

DISSERTATION
DEFENSE

Privacy Preserving Medical Data Publishing

James Gardner
Emory University

Abstract: There is an increasing need for sharing of medical information for public health research. Data custodians and honest brokers have an ethical and legal requirement to protect the privacy of individuals when publishing medical datasets. This dissertation presents an end-to-end Health Information DE-identification (HIDE) system and framework that promotes and enables privacy preserving medical data publishing of textual, structured, and aggregated statistics gleaned from electronic health records (EHRs). This work reviews existing de-identification systems, personal health information (PHI) detection, record anonymization, and differential privacy of multi-dimensional data. HIDE integrates several state-of-the-art algorithms into a unified system for privacy preserving medical data publishing. The system has been applied to a variety of real-world and academic medical datasets. The main contributions of HIDE include: 1) a conceptual framework and software system for anonymizing heterogeneous health data, 2) an adaptation and evaluation of information extraction techniques and modification of sampling techniques for protected health information (PHI) and sensitive information extraction in health data, and 3) applications and extension of privacy techniques to provide privacy preserving publishing options to medical data custodians, including de-identified record release with weak privacy and multidimensional statistical data release with strong privacy.

Monday, March 26, 2012, 12:00 pm
Mathematics and Science Center: E406

Advisor: Li Xiong

MATHEMATICS AND COMPUTER SCIENCE
EMORY UNIVERSITY