Introduction

Jingchen (Monika) Hu

Vassar College

Statistical Data Privacy

Outline

Outline

General info

Instructor: Jingchen (Monika) Hu - jihu@vassar.edu
RH 403

Lecture: Mondays 3:10-5:10pm
RH 308

Lab: Some lectures will be used as labs.

Office hours: Tuesday 9:00am-11:00am & Thursday 2:00pm-4:00pm, or by appointment.
Note that the first hour is in person in RH 403 and the second half is on Zoom at https://vassar.zoom.us/j/9108856270.

Required materials

Prerequisite: MATH 242 or any of the 300-level statistics courses, and interests in statistical data privacy

Readings: A collection of selected journal articles and manuscripts

Software: We will use the software R/RStudio for labs and project. Download R from *http://www.r-project.org/* and RStudio, from *https://www.rstudio.com/*

Webpage: Vassar's Moodle: tentative schedule, slides, references, datasets, R scripts...

Course topics

- Overview of synthetic data
- Introduction to Bayesian modeling
- Bayesian synthesis models: continuous, binary, categorical, and count
- Evaluation of synthetic data
  - data utility
  - disclosure risks
- Overview of differential privacy (DP)
  - basics
  - mechanisms
- DP for common statistics: descriptive statistics and regressions

Course components

- Two parts of material: synthetic data, differential privacy

Course components

- Two parts of material: synthetic data, differential privacy
- Each part is a combination of a selection of the following: readings, lectures, labs, homework, discussions (in-class and online), and project

Course components

- Two parts of material: synthetic data, differential privacy
- Each part is a combination of a selection of the following: readings, lectures, labs, homework, discussions (in-class and online), and project
- The course project (individual or in pair) is a final product, using either synthetic data method or differential privacy method to provide privacy protection of a dataset of your interest

Course components

- Two parts of material: synthetic data, differential privacy
- Each part is a combination of a selection of the following: readings, lectures, labs, homework, discussions (in-class and online), and project
- The course project (individual or in pair) is a final product, using either synthetic data method or differential privacy method to provide privacy protection of a dataset of your interest
- Weekly check-ins

Course project

- Find a dataset and demonstrate disclosure risks
- Design and implement protection methods
- Evaluate protection methods
- Write a paper and present the results

Course project examples from previous semester

- Protecting income information in the National Health Interview Survey (NHIS)
- Protecting individual's privacy (political opinion) in the AP VoteCast
- Protecting price and available days information of Airbnb listings in New York City
- Protecting individual's privacy (demographics) in the STEM Labor Force survey

Course project examples from previous semester

- Protecting income information in the National Health Interview Survey (NHIS)
- Protecting individual's privacy (political opinion) in the AP VoteCast
- Protecting price and available days information of Airbnb listings in New York City
- Protecting individual's privacy (demographics) in the STEM Labor Force survey

You can find recordings of these projects and other projects at (towards the end of the playlist): https://www.youtube.com/playlist?list=PL_lWxa4iVNt0XPY0E0MDuGhKvbq_767mr

Course project examples from previous semester

- Protecting income information in the National Health Interview Survey (NHIS)
- Protecting individual's privacy (political opinion) in the AP VoteCast
- Protecting price and available days information of Airbnb listings in New York City
- Protecting individual's privacy (demographics) in the STEM Labor Force survey

You can find recordings of these projects and other projects at (towards the end of the playlist): https://www.youtube.com/playlist?list=PL_lWxa4iVNt0XPY0E0MDuGhKvbq_767mr

Questions? Ideas to discuss and share?