

CSE 598D / STAT 598B: Data Privacy

Course Announcement

Instructors: Aleksandra Slavkovic & Adam Smith

Time & Place: Thursday, 1-3:45pm, IST Building Room 333. Listed as CSE 598D, schedule # 947575 and STAT 598B, schedule # 947500.

Website: <http://www.cse.psu.edu/~asmith/courses/privacy598d/>

Prerequisites: Background in probability and statistics recommended. Undergraduates require permission of an instructor.

Syllabus: Data privacy has become a fundamental problem of the modern information infrastructure. Collections of personal and sensitive data, previously the purview of governments and statistical agencies, have become ubiquitous as database systems have become larger and cheaper. Increasing volumes of information are collected and archived by health networks, financial organizations, search engines, intrusion detection systems, social networking systems, retailers and other enterprises. The potential social benefits from analyzing these databases are enormous. The main challenge is to learn the properties of the databases as a whole while protecting the privacy of individual contributors. The ultimate goal is to allow for useful statistical inference while minimizing the disclosure of sensitive information, that is, to strike a balance between data utility and disclosure risk. The problem is variously known as “statistical disclosure limitation”, “privacy-preserving data mining”, “anonymization”, “private data analysis”, or simply “data privacy”.

The course may be of interest to students and faculty from

- computer science and statistics
- fields of application dealing with sensitive or private information such as social sciences and health sciences.

The focus of the course is a survey of statistical and computational techniques used in data privacy, with emphasis on methods from statistics, data mining, and cryptography. We may also discuss the legal, policy and economic issues that arise in the use of sensitive information. The exact topics will be determined in part by the students interests.

Topics of interest include:

- What is privacy? Defining the problem and teasing apart the multiple meanings of “privacy”

- Statistical disclosure limitation (SDL) methods & practices for tabular data, e.g., perturbation techniques, data swapping, cell rounding, cell suppression
- SDL methods & practices for microdata, e.g., synthetic data, remote servers, multiple-imputation methods
- Assessment & estimation of disclosure risk, e.g., Risk-Utility maps, Bayesian frameworks
- Relevant cryptographic tools: secure function evaluation (SFE) protocols, authentication, efficient SFE protocols for specific data mining tasks
- Privacy-preserving data mining techniques: pre- and post-randomization, output perturbation
- Algorithmic questions: can we apply data privacy techniques on massive databases? To what extent can decisions be automated?
- Reidentification and deanonymization techniques, e.g. record linkage methods, information fusion.
- Privacy issues in specific areas of application, e.g. e-commerce, healthcare, official statistics, social networks.
- Ethics and privacy: balancing homeland security, economic utility and the right to privacy

Evaluation Students will be evaluated mainly on class participation, reading assignments and a term project (including a presentation).