacy and Security Software.

Tomas Sander, HP Labs, organizer of the Working Group on Privacy / Confidentiality of Health Data.

Stefan Savage, UC San Diego, organizer of the Workshop on Large-scale Internet Attacks.

Stuart Staniford, Silicon Defense, organizer of the Workshop on Large-scale Internet Attacks.

David Wagner, UC Berkeley, member of the Special Focus on Communication Security and Information Privacy organizing committee.

Rebecca Wright, Stevens Institute of Technology, organizer of the Workshop and Working Group on Privacy-Preserving Data Mining and organizer of the Tutorial on Applied Cryptography and Network Security.

Partner Organizations

Telcordia Technologies: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning.

AT&T Labs - Research: Collaborative Research; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and research and workshop/working group organization.

NEC Laboratories America: Collaborative Research; Personnel Exchanges

Partner organization of DIMACS. Individuals from the organization participated in the program planning and research.

Lucent Technologies, Bell Labs: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning.

Princeton University: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning.

Avaya Labs: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning.

HP Labs: Collaborative Research; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and research and workshop/working group organization.

IBM Research: Collaborative Research; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and research and workshop/working group organization.

Microsoft Research: Collaborative Research; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and research and workshop/working group organization.

Centers for Disease Control and Prevention: Collaborative Research; Personnel Exchanges. Individuals from the organization participated in the program planning and research and workshop/working group organization.

Activities and Findings

Overview

Vitally important aspects of our modern society have become dependent on rapid and secure communication, which is increasingly electronic. The new electronic age offers vast potential for new services and applications, but gives rise to serious new vulnerabilities and security threats. Moreover, many of the most important new applications come at the price of threats to privacy. The “special focus” on Communication Security and Information Privacy, which began in summer 2003, is exploring the new vulnerabilities and threats and new methods for dealing with them.

Within the last decade a tremendous transition has taken place in communications networks. Previously, nearly all communication, whether data, voice or other media, was carried over private networks. Anyone who was not a customer of the network provider

was not given physical access to the network. Securing such networks was relatively straightforward. While a great deal of data and media traffic still run over circuit switched or packet switched ATM or Frame Relay private networks, a huge amount and variety of data and media traffic now run over the public Internet, so much so that the Internet is now an important national infrastructure whose integrity is vital to the functioning of our economy, culture, and government. The migration of communication services to the Internet is still very much in progress. This migration brings with it new and complex challenges for maintaining communication security.

There are many factors driving the migration to the Internet. One is universal connectivity. The Internet protocol allows users with many different types of local area network technologies (e.g., Ethernet, and 802.11) to be integrated into a single large network. This allows for a type of positive feedback often referred to as the "network effect." The network grows quickly because the number of users, servers, and devices that are already reachable on the Internet make it very valuable to any new IP device. A second factor is unification. Unlike the circuit switched world for which signaling and data/media were carried by two separate networks, signaling and data/media can both be carried over the Internet. For network providers, migrating their services onto an Internet backbone means that they need only deploy, manage, and control a single network, thereby reducing their cost of providing services. Finally, the ultimate promise of the Internet is as a platform for integrating a variety of services such as voice, instant messaging, mobile presence, multimedia, Web and data services. While these are powerful factors driving the migration to IP communications, they have serious security repercussions. Indeed, securing an extremely large, shared services, packet-based IP network with a large number of administrative domains is a much more complex task than securing segregated/circuit switched networks.

Furthermore, through the collection and dissemination of vast amounts of data, the Internet allows users to take advantage of new functionalities that inherently require new notions of security. For example, new issues of privacy for Internet users and applications are arising due to the multitude of data available online. This new electronic reality and the vast potential for interaction between users and computers give rise to new digital applications and services once thought possibly only in the physical tangible world. This, in turn, creates the need for the invention and implementation of new security and cryptographic techniques. Enabling secure electronic commerce and securing digital rights management are some central examples of the new challenges faced in the security area.

Some of the most exciting progress in the fields of communication security and information privacy has come because of the interconnections of practitioners in these fields with researchers developing relevant methods of theoretical computer science and mathematics. We are exploring these interconnections in order to address some of the fundamental challenges to communication security and information privacy posed by the rapid transition and remarkable growth of new applications in today's communication networks. The project is centered around workshops and research “working groups,” with a tutorial, visitor program, and graduate student program.

The Themes of the Special Focus Include:

- Studying protocol and host vulnerabilities related to Internet communication. Among them are: the weakness or total lack of source authentication for the base protocols in the IP suite, lack of admission control mechanisms, vulnerability of hosts to implementation and configuration errors. What is more, protocol and host vulnerabilities can be exploited in tandem to create serious attacks such as distributed denial of service attacks.

- Securing the protocol layer. The special focus will analyze a wide range of security issues related to newer technologies such as wireless access at the lower layers of the protocol stack, or Web services at the higher layer of the protocol stack, including issues dealing with ad-hoc trust establishment, secure roaming between overlay networks, the controlled execution of untrusted code, and peer-to-peer connection in pervasive networking scenarios.

- Seamless data movement vs. privacy and property rights. The power of service providers to automatically log and analyze information on site visitors or customers for collection and dissemination is so great that it must be properly managed or else there is a significant potential for abuse. The special focus is examining both violation of property rights and violation of privacy both in the general context and in more specialized applications such as health care data and electronic voting.

- Cryptography and secure protocols. As technology evolves, cryptography faces the task of developing new security models and techniques such as developing a complete suite of solutions that can handle the concurrency and asynchrony of the Internet and obtaining information from multiple data sets while protecting privacy and confidentiality.

Tutorials, Workshops, and Working Groups

Tutorial: Applied Cryptography and Network Security
Date: August 4 - 7, 2003
Location: DIMACS Center, CoRE Building, Rutgers University
Organizer: Rebecca Wright, Stevens Institute of Technology
Attendance: 46

The intention of this tutorial was two-fold. One, it was a stand-alone condensed course on cryptography and its applications to secure networking and electronic commerce, giving an introduction to some of the fundamental issues in this field. Two, it was designed to provide background knowledge to researchers and graduate students who wished to participate in the DIMACS Special Focus on Communication Security and Information Privacy. The tutorial appears to have been successful in both regards. For example, several graduate student participants have returned or plan to return for later workshops in the special focus and have said that they feel the tutorial helped them to get more out of the workshops than they otherwise would have. Another participant, a professor at an

undergraduate college, was attending the tutorial in order to help him with curriculum development in cryptography and security at his college. On a more personal level, due to interactions initiated at the tutorial, the graduate student author of the report on the tutorial is now a Ph.D. student at Stevens Institute of Technology under the guidance of tutorial organizer Dr. Wright. Each tutorial day included both lectures and problem sessions. The following were the topics of the lectures:

- cryptographic primitives and protocols: symmetric key cryptography, public key cryptography, authentication, and key exchange protocols
- key management and access control: public key infrastructures and trust management
- network security: snooping, spoofing, distributed denial of service attacks, SSL, SSH, IPsec.
- electronic commerce: electronic payments protocols, auctions

There were 41 participants in the tutorial in addition to the 5 lecturers. The academic participants included slightly more students than faculty. About a quarter of the participants were from industry. In addition to the United States participants, there were students and faculty from South Korea and Canada.

The lectures were:

Principles of Security and Modern Cryptography, Symmetric Encryption, Amir Herzberg, Bar-Ilan University
Rijndael, Arta Doci, University of Colorado
Hashing, Amir Herzberg, Bar-Ilan University
Message Authentication Codes (MAC), Hugo Krawczyk, Technion
Public Key Cryptography, Rebecca Wright, Stevens Institute of Technology
Public Key Infrastructures (PKI), Access Control, Trust Management, Amir Herzberg, Bar-Ilan University
Resilience to Key Exposure: Revocation, Forward Security, Secret Sharing, Threshold and Proactive Security, Amir Herzberg, Bar-Ilan University
Distributed Denial of Service Attacks, Software Security, Angelos Keromytis, Columbia University
Internet Crypto Tools, Amir Herzberg, Bar-Ilan University
Electronic Commerce: Payment Protocols and Fair Exchange, Markus Jakobsson, RSA Laboratories

In addition to the lectures, each day included a one-hour "problem session." During these sessions, Nelly Fazio interactively led the tutorial participants in working through a number of exercises, some of which had been handed out the day before. These sessions were designed to make the learning experience more complete by giving participants a chance to try to solve some problems on their own and then see how well they had done.

An extensive report on this activity is given in *Report on DIMACS Tutorial on Applied Cryptography and Network Security*, which can be found at http://dimacs.rutgers.edu/Workshops/ComputerSecurity/appl-crypt-8-03.pdf

Workshop: Large-scale Internet Attacks

Date: September 23 - 24, 2003
Location: DIMACS Center, CoRE Building, Rutgers University
Organizer: Vern Paxson, ICSI Center for Internet Research; Steve Bellovin, AT&T Labs - Research; Stuart Staniford, Silicon Defense; Stefan Savage, UC San Diego
Attendance: 65

With the increasing size of the Internet, we have seen an increasing number of attacks that take advantage of the network's large scale. These kind of large-scale Internet attacks are usually difficult to counter because of the difficulties in tracing them back or deploying widespread defensive measures. This workshop explored four general types of large-scale attacks and the possible countermeasures:

(1) Distributed Denial of Service (DDoS), in which collections of hundreds or thousands of compromised machines are coordinated to simultaneously send floods of bogus traffic towards a target, completely overwhelming the target's resources, or those of the target's network;
(2) Self-propagating Malicious Code, or Worms, which have in recent years compromised hundreds of thousands of Internet hosts in a matter of hours (with recent work arguing that future worms will likely be even more rapid, and/or much harder to detect);
(3) Infrastructure Attacks, which attempt to subvert the key components of the Internet's underlying infrastructure (domain name system, routing);
(4) Attacks on Large-scale Services, which take advantage of the fact that the Internet's growth has seen the rise of some very large, publicly accessible services (such as portals, search engines, and auctions), which gain their utility by their very scale, but generally do so by making access to the service extremely cheap and thus open to a new class of sophisticated, highly automated attacks.

The workshop included the following talks:

Experiences with large-scale attacks: A Large-scale View of Large-scale Attacks, Sean Donalen, SBC Internet Services
Infrastructure attack trends, Craig Labovitz, Arbor Networks
Attacks on services - unofficial perspectives from a CDN
Attacks on services - unofficial perspectives from a large website
DDoS: Overview, John Ioannidis, AT&T Labs - Research
Defense, Angelos Keromytis, Columbia University
Source address filtering, Avi Freedman
Techniques: Telescopes, David Moore, UCSD
Sampling techniques, George Varghese, UCSD
P2P techniques, large-scale coordination, Joel Sandin, Stanford University
Honeynet, Dave Dittrich, University of Washington
Worms: Overview, Stuart Staniford, Silicon Defense
Diverse axes of scaling, Dan Ellis, MITRE
Modeling/detecting worm propagation, Lixin Gao, University of Massachusetts
Topological worm defenses, Nick Weaver, UCB
Pulsing attacks on router, Avi Freedman

Auto-patching, Angelos Keromytis, Columbia University
Attacks on routing: BGP attack, Avi Freedman
Targeted link attack, Steve Bellovin, AT&T Labs - Research
Authentication and robustness, Alex Snoeren, UCSD

Stefan Savage, UCSD, led a discussion of future research challenges. Many were identified, including:

- How do we accurately detect the large-scale attacks in the Internet? Sometimes, it is difficult to distinguish the normal activities from the attacks.
- How do we collect and analyze the huge amount of attack monitoring information and do it in real time?
- How do we divide the responsibility and obligation between the network service providers and their customers with regard to the response to the attacks, and which kind of business service model is needed here?
- How do we build a formal collaboration between different sites, ISPs and various agencies, which can automate the coordination among related parties in the case of being attacked?
- How do we motivate the ISPs to deploy the defensive measures, e.g., source address filtering to defend against the DDoS attacks?
- How do we protect the network infrastructure itself from the possible attacks?
- How do we deal with the polymorphism of the worms, including the syntactic and semantic polymorphism?
- How do we build a secure worm detection system in the distributed systems?
- Given the possible high spreading speed of the future worms, how do we deploy the worm containment mechanism to contain them?

An extensive report on this activity is given in *Report on DIMACS Workshop on Large-scale Internet Attacks*, which can be found at
http://dimacs.rutgers.edu/Workshops/Attacks/internet-attack-9-03.pdf

Working Group Meeting: Privacy / Confidentiality of Health Data
Date: December 10 - 12, 2003
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Rakesh Agrawal, IBM Almaden; Larry Cox, CDC; Joe Fred Gonzalez, CDC; Harry Guess, University of North Carolina
Attendance: 37

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, including:

Overview of Statistical Disclosure Limitation, Lawrence H. Cox, Associate Director, ORM, NCHS, CDC

Legal and Regulatory Framework in the United States and the European Union, Oliver Johnson, Merck and Co., Inc.

The Health Insurance Portability and Accountability Act (HIPAA) and its Implications on Epidemiological Research Using Large Databases, K. Arnold Chan, Harvard University

Health Care Databases under HIPAA: Statistical Approaches to De-identification of Protected Health Information, Judith Beach, Quintiles Transnational

Protecting the Privacy of Healthcare Data While Preserving the Utility of Geographic Location Information for Epidemiologic Research, Daniel Barth-Jones, Center for Healthcare Effectiveness Research and Department of Medicine, Wayne State University, School of Medicine

Privacy Technologies and Challenges in their Deployment, Tomas Sander, HP Labs

Software Demonstration of the use of Hippocratic Database Technology in Supporting a Health Care Provider. Tyrone Grandison, IBM

Cryptographic Techniques for Confidentiality of Aggregate Statistics on Health Data, Giovanni DiCrescenzo, Telecordia

Tutorial on Data Mining, David Madigan, Rutgers University

Using Data Mining Techniques to Harvest Information from Clinical Trials, Richard D. De Veaux, Williams College, Williamstown, MA.

Experimental Results on Privacy-Preserving Statistics Computation, Rebecca Wright, Stevens Institute of Technology

Semantic Web Services for Privacy/Confidentiality of Health Care Data, Nabil Adam, Rutgers University

Privacy/Confidentiality Issues in Collecting Agricultural Data, Gary Smith, University of Pennsylvania

Private Analysis of Data Sets, Benny Pinkas, HP Labs, NJ

Overview of Masking Schemes for Microdata, Jay J. Kim, ORM, NCHS, CDC

Statistical Disclosure Limitation in Tabular Data and Related Mathematical and Computational Problems, Lawrence H. Cox, ORM, NCHS, CDC

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?

An extensive report on this activity is given in *Report on DIMACS Working Group on Privacy / Confidentiality of Health Data*, which can be found at http://dimacs.rutgers.edu/Workshops/Health/priv-health-data-12-03.pdf

Workshop: Privacy-Preserving Data Mining

Date: March 15 - 16, 2004

Location: DIMACS Center, CoRE Building, Rutgers University

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

Attendance: 88

This workshop and associated working group meeting brought together researchers and practitioners in cryptography, data mining, and other areas to discuss privacy-preserving data mining. The workshop consisted of invited talks and discussion, including:

From Idiosyncratic to Stereotypical: Toward Privacy in Public Databases, Shuchi Chawla, CMU

Privacy-Preserving Datamining on Vertically Partitioned Databases, Kobbi Nissim, Microsoft Research

Confidentiality in Tables Viewed from an Algebraic Perspective, Lawrence H. Cox, CDC

Privacy Preserving Computation of the k'th Ranked Element, Gagan Aggarwal, Stanford University

Efficient Private Matching and Set Intersection, Mike Freedman, NYU

An Experimental Study of Association Rule Hiding Techniques, Emmanuel Pontikakis, University of Patras
Public-Key Encryption with Keyword Search, Giovanni DiCrescenzo, Telcordia
Privacy-Enhanced Searches Using Encrypted Bloom Filters, Steve Bellovin, AT & T Research
Secure indexes, Eu-Jin Goh, Stanford University
Privacy Preserving Keyword Searches on Remote Encrypted Data, Yan-Cheng Chang, Harvard University
Completeness in Two-Party Secure Computation - A Computational View, Moni Naor, Weizmann Institute
Data Mining and Information Privacy - New Problems and the Search For Solutions, Tal Zarsky, Yale University
On the Difficulty of Defining Ideal Functionalities for Privacy Preserving Data Mining: Why Naive Secure Multiparty Computation Fails, Yehuda Lindell, IBM
Extending Oblivious Transfers Efficiently, Yuval Ishai, Technion
Amortized PIR via Batch Codes, Eyal Kushilievitz, Technion
Private Inference Control, David Woodruff, MIT
Cryptographic Randomized Response Techniques, Markus Jakobsson, RSA Security
Privacy as Contextual Integrity, Helen Nissenbaum, NYU
Trading Entropy for Privacy, or Unconditional Security When Information Leakage is Unavoidable, Adam Smith, MIT
Calypso: UCSD's Project on Privacy in Database Publishing, Alin Deutsch, UCSD
When can the Randomization Fail to Protect Privacy?, Kevin Du, Syracuse University
Computing Sketches of Matrices Efficiently and Applications to Privacy Preserving Data Mining, Petros Drineas, Rensselaer Polytechnic Institute
Information Leakage and Privacy in Data Mining, Poorvi Vora, GWU
Random Encodings, Privacy Loss, and Some Possible Solutions: A Coding Theory Perspective, Hillol Kargupta, University of Maryland
Secure Regression on Distributed Databases, Allan Karr, National Institute of Statistical Sciences
Tabular Data: Releases of Conditionals and Marginals, Aleksandra Slavkovic, Carnegie Mellon University
Private Data Mining Based on Randomized Linear Projections, Martin Strauss, AT&T Research

Shuchi Chawla concluded her talk by presenting some future directions such as extending privacy arguments to various distributions, characterizing acceptable auxiliary information to other interesting macroscopic properties.

Kobbi Nissim briefly indicated 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 SuLQ DBs.
- Obtaining strong privacy definition and rigorous privacy proof in SuLQ.

- Gaining usability for the data miner for both single and vertically split databases

Michael Freedman presented some open problems such as finding a more computationally efficient protocol, and, in fuzzy matching, obtaining protocols which are secure against malicious adversaries.

Emmanuel Pontikakis' open questions, included:

- 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.

Yan-Cheng Chang mentioned possible directions of research such as ensuring file integrity, preventing file omission, Boolean searches, pattern matching and new applications like P2P.

Yehuda Lindell stated 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.

Helen Nissenbaum asked two questions. Can we develop systematic ways to inform the technical mission of privacy-preserving data transactions with contextual norms? How do we establish meaningful, ongoing conversation across the disciplines - despite vast differences in knowledge bases and methodologies?

Problems posed by Wenliang Du include: How do we improve randomization to reduce information disclosure? (Making the noises correlated is a possible solution.) How do we combine principle component analysis with univariate data reconstruction?

An extensive report on this activity is given in* Report on DIMACS/PORTIA Workshop and Working Group on Privacy-Preserving Data Mining*, which is in preparation.

Working Group Meeting: Privacy-Preserving Data Mining
Date: March 17, 2004
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Rebecca Wright, Stevens Institute of Technology; Benny Pinkas, HP Labs; Cynthia Dwork, Microsoft
Attendance: 53

The working group identified and explored approaches that could serve as the basis for more sophisticated algorithms and implementations than presently exist, and discussed directions for further research and collaboration. Some of the individual presentations included:

Some Successes and Some Open Questions in Privacy-Preserving Data Mining, Rafi Ostrovsky, UCLA

Privacy-Preserving Data Sharing in Peer-to-Peer Network -- A Research Agenda, Michael Fischer and Hong Jiang, Yale University
Overview of Database Privacy Research at Stanford, Krishnaram Kenthapadi and Dilys Thomas, Stanford University
The PORTIA Project, Rebecca Wright, Stevens Institute of Technology
When do Data Mining Results Violate Privacy?, Chris Clifton, Purdue University
Handling Incompatible Formats and Erroneous Data in the Context of Privacy-Preserving Data Mining, Arta Doci, University of Colorado

Much of the meeting consisted of the discussion of open research areas and future working group plans. Some of the ideas presented include:

Problems of privacy in peer-to-peer systems should be carefully defined in a rigorous model that closely reflects practical systems.

Research in privacy-preserving data mining should include the development of 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.

There are challenges in applying secure multiparty computation in practical scenarios. A prototype must be built for a realistic scenario. Implementation is essential for determining usability since many real problems may only be revealed upon implementation.

Digital rights management solutions can be applied to privacy problems. 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.

Discussions about privacy policy need to occur at a different level than just the technical level. How can the technical community influence policy? How can the technical ideas be made practical?

A considerable portion of the total cost of owning a desktop PC goes to troubleshooting, and the typical cause of application failure is mis-configuration. Instead of using a centralized approach, which could potentially enable aggregation of private data, we could use a peer-to-peer approach. We can use techniques like random walk to provide anonymity in peer-to-peer applications. This is an important approach to study, and should be generalized.

Some open problems in secure function evaluation for research include developing private approximation for objective functions of NP-hard problems, developing private approximation for hard search problems, and studying whether small leakage in one context may help the adversary in other contexts.

There are two approaches to the study of private information retrieval (PIR): computational PIR and information-theoretic PIR. Open questions concern the time complexity and the problem of finding better upper/lower bounds for information-theoretic PIR.

How do we prevent queries of certain patterns that allow the adversary to infer private information? There is also the problem of lying in a multi-party computation.

An extensive report on this activity is given in* Report on DIMACS/PORTIA Workshop and Working Group on Privacy-Preserving Data Mining*, which is in preparation.

Additional workshops that have been scheduled for the time frame of this report are:

Workshop: Electronic Voting -- Theory and Practice
Dates: May 26 - 27, 2004
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Markus Jakobsson and Ari Juels, RSA Laboratories

Workshop: Security Analysis of Protocols
Dates: June 7 - 9, 2004
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: John Mitchell, Stanford and Ran Canetti, IBM Hawthorne

Workshop: Usable Privacy and Security Software
Dates: July 7 - 8, 2004
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Lorrie Cranor, AT&T; Mark Ackerman, University of Michigan; Fabian Monrose, Johns Hopkins University; Andrew Patrick, NRC Canada; Norman Sadeh, Carnegie Mellon University

Working Group Meeting: Usable Privacy and Security Software
Date: July 9, 2004
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Lorrie Cranor, AT&T; Mark Ackerman, University of Michigan; Fabian Monrose, Johns Hopkins University; Andrew Patrick, NRC Canada; Norman Sadeh, Carnegie Mellon University

Findings

<u>Efficient Private Matching and Set Intersection</u>

Kobbi Nissim, Microsoft Research, Benny Pinkas, HP Labs and Michael Freedman, New York University, are considering the problem of computing the intersection of private datasets of two parties, where the datasets contain lists of elements taken from a large domain. This problem has many applications for online collaboration. They have

developed protocols, based on the use of homomorphic encryption and balanced hashing, for both semi-honest and malicious environments. For lists of length k, they obtained O(k) communication overhead and O(k ln ln k) computation. The protocol for the semi-honest environment is secure in the standard model, while the protocol for the malicious environment is secure in the random oracle model. They are also considering the problem of approximating the size of the intersection, have found a linear lower bound for the communication overhead of solving this problem, and have developed a suitable secure protocol. Lastly, they are investigating other variants of the matching problem, as well as the problem of approximate matching. Their paper "Efficient Private Matching and Set Intersection," will be the opening paper at the Eurocrypt 2004 conference at Interlaken, Switzerland, 2-6 May 2004. Eurocrypt 2004 is a scientific conference that focuses on research in cryptology. It is organized by the International Association for Cryptologic Research (IACR) in cooperation with the Network Security and Cryptography Group at the IBM Zurich Research Laboratory**.**

Masking Microdata

Joe Fred Gonzalez, Jr. and Lawrence H. Cox, co-organizers for the Working Group on Privacy and Confidentiality of Health Data, one of the invited speakers, Jay J. Kim, National Center for Health Statistics (NCHS), and one of the participants, Myron Katzoff, NCHS, have initiated research and obtained results in a couple of areas dealing with masking microdata. They have studied the effects of rounding continuous data using specific rules. Data such as incomes are frequently rounded. Rounding may be done to protect the confidentiality of records in a file or to enhance readability of the data, or by the notion that the digits subject to rounding are inconsequential. The rounding may not have any effect on the bias of an estimator, but may have a large impact on variance. Integers can be expressed as $x = qB + r$, where q is the quotient, B is the base, and r is the remainder. B is a constant, but q and r are random variables. They use four rules for rounding r to observe the effects of rounding on bias and variance. They assume a uniform distribution on r, but no specific distributional assumption is made on q. When $q = 0$, they show that the variance after rounding is three times the variance before rounding. As the variance of q gets larger, the effect of rounding on the variance decreases. They have computed the disclosure risk in terms of the posterior probability $P(x|qB)$.

Another problem this group has studied is the effects of grouping data on first and second distribution moments. Data such as income are often grouped and released as interval data, considered to be one of the best ways of summarizing data that has disclosure risk implications as well. Class marks (midpoints) of intervals are then used to calculate the mean and variance of the grouped data. In most situations, using midpoints for every observation in the interval smoothes the data, thereby reducing the variance. It can be shown, as in analysis of variance, that, using midpoints, we lose the within-interval variance component if within-interval data have a uniform distribution. However, if distributions within some intervals are peaked or skewed, use of the midpoints of the interval data can result in higher variance estimates than would be obtained with the raw data. Moreover, for those data, the mean of the grouped data based on the use of

midpoints is biased. If class (conditional) means are used for calculating overall mean and variance, the mean of the raw data can be recaptured and the variance will be lower. They have obtained some initial results from their study of the impact of accepted practices for approximating moments with summarized data.

As a result of their work, Gonzalez, Cox, Kim, and Katzoff will present two papers at the American Statistical Association Joint Statistical Meetings (JSM 2004), *Statistics as a Unified Discipline*, August 8-12, 2004 in Toronto, Canada. The papers are "Effects of Rounding Continuous Data Using Specific Rules" and "Effects of Grouping Continuous Data on First and Second Distribution Moments." These papers will be published in the ASA Survey Research Methods Section proceedings.

Books

Amir Herzberg, *Introduction to Secure Communication and Commerce, with Applied Cryptography*, in preparation.

Papers

Freedman, M., Nissim, K., and Pinkas, B., "Efficient Private Matching and Set Intersection," to appear in the Proceedings of **Eurocrypt '2004**, May 2-6, 2004.

Cox, L., Gonzalez, Jr, J. F., and Katzoff, M., "Effects of Rounding Continuous Data Using Specific Rules," to appear in **Proceedings of the ASA Joint Statistical Meetings Survey Research Methods Section**.

Cox, L., Gonzalez, Jr, J. F., and Katzoff, M., "Effects of Grouping Continuous Data on First and Second Distribution Moments," to appear in **Proceedings of the ASA Joint Statistical Meetings Survey Research Methods Section**.

Herzberg, A. "Preventing Spoofing, Spamming and Phishing," in preparation.

Main website

http://dimacs.rutgers.edu/SpecialYears/2003_CSIP/

Other Specific Products

Web pages

DIMACS Tutorial on Applied Cryptography and Network Security
http://dimacs.rutgers.edu/Workshops/ComputerSecurity/

DIMACS Workshop on Large-scale Internet Attacks
http://dimacs.rutgers.edu/Workshops/Attacks/

DIMACS Working Group on Privacy / Confidentiality of Health Data
http://dimacs.rutgers.edu/Workshops/Health/

DIMACS/PORTIA Workshop on Privacy-Preserving Data Mining
http://dimacs.rutgers.edu/Workshops/Privacy/

DIMACS/PORTIA Working Group Meeting on Privacy-Preserving Data Mining
http://dimacs.rutgers.edu/Workshops/WGDatasets/

DIMACS Workshop on Electronic Voting -- Theory and Practice
http://dimacs.rutgers.edu/Workshops/Voting/

DIMACS Workshop on Security Analysis of Protocols
http://dimacs.rutgers.edu/Workshops/Protocols/

DIMACS Workshop on Mobile and Wireless Security
http://dimacs.rutgers.edu/Workshops/MobileWireless/

DIMACS Workshop on Usable Privacy and Security Software
http://dimacs.rutgers.edu/Workshops/Tools/

DIMACS Working Group on Usable Privacy and Security Software
http://dimacs.rutgers.edu/Workshops/WGTools/

Reports

Report on DIMACS Tutorial on Applied Cryptography and Network Security
Report Author: Geetha Jagannathan, Department of Computer Science, SUNY at Stony Brook
Date of report: December 3, 2003
http://dimacs.rutgers.edu/Workshops/ComputerSecurity/appl-crypt-8-03.pdf

Report on DIMACS Workshop on Large-scale Internet Attacks
Report Author: Xuhui Ao, Dept. of Computer Science, Rutgers University
Date of report: November 30, 2003
http://dimacs.rutgers.edu/Workshops/Attacks/internet-attack-9-03.pdf

Report on DIMACS Working Group on Privacy / Confidentiality of Health Data
Report Authors: Hiran Subramaniam and Zhiqiang Yang, Department of Computer Science, Stevens Institute of Technology
Date of Report: December 20, 2003
http://dimacs.rutgers.edu/Workshops/Health/priv-health-data-12-03.pdf

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

Report Authors: Geetha Jagannathan, Department of Computer Science, SUNY at Stony Brook and Hong Jiang, Department of Computer Science, Yale University

In preparation.

Contributions

Contributions within Discipline

The "discipline" of this project is computer science, broadly speaking, with related areas of mathematics, statistics and electrical engineering. The main contribution of this project at this stage is threefold.

1. The Tutorial on Applied Cryptography and Network Security has provided its participants with an introduction to some of the fundamental issues in this field, giving them the background knowledge to enable them to participate in the Special Focus.
2. Each of the activities has contributed to the explicit description of a myriad of open questions and research challenges. These have been discussed in great detail in the section on Activities.
3. The interactions among the participants have already led to new research collaborations and potential collaborations as well as new research directions for existing research groups. See section on Contributions to Human Resource Development.

Even at this early stage of the project there have been research results in computer science. For example, Kobbi Nissim, Microsoft Research, Benny Pinkas, HP Labs and Michael Freedman, New York University have obtained results on the problem of computing the intersection of private datasets of two parties, where the datasets contain lists of elements taken from a large domain. More details on these results are given in the section on Findings. We expect many more results as the project progresses.

Additional results of the project are presentations at professional meetings, new courses and units of courses by participants back on their own campuses, and the development of research directions for graduate students and new research directions for researchers.

Contributions to Other Disciplines

This project has already produced results in statistics that are closely related to privacy concerns arising from data in a wide variety of fields. The primary motivation for this work involves health data privacy. The work of Joe Fred Gonzalez, Jr., Lawrence H. Cox, Jay J. Kim, and Myron Katzoff on the effects of rounding continuous data using specific rules and this same group's results on the effects of grouping data on first and second distribution moments are described in the section on Findings.

Contributions Beyond Science and Engineering

In a very real sense, this project's most significant contributions are beyond science and engineering. The computer science, mathematics, and statistics of this project are motivated by vitally important problems in our modern society. Our society has become dependent on rapid and secure communication, which is increasingly electronic. The new electronic age offers vast potential for new services and applications, but gives rise to serious new vulnerabilities and security threats. Moreover, many of the most important new applications come at the price of threats to privacy. Within the last decade a tremendous transition has taken place in communications networks. A huge amount and variety of data and media traffic now run over the public Internet, so much so that the Internet is now an important national infrastructure whose integrity is vital to the functioning of our economy, culture, and government. The migration of communication services to the Internet brings with it new and complex challenges for maintaining communication security. Furthermore, through the collection and dissemination of vast amounts of data, the Internet allows users to take advantage of new functionalities that inherently require new notions of security. For example, new issues of privacy for Internet users and applications are arising due to the multitude of data available online. This new electronic reality and the vast potential for interaction between users and computers give rise to new digital applications and services once thought possible only in the physical tangible world. This, in turn, creates the need for the invention and implementation of new security and cryptographic techniques. Enabling secure electronic commerce and securing digital rights management are some central examples of the new challenges faced in the security area.

The activities in this project have created a dialogue among the principal players protecting against Internet attacks, conforming to new laws safeguarding health data privacy, and enabling collaborative research using privacy protected data. Lawyers, epidemiologists, cryptographers, and statisticians are sharing their areas of expertise to define the problems and the approach to their solutions. Private sector Internet service providers are sharing information and data with computer scientists and statisticians to understand the nature of past Internet attacks and forecast and protect against future attacks.

Contributions to Human Resources Development

As stated in the section on Research and Education Activities, as a result of the *Tutorial: Applied Cryptography and Network Security* several graduate student participants have returned or plan to return for later workshops in the special focus and have said that they feel the tutorial helped them to get more out of the workshops than they otherwise would have. Another participant, a professor at an undergraduate college, was attending the tutorials to help him with curriculum development in cryptography and security at his college.

Graduate students have authored reports for each of the program activities. To produce the reports, the students engaged in significant interaction with the organizers and the speakers, making contacts that would almost surely not have developed otherwise. Due to interactions initiated at the Tutorial on Applied Cryptography and Network Security, the graduate student author of the report is now a Ph.D. student at Stevens Institute of Technology under the guidance of workshop organizer Rebecca Wright.

Both graduate students and undergraduates have been given the opportunity to make presentations. For example, the Workshop on Privacy-Preserving Data Mining included the talk by an undergraduate from the University of Patras, Emmanuel Pontikakis, *An Experimental Study of Association Rule Hiding Techniques.* In the associated working group meeting, Hong Jiang, a graduate student at Yale University, gave a presentation on *Privacy-Preserving Data Sharing in Peer-to-Peer Network -- A Research Agenda.*

The following is a typical reaction of a Ph.D. student participating in the special focus.

> "I am a third year PhD student at Stanford, advised by Prof. Rajeev Motwani. I attended the DIMACS/PORTIA workshop/working group on Privacy-Preserving Data Mining in March this year. It was very helpful to learn about the current research results / future directions. Interaction with other researchers in the field was also very useful.
>
> The workshop involved talks by researchers from diverse fields such as cryptography (secure multiparty computation), databases, statistics and law. This provided a broad view of the area. As I am starting to work in this area, the workshop helped me to know about the current as well as earlier results. For example, Eyal Kushilevitz's talk enlightened me about some lower bound results in Private Information Retrieval on which I have been working. The open problem session in the working group meeting was especially helpful." - Krishnaram Kenthapadi, Stanford

Effects on more established researchers have also already been demonstrated. Many new collaborations have begun. One example is the following:

> Martin Strauss, AT&T Shannon Laboratory, and Petros Drineas, Assistant Professor of Computer Science at Rensselaer Polytechnic Institute, attended the Privacy-Preserving Data Mining workshop and working group meeting co-organized by Rebecca Wright of Stevens Institute of Technology. As a result of these meetings, two collaborations began, one involving Martin Strauss and Rebecca Wright and a second involving Martin Strauss, Petros Drineas, and Michael Mahoney of Yale. In both cases, the topic is secure multiparty computations of approximations. The idea is: for some target function f(x,y), where the various inputs are each owned by suspicious players, compute some other function g(x,y), such that:
>
> - g(x,y) is a good approximation to f(x,y).

- g(x,y) is efficiently computable--typically much more efficient than f(x,y). This includes communication efficiency in the case where x and y are massive data sets.
- g(x,y) is private, in the sense that no party learns anything more than what they "ought." That is, each party knows their own input and "ought" to learn the exact output f(x,y); neither the output g(x,y) nor the intermediate messages of g(x,y)'s computation gives additional information.

The current (preliminary) research of Strauss, Drineas, and Mahoney focuses on private computations of singular values and (perhaps) singular vectors of matrices. The goal is to allow multiple parties that possess different parts of a matrix to perform such computations while leaking the minimum amount of information. Applications are numerous; namely, all the areas where distributed singular value and singular vector computations are useful (e.g. Internet, Social Networks, etc.).

To give a second example of the impact of the special focus on a more established research participant, we note the following from Amir Herzberg, Computer Science Department, Bar Ilan University:

> "I was the principal lecturer in DIMACS Tutorial on Applied Cryptography and Network Security, August 4 - 7, 2003. I prepared a substantial amount of lectures for this course, and I am still using these lectures in my university classes and in other forums; I also make all of my lectures (including these) available in my website (with a lot of downloads). I am now working on a book on *Introduction to Secure Communication and Commerce, with Applied Cryptography*, and I believe the lectures I gave will help me in this project.
>
> I also believe that preparing these lectures, and esp. the PKI and SSL/TLS lectures, helped me identify important problems and solutions in security of web and e-mail, which I am now describing in a paper tentatively titled "Preventing Spoofing, Spamming and Phishing."
>
> In summary: I hope this course gave the participants as much as it helped me... I'll definitely be happy to repeat the experience!"