Data Confidentiality and Statistical Disclosure Control
Problems with traditional approaches
and alternatives based on synthetic data
SURV 735
2 credits/4 ECTS
Spring 2019, Section 1

Instructor
Jörg Drechsler, joerg.drechsler@iab.de

Video lectures by Jörg Drechsler

Short Course Description
This course will provide a gentle introduction to statistical disclosure control with a focus on generating synthetic data for maintaining the confidentiality of the survey respondents. The first part of the course will introduce several traditional approaches for data protection that are widely used at statistical agencies. Some limitations of these approaches will also be discussed. The second part of the course will introduce synthetic data as a possible alternative. This part of the course will discuss different approaches to generating synthetic datasets in detail. Possible modeling strategies and analytical validity evaluations will be assessed and potential measures to quantify the remaining risk of disclosure will be presented. To provide the participants with hands on experience, all steps will be illustrated using simulated and real data examples in R.

Course and Learning Objectives
By the end of the course, students will...
• know which measures are typically taken by statistical agencies to guarantee confidentiality for the survey respondents if data are disseminated to the public.
• be aware of potential limitations of these measures.
• have a practical understanding of the concept of synthetic data.
• be able to judge in which situations the approach could be useful.
• know how to generate synthetic data from their own data.
• have a number of tools available to evaluate the analytical validity of the synthetic datasets.
• know how to assess the disclosure risk of the generated data.

Prerequisites
The statistical software R will be used for illustrations and for (some of) the homework assignments. Thus, knowledge of R is required to be able to complete the assignments. Some background regarding general linear modelling is expected. Familiarity with the concept of Bayesian statistics is helpful but not required.

Class Structure and Course Concept
This is an online course using a flipped classroom design. It covers the same material and content as an on-site course but runs differently. In this course, you are responsible for watching video recorded lectures and reading the required literature for each unit and then “attending” mandatory weekly one-hour online meetings where students have the chance to discuss the materials from a unit with the instructor. Just like in an on-site course, homework will be assigned and graded and there will be a final exam at the end of the course.

Although this is an online course where students have more freedom in when they engage with the course materials, students are expected to spend the same amount of time overall on all activities in the course – including preparatory activities (readings, studying), in-class-activities (watching videos, participating in online meetings), and follow-up activities (working on assignments and exams) – as in an on-site course. As a rule of thumb, for each credit offered by a course, students can expect to spend one hour per week on in-class activities and three hours per week on out-of-class activities over the span of a full 12-week term. This is a 1-credit course that runs for 4 weeks. Hence, the total average workload is about 12 hours per week.

Mandatory Weekly Online Meetings
Wednesday, 8:00 – 8:50 PM CET, starting March 20
Meetings usually start at 2:00 pm ET. However, the time changes depending on the time change in the USA and Europe. We have highlighted the different times in the syllabus in red. This also applies to homework deadlines, etc.

Meetings will be held online through Zoom. Follow the link to the meeting sessions on the course website on https://www.elms.umd.edu/. If video participation via Internet is not possible, arrangements can be made for students to dial in and join the meetings via telephone.

In preparation for the weekly online meetings, students are expected to watch the lecture videos and read the assigned literature before the start of the meeting. In addition, students are encouraged to post questions about the materials covered in the videos and readings of the week in the forum before the meetings (deadline for posting questions is Wednesday 8:00 AM EST/2:00 PM CEST).
Daylight saving time begins in the US on March 10, 2019 and clocks are turned forward 1 hour. Daylight saving time begins in Europe on March 31, 2019. Therefore, look carefully at the times of meetings and deadlines!

Students have the opportunity to use the Zoom meeting room set up for this course to connect with peers outside the scheduled weekly online meetings (e.g., for study groups). Students are encouraged to post the times that they will be using the room to the course website forum to avoid scheduling conflicts. Students are not required to use Zoom and can of course use other online meeting platforms such as Google Hangout or Skype.

Grading
Grading will be based on:
- 2 quizzes (worth 15% total)
- Participation in the weekly online meetings, engagement in discussions during the meetings and/or submission of questions via e-mail (10% of grade)
- Three homework assignments (45%)
- A final online exam (30% of grade)

Dates of when assignment will be due are indicated in the syllabus. There will be a grace period for late assignments (not for quizzes), but late assignments will be penalized according to the following rules:
- 1 day late: 10% off
- 2 days late: 25% off
- 3 days late: 50% off
- 4+ days late: no credit

Technical Equipment Needs
The learning experience in this course will mainly rely on the online interaction between students and the instructor during the weekly online meetings. Therefore we encourage all students in this course to use a web camera and a headset. Decent quality headsets and web cams are available for less than $20 each. We ask students to refrain from using built-in web cams and speakers on their desktops or laptops. We know from our experience in previous online courses that this will reduce the quality of video and audio transmission and therefore will decrease the overall learning experience for all students in the course. In addition, we suggest that students use a wire connection (LAN), if available, when connecting to the online meetings. Wireless connections (WLAN) are usually less stable and might be dropped.
Long Course Description
Statistical agencies and other data collecting institutions constantly face the dilemma between providing broad access to their data and maintaining the confidentiality of the individuals included in the collected data. To address this trade-off various statistical disclosure control (SDC) methods have been developed which help to ensure that no sensitive information can be disclosed based on the disseminated data. However, applying these methods usually comes at the price of information loss or potentially biased inferences based on the protected data.

This course will introduce the data protection strategies that are commonly used by statistical agencies and discuss their advantages and limitations. We will also briefly look at the computer science perspective on data privacy. We will discuss the differences to the SDC perspective and what the SDC community could learn from the approaches developed in computer science.

The main part of the course will focus on a relatively new approach to statistical disclosure control that has been implemented successfully for some data products recently: Generating synthetic data. With this approach statistical models are fitted to the original data and draws from these models are released instead of the original data. If the synthesis models are selected carefully, most of the relationships found in the original data are preserved.

You will learn about the general idea of synthetic data and the two main approaches for generating synthetic datasets. The close relationship to multiple imputation for nonresponse will also be discussed.

The quality of the synthetic data crucially depends on the quality of the models used for generating the data. Thus, the course will present various parametric and nonparametric modeling strategies in great detail.

The quality needs to be evaluated in two dimensions: (i) How well is the analytical validity preserved, i.e. how close are analysis results based on the synthetic data to results obtained from the original data? (ii) What is the remaining risk of disclosure for the released data?

Several strategies to measure these two dimensions will be introduced. All steps of the synthesis process from generating the data, over analyzing the data, to evaluating the analytical validity and disclosure risk will be illustrated using simulated and real data examples in R.

Readings
All readings will be made available on the course website https://www.elms.umd.edu/.

Interested students might find the following additional recommended book helpful in preparing for the course:

Academic Conduct
Clear definitions of the forms of academic misconduct, including cheating and plagiarism, as well as information about disciplinary sanctions for academic misconduct may be found at

https://www.president.umd.edu/sites/president.umd.edu/files/documents/policies/III-100A.pdf (University of Maryland) and

Knowledge of these rules is the responsibility of the student and ignorance of them does not excuse misconduct. The student is expected to be familiar with these guidelines before submitting any written work or taking any exams in this course. Lack of familiarity with these rules in no way constitutes an excuse for acts of misconduct. Charges of plagiarism and other forms of academic misconduct will be dealt with very seriously and may result in oral or written reprimands, a lower or failing grade on the assignment, a lower or failing grade for the course, suspension, and/or, in some cases, expulsion from the university.

Accommodations for Students with Disabilities
In order to receive services, students at the University of Maryland must contact the Disability Support Services (DSS) office to register in person for services. Please call the office to set up an appointment to register with a DSS counselor. Contact the DSS office at 301.314.7682; http://www.counseling.umd.edu/DSS/.

Students at the University of Mannheim should contact the Commissioner and Counsellor for Disabled Students and Students with Chronic Illnesses at http://www.uni-mannheim.de/studienbueros/english/counselling/disabled_persons_and_persons_with_chronic_illnesses/.

Course Evaluation
In an effort to improve the learning experience for students in our online courses, students will be invited to participate in an online course evaluation at the end of the course (in addition to the standard university evaluation survey). Participation is entirely voluntary and highly appreciated.

Class Schedule
Please note that assignments and dates are subject to change. Information (e.g., articles and assignments) posted to the course website supersedes the information noted here.
Unit 1: A Brief History of Data Confidentiality & Traditional Approaches for Data Protection

Video lecture: available online Wednesday, March 13, 2019

Online meeting: Wednesday, March 20, 2019, 3:00 PM EDT / 8:00 PM CET

Daylight saving time begins in the US on March 10, 2019 and clocks are turned forward 1 hour. Daylight saving time begins in Europe on March 31, 2019. Therefore, look carefully at the times of meetings and deadlines!

Readings:

Recommended (optional):

Unit 2: The Computer Science Perspective on Data Privacy & Introduction to Multiply Imputed Synthetic Datasets

Video lecture: available online Wednesday, March 20, 2019

Online meeting: Wednesday, March 27, 2019, 3:00 PM EDT / 8:00 PM CET

First homework assignment: due Saturday, March 30, 2019, 6:59 PM EDT / 11:59 PM CET

Daylight saving time begins in the US on March 10, 2019 and clocks are turned forward 1 hour. Daylight saving time begins in Europe on March 31, 2019. Therefore, look carefully at the times of meetings and deadlines

Readings:

Recommended (optional):

**Unit 3: Analyzing Synthetic Datasets & Relationship to Multiple Imputation for Nonresponse**

Video lecture: available online Wednesday, March 27, 2019

Online meeting: Wednesday, April 3, 2019, 2:00 PM EDT/8:00 PM CEST

Quiz 1: due Saturday, April 6, 2019, 5:59 PM EDT/11:59 PM CEST

**Daylight saving time begins in the US on March 10, 2019 and clocks are turned forward 1 hour. Daylight saving time begins in Europe on March 31, 2019. Therefore, look carefully at the times of meetings and deadlines**

Readings:

Recommended (optional):

**Unit 4: Synthesis Models Part I (Univariate and Linear Regression Models)**

Video lecture: available online Wednesday, April 3, 2019

Online meeting: Wednesday, April 10, 2019, 2:00 PM EDT/8:00 PM CEST

Second homework assignment: due Saturday, April 13, 2019, 5:59 PM EDT/11:59 PM CEST

Readings:
Recommended (optional):

*** No online meeting on Wednesday, April 17, 2019 ***

**Unit 5: Synthesis Models Part II (Models for Categorical Variables and Nonparametric Models) & Modeling Strategies**

Video lecture: available online Wednesday, April 10, 2019

Online meeting: Wednesday, April 24 2019, 2:00 PM EDT/8:00 PM CEST

Quiz 2: due Saturday, April 27, 2019, 5:59 PM EDT/11:59 PM CEST

Readings:


Recommended (optional):

**Unit 6: Analytical Validity & Disclosure Risk Part I (Theory)**

Video lecture: available online Wednesday, April 24, 2019

Online meeting: Wednesday, May 1, 2019, 2:00 PM EDT/8:00 PM CEST

Third homework assignment: due Saturday, May 4, 2019, 5:59 PM EDT/11:59 PM CEST

Readings:

Recommended (optional):

**Unit 7: Disclosure Risk Part II (Examples in R) & Discussion of the Chances and Obstacles of the Synthetic Data Approach**

Video lecture: available online Wednesday, May 1, 2019

Online meeting: Wednesday, May 8, 2019, 2:00 PM EDT/8:00 PM CEST

Readings:


   Recommended (optional):

**Unit 8: Discussion of the Third Homework Assignment**

Video lecture: no video lecture

Online meeting: Wednesday, May 15, 2019, 2:00 PM EDT/8:00 PM CEST

Readings:

Final Exam Take Home

Final Exam due Wednesday, May 22, 2019, 5:59 PM EDT/11:59 PM CEST

Note: Student access to the course website will be revoked two weeks after the final exam.