Introduction to Privacy and Statistical Disclosure
Fall 2014

Instructor: Rebecca C. Steorts, beka@cmu.edu
Visiting Assistant Professor
Carnegie Mellon University
Department of Statistics

Course webpage: `http://www.stat.cmu.edu/~rsteorts/`

Large data-sets of sensitive personal information are becoming increasingly common – no longer are they the domain only of census agencies: now hospitals, social networks, insurance companies, and search engines all possess vast amounts of sensitive data. These organizations face legal, financial, and moral pressure to make their data available to the public, but also face legal, financial, and moral pressure to protect the identities of the individuals in their dataset. Formal guarantees trading off privacy and utility are therefore of the utmost importance.

Unfortunately, the history of data privacy is filled with the failed attempts at data privacy. The most egregious of these attempts at anonymization simply scrub the dataset of obviously identifiable information, and release the remaining data in its complete form. Furthermore, there have been many failed attempts, which are called linkage attacks. Linkage attacks are where identifying information is associated with anoymous records by cross referencing them with some additional source of information.

Over this course, we define what is mean by statistical disclosure and introduce what a formal constraint of privacy should encode and entail based upon Delanius, 1977 and why this fails Dwork (2006). This led to the introduction of differential privacy, an information theoretic guarantee on the distribution from which the private outputs of an algorithm are generated, that is independent of whatever auxiliary information an attacker may have available to him. We will then cover methods that many agencies use for releasing personal data. Such methods include data swapping and creating synthetic data for a sample or population. The most common pre-tabular method of SDC for Census tables is data swapping on the microdata prior to tabulation where values of variables are exchanged between pairs of households. (Scholmo, Tudor, and Groom 2010). We will then discuss alternative methods for releasing sensitive micro data when data steward may not be able to protect confidentiality by suppressing/perturbing only a small fraction of values. This approach involves generating/releasing fully synthetic data (Rubin, 1993; Fienberg, 1994; Reiter, 2002, 2005b, 2009; Raghunathan et al., 2003; Reiter and Raghunathan, 2007; Hu and Reiter, 2014).

More advanced topics will be covered as time permits including record linkage, private blocking, and privacy preserving record linkage, just to name a few.

Prerequisites: Students are expected to be very familiar with R. All reports, and exams, and write ups. should be submitted in Latex .pdf format.

Required Texts:

The Algorithmic Foundations of Differential Privacy , Cynthia Dwork and Aaron Roth, (2014), Now. `http://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf`.

Highly Recommend Papers:

These will be posted on the course webpage.

Grading Policy:

Scribing	15%
Class Discussions	15%
Final Project	70%

Topics covered (which are subject to change)

- Introduction to Statistical Disclosure Control
  - Data Anonymization
  - What is disclosure?
  - How to measure disclosure risk (identity and attribute disclosure)
- Differential Privacy
- Data Perturbation Methods
  - Data Swapping
  - Synthetic Data
- Privacy Preserving Linkage (PPRL)
  - Record Linkage
  - Privacy Blocking
  - Specific Types of PPRL and Limitations

Course Policies: All assignments will be announced in class (along with the due date). It must be turned in at the beginning of the lecture on the due date. Late assignments will not be accepted.

Scribing is a form of taking notes. You will scribe once during the course and this will count as part of your final grade. Each scribe will write notes for 50 minutes. Please prepare one set of notes for scribing that will be uploaded to the course webpage for the course to view. Please use LaTex to prepare scribe notes, and please use the template file on the course webpage.[1] Two of

[1]If you are not familiar with Latex (please see http://www.latex-project.org/ for more information and downloading for your OS). This is a great way to write up reports and display mathematical equations and graphical plots.

you will be randomly chosen to scribe on the day of lecture and you will have one week to prepare the notes with your classmate. The combined scribed notes should be emailed to the instructor by before the start of the next lecture. You are not allowed to switch with other students on the day you are scheduled to scribe.

Makeup exams must be approved before the time of the exam and will be given only in case of medical or family emergencies (which must be appropriately documented). All work turned in for a grade must be entirely your own. This particularly relates to the final project.
Furthermore, you are responsible for everything from lecture. Do not depend on the course web page for announcements regarding due dates for homework, changes in schedules, etc. Such announcements will be made in class. Homework assignments will be uploaded to the course webpage along with course readings (please check here frequently for updates).

Cell phones should be turned off (or set on silent). Laptops are allowed when we are doing applied examples or labs in class, but otherwise should not be out or being used.