

Data Privacy in Biomedicine (BMIF-380 / CS-396)

Instructor: Bradley Malin

Semester: Spring

Website: <http://people.vanderbilt.edu/~b.malin/BMIF380/index.html>

DESCRIPTION

The integration of information technology into biomedical environments has enabled unprecedented advances in the collection, storage, analysis, and rapid dissemination of patient-specific data to physicians and researchers. Given the potential wealth of such detailed databases for further advances in healthcare, many organizations share, or anticipate sharing, their collections for various purposes related to quality assurance, public health, and research. However, in the face of today's complex networked environments, many organizations are finding it increasingly difficult to share biomedical data due to concerns about patient privacy and anonymity. For instance, how can we share patient-specific data without revealing the identity of the patient? Security practices, such as role based access control and encrypted communications ensure authentication and secure communications, but they do not prevent the leakage of inferences from the data after it has been accessed or transmitted. Thus, this course is concerned with the analysis and protection of data privacy. The goal of this course is to introduce students to the computational challenges, as well as formal privacy protection solutions, for data privacy in healthcare and biomedical research environments. The topology of data privacy is a highly interdisciplinary landscape and material in this course will touch on issues and methodologies from bioinformatics, cryptography, data mining, databases, distributed systems, law, machine learning, medical informatics, policy, and statistics.

OBJECTIVES

After this course, students will be able to analyze data privacy issues from three non-exclusive perspectives:

1. *Data Detectives*: Oftentimes data is shared with false beliefs about privacy and data protection. From this perspective students will learn how seemingly private information, can be learned using automated strategies.
2. *Data Protectors*: Students will learn how to construct privacy protection technologies that provide formal computational guarantees of privacy in data collection and sharing.
3. *Technology Policy Designers*: Computational models provide a basis for protection, but in order to implement such technology in the real world, it must support, and not circumvent, existing policy specification. From this perspective, students will learn how to develop privacy protection solutions which complement policy regulations.

PREREQUISITES

Required: Students are expected to have proficiency in designing and writing software programs. There is no programming language requirement for this class, though experience with object orientation is beneficial.

Recommended: Students should be comfortable with learning about basic statistics, data structures, and algorithm analysis. When appropriate, quantitative and computational methodology will be reviewed. Knowledge of, and prior

experience with, security principles is NOT a prerequisite for this course.

GRADING

Criteria	Percent of Grade
Project	50%
<i>(Initial Proposal)</i>	<i>(10%)</i>
<i>(Status Report)</i>	<i>(10%)</i>
<i>(Final Report & Presentation)</i>	<i>(30%)</i>
Homework Assignments (3 assignments, 10% each)	30%
Reading Summaries	10%
Class Participation	10%
	100%

Required Reading Assignments: There is no primary textbook for this course. Reading assignments will be selected from various periodicals. Students will be required to read and submit brief summaries of assigned readings.

Project: In lieu of a final exam, each student must complete an independent project on a data privacy issue in biomedicine. Projects should investigate a topic of interest to the student, and must demonstrate analysis and critical thinking in data privacy. The project will require a significant commitment and contribute to a substantial part of the final grade. A list of sample project topics will be made available and reviewed in class.

SCHEDULE (*Tentative and Subject to Change*)

Week 1: Course Overview and Introduction to Data Collection and Privacy

What is data privacy? How does it relate to data security principles, such as authorization, access control, and authentication? What are the legal and policy precedents for privacy in modern healthcare environments and society? Who collects medical information and when do patients have control over their privacy? Can policy and specification of privacy protections be automated?

Potential Readings:

- *National Research Council. For the Record: Protecting Electronic Health Information. National Academy Press. Washington, DC. 1997. Chapters 1 & 2.*
- *U.S. Code of Fair Information Practices. 1973.*
- *Gostin L. Health care information and the protection of personal privacy: ethical and legal considerations. 1997; 127(5, Pt. 2): 683-690.*

Week 2: De-identification, Uniqueness, and Re-identification

Many privacy regulations and policies protect patient privacy through the “de-identification” of data. This week, we will look into what de-identification entails and how it relates to “anonymity”. Students will learn how to characterize uniqueness in data, both at elemental and population levels of granularity.

Potential Readings:

- *Summary of the HIPAA Privacy Rule. <http://www.hhs.gov/ocr/privacysummary.pdf>*

- *Latanya Sweeney. Simple demographics often identify people uniquely. Data Privacy Laboratory Working Paper LIDAP-3, Carnegie Mellon University. 2000.*
- *Ohno-Machado L, Silveira P, Vinterbo S. Protecting patient privacy by quantifiable control of disclosures in disseminated databases. International Journal of Medical Informatics, 2004; 73 (7-8), 599-606.*

Week 3: Availability of Personal Information and Identifiers

Personal information is available in many different resources both on- and offline. Where is this information? How do we automatically capture and organize it for privacy assessments? This week we will look at various information repositories, such as vital records and statistics (including birth records, death records, marriage records, court documents) and the Social Security Death Index. We will also discuss issues such as the potential of unique numbers for persistent patient identifiers and the history of the Social Security Number.

Potential Readings:

- *Robert Ellis Smith. Ben Franklin's Website: Privacy and Curiosity from Plymouth Rock to the Internet. Providence: Privacy Journal. 2000. Chapters 12 - 14 (Numbers, Databanks, and Cyberspace).*
- *Simson Garfinkel. Database Nation. New York: O'Reilly. 2000. Chapter 3 (Absolute Identification), Chapter 8 (Who Owns Your Information).*

Week 4: Record Linkage

This section of the course will present concepts and methodology associated with the linkage of data in disparate databases. Methods will be drawn from deterministic and probabilistic frameworks. We will also discuss how linkage methods can be automated.

Potential Readings:

- *Fellegi I and Sunter. A theory for record linkage. American Statistical Association Journal. 1969; 64: 1183-1210.*
- *Winkler, W. E. Matching and Record Linkage. In B. G. Cox et al., ed. Business Survey Methods, New York: J. Wiley, 355-384. Online: <http://www.fcsm.gov/working-papers/wwinkler.pdf>.*
- *Grannis SJ, Overhage JM, Hui S, McDonald CJ. Analysis of a probabilistic record linkage technique without human review. . Proceedings of the American Medical Informatics Association Annual Symposium. 2003: 259-363.*
- *Grannis S, Overhage J, McDonald C. Analysis of identifier performance using a deterministic linkage algorithm. Proceedings of the American Medical Informatics Association Annual Symposium. 2002: 305-309.*

Week 5: Trails and Graph-Based Approaches to Privacy

People leave information behind in many different organizations. Simple methods of data protection and de-identification appear sufficient to protect information from privacy compromise. However, simple automated strategies can be constructed to link information across databases. This week will discuss the evolution of the “trail” re-identification attack, which will be used to illustrate a formal model of data re-

identification. At this point, we will explore how graph-based modeling can be applied to represent privacy problems in general.

Potential Readings:

- *Bradley Malin. Betrayed by my shadow: compromising privacy with trail matching. Journal of Privacy Technology. 2005.*
- *Bradley Malin and Latanya Sweeney. How not to protect genomic privacy in a distributed environment. Journal of Biomedical Informatics. 2005; 37(3): 179-192.*

Week 6: Text Scrubbing

A large amount of information collected in biomedical setting is in the form of free text: doctor's notes, laboratory reports, discharge reports, and more. How can we de-identify text information? Can we ever achieve "anonymized" text? This section of the course will review various methods and software for the discovery and replacement of personal identifiers.

Potential Readings

- *Sweeney L. Replacing personally-identifying information in medical records, the Scrub system. In Proceedings of the 1996 American Medical Informatics Association Annual Fall Symposium. 1996: 333–337.*
- *Ruch P, Baud RH, Rassinoux AM, Bouillon P, Robert G. Medical document anonymization with a semantic lexicon. In Proceedings of the 2000 American Medical Informatics Association Annual Fall Symposium. 2000; 729-733.*
- *Berman JJ. Concept-match medical data scrubbing. Archives of Pathology and Laboratory Medicine. 2003; 127 (6): 680–686.*
- *Gupta D, Saul M, Gilbertson J. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. American Journal of Clinical Pathology. 2004; 121: 176–86.*

Week 7: Formal Models of Anonymity

This week will look into formal models of anonymity protection, such as k -map and k -anonymity. We will also look at ways in which formal models can be satisfied through computational transformations of data, such as generalization, suppression, and aggregation. We will study the computational complexity of the anonymizing data with minimal changes to the data and review various heuristics and strategies for data protection.

Potential Readings

- *Sweeney L. k -anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.*
- *Sweeney L, k -anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 571-588.*
- *Machanavajjhala A, Gehrke J, Kifer D, Venkatasubramanian M. l -diversity: privacy beyond k -anonymity. ACM Transactions on Knowledge Discovery from Data. 2007; 1(1): a3.*

- Li N, Li T, Venkatasubramanian S. *t-Closeness: Privacy Beyond k-Anonymity and l-Diversity*. In *Proceedings of the 23rd IEEE International Conference on Data Engineering*. 2007.
- Ohrn A, Ohno-Machado L. *Using Boolean reasoning to anonymize databases*. *Artificial Intelligence in Medicine*. 1999; 15(3): 235-254.

Week 8: Privacy Preserving Biosurveillance and Geospatial Information

Public health and epidemiology require geographic information regarding the presence of clinically interesting cases to detect potential outbreaks and bioterrorist activities. However, the sharing of geographic and spatiotemporal information may lead to re-identification. This week we will investigate various approaches by which such information may be protected during data sharing.

Potential Readings

- Curtis A, Mills J, Leitner M. *Keeping an eye on privacy issues with geospatial data*. *Nature*. 2006; 441, 150.
- Brownstein JS, Cassa CA, Mandl KD. *No place to hide - reverse identification of patients from published maps*. *New England Journal of Medicine*. 2006; 355(16): 1741-1742.
- Olson KL, Grannis SJ, Mandl KD. *Privacy protection versus cluster detection in spatial epidemiology*. *American Journal of Public Health*. 2006; 96(11): 2002-2008.
- Cassa CA, Grannis SJ, Overhage JM, Mandl KD. *A context-sensitive approach to anonymizing spatial surveillance data: impact on outbreak detection*. *Journal of the American Medical Informatics Association*. 2006;13(2):160-165.
- Gedik B and Liu L. *Location privacy in mobile systems: a personalized anonymization model*. In *Proceedings of the 25th International Conference on Distributed Computing Systems*. 2005; 620-629.

Week 9: Privacy in Biological and Genomic Databases

As high-throughput technologies become further ingrained in the clinical environment, the collection and sharing of biological information, such as DNA data, is becoming more common. In this section of the course, we will investigate ways in which patient identity in genomic data is protected, how it is re-identified and how it can be formally protected.

Potential Readings

- National Institutes of Health. *Request for Information (RFI): Proposed Policy for Sharing of Data obtained in NIH supported or conducted Genome-Wide Association Studies (GWAS)*. NOT-OD-06-094. August 30, 2006.
<http://grants.nih.gov/grants/guide/notice-files/not-od-06-094.html>
- Malin B. *An evaluation of the current state of genomic privacy protection technologies and a roadmap for the future*. *Journal of the American Medical Informatics Association*. 2005;
- McGuire AL, Gibbs RA. *Genetics. No longer de-identified*. *Science*. 2006 Apr 21;312 (5772): 370-371.

- Lin Z, Hewett M, Altman R. Using Binning to Maintain Confidentiality of Medical Data. In *Proceedings of the 2002 American Medical Informatics Association Annual Symposium*. 2002: 454-458.
- Lin Z, Owen A, Altman O. Genomic research and human subject privacy. *Science*. 2004 Jul 9; 305 (5681): 183.
- Malin B. Protecting genomic sequence anonymity with generalization lattices. *Methods of Information in Medicine*. 2005; 44(5): 687-692.

Week 10: Privacy in Population-Based Research / Project Status Report Presentations

The first lecture of this week will be dedicated to data privacy issues associated with the collection and sharing of information for population-based investigations. Does this information pose a threat to privacy and if so, how can we formally prevent such a threat? The second lecture of this week will be dedicated to student projects. Students will write a short summary of their problem statement, initial research design, and make a short presentation on the status of their projects for an in-class evaluation.

Potential Readings

- Lako CJ. Privacy protection and population-based health research. *Social Science and Medicine*. 1986; 23(3): 293-295.
- Botkin J, McMahon W, Smith K, Nash J. Privacy and Confidentiality in the Publication of Pedigrees. *Journal of the American Medical Association*. 1998; 279(22): 1808-1812.
- Malin B. Re-identification of familial database records. *Proceedings of the 2006 American Medical Informatics Association Annual Symposium*. 2006; 524-528.
- Gulcher JR, Kristjánsson K, Gudbjartsson H, Stefánsson K. Protection of privacy by third-party encryption in genetic research in Iceland. *European Journal of Human Genetics*. 2000 Oct; 8(10):739-742.

Week 11: Private Record Linkage and Secure String Comparison

How can we integrate information on people with revealing their identifying information. This question has been studied for years in biomedical environments and beyond. Various solutions have been proposed, including one-way hashing, keyed encryption, and oblivious transfers. In this week's lectures, we will review how obscured personal identifiers can be compared.

Potential Readings

- Churches T, Christen P. Some methods for blindfolded record linkage. *BMC Medical Informatics and Decision Making*. 2004 Jun 28; 4: 9.
- O'Keefe CM, Yung M, Gu L, and Baxter R. Privacy-preserving data linkage protocols. *Proceedings of the ACM Workshop on Privacy in the Electronic Society*. 2004: 94-102.
- Quantin C, Bouzelat H, Allaert FA, Benhamiche AM, Faivre J, Dusserre L. Automatic record hash coding and linkage for epidemiological follow-up data confidentiality. *Methods of Information in Medicine*. 1998 Sep; 37(3): 271-277.

- *Berman J. Zero-check: a zero-knowledge protocol for reconciling patient identities across institutions. Archives of Pathology and Laboratory Medicine. 2004 Mar; 128(3): 344-346.*
- *Agrawal R, Asonov D, Kantarcioglu M, Li Y. Sovereign joins. In Proceedings of the 22nd IEEE International Conference on Data Engineering. 2006: 26.*

Week 11: Secure Multiparty Computation and Its Applications to Distributed Data Mining

The traditional application of cryptography is framed from a two-party viewpoint in which two participants, Alice and Bob, exchange information, such as a patient's medical record, over an unsecured channel. An extension to the traditional model is secure multiparty computation (SMC), which is concerned with the interaction of two or more participants that need to exchange information to construct a result without revealing private information. This week we will look at how secure multiparty computation is used in the context of data mining across a set of data holders with different privacy constraints.

Potential Readings

- *Kantarcioglu M and Clifton C. Privacy-preserving distributed mining of association rules on horizontally partitioned data. IEEE Transactions on Knowledge and Data Engineering. 2004; 16(9): 1026-1037.*
- *Vaidya J and Clifton C. Secure set intersection cardinality with application to association rule mining. Journal of Computer Security. 2005; 13(4): 593-622.*
- *Zhong S, Yang Z, Wright R. Privacy-Enhancing k-anonymizations of customer data. Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. 2005; 139-147.*

Week 12: Image and Video Privacy

Video is increasingly used for monitoring and surveillance in health care environments, such as managed care facilities. This week we will investigate several procedures and principles for removing personally identifying features, e.g., an individual's face, from video streams. We will also investigate how images, e.g., the picture of a face, are a special case of video streams, can be protected using formal models of anonymity.

Potential Readings

- *Newton E, Sweeney L, Malin B. Preserving privacy by de-identifying Facial Images. IEEE Transactions on Knowledge and Data Engineering. 2005; 17(2): 232-243.*
- *Gross R, Airoidi E, Malin B, and Sweeney L. Integrating utility into face de-identification. Lecture Notes in Computer Science: Proceedings of the 5th Privacy Enhancing Technologies Conference. Berlin: Springer. 2005; 3856: 227: 252.*
- *Senior A, Pankanti S, Hampapur A, Brown L, Tian Y, and Ekin A. Enabling video privacy through computer vision. IEEE Security & Privacy. 2005; 3(3):50-57.*
- *Schiff J, Meingast M, Mulligan D, Sastry S, and Goldberg K. Respectful cameras: detecting visual markers in real-time to address privacy concerns. In Proceedings of the International Conference on Intelligent Robots and Systems. 2007.*

Week 13: Trail Anonymization

In this section of the course, students will learn how to integrate formal privacy models with private record linkage to thwart the re-identification problems. To demonstrate this approach, we will study how the trail re-identification problem can be thwarted.

Potential Readings

- *Malin B. A computational model to protect patient data from location-based re-identification. Artificial Intelligence in Medicine. 2007: doi:10.1016/j.artmed.2007.04.002.*
- *Malin B. A secure protocol to distribute unlinkable health data. In Proceedings of the 2005 American Medical Informatics Association Annual Symposium. 2005: 485-489.*
- *Malin B, Airoidi E. The effects of location access behavior on re-identification risk in a distributed environment. Lecture Notes in Computer Science: Proceedings of the 6th Privacy Enhancing Technologies Conference, Revised Selected Papers. Berlin: Springer. 2006; 4258: 413-429.*

Week 14: Privacy Preserving Database Integration and Querying / Final Presentations

In the first lecture we will study how to integrate data from a clinical network without revealing where the data came from. This section looks at several architectures for data matching and query execution that leverage various data mining and database-centric approaches. The final lecture will be dedicated to students' presentations on

Potential Readings

- *Clifton C, Doan A, Elmagarmid A, Kantarcioglu M, Schadow G, Suciu D, Vaidya J. Privacy-preserving data integration and sharing. Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. 2004: 19-26.*
- *Privacy-Preserving Distributed Queries for a Clinical Case Research Network. IEEE Workshop on Privacy, Security, and Data Mining. 2002.*