Flexible Formal Privacy for Public Data Curation

Researchers rely extensively on public datasets disseminated by official statistics agencies, universities, non-governmental organizations, and other data curators. With the increasing availability of data and computing power comes increased threats to privacy, as published statistics can more easily be used to reconstruct sensitive personal data. Formal privacy (FP) methods, like differential privacy (DP), provably limit such information leakage by injecting carefully chosen randomized noise into published statistics. However, the way DP accounts for privacy degradation requires this noise be injected into every statistic dependent on the confidential dataset. This fails to reflect data curator needs, social, legal or ethical requirements, and complex dependency structures between public and confidential datasets. In this talk, I'll discuss statistical methodology that addresses these problems. We propose a FP framework with novel characterizations of disclosure risk when assessing collections of statistics wherein only some statistics are published with DP guarantees. We demonstrate FP properties maintained by our proposed framework, propose data release mechanisms which satisfy our proposed definition, and prove the optimality properties of downstream statistical estimators based on these mechanism outputs. For this talk, I'll discuss a few end-to-end data analysis examples in public health and surveys, showing how theoretical trade-offs between privacy, utility, and computation time manifest in practice when assessing disclosure risks and statistical utility. I'll conclude with a discussion on the implications of this work for survey researchers, focusing on opportunities to incorporate privacy by design in survey planning, experimental design, and other data collection operations.

Jeremy Seeman is a Michigan Data Science Fellow at the Michigan Institute for Data Science (MIDAS) and MPSDS. He recently graduated with his PhD in statistics from Penn State University. Jeremy's research focuses on statistical data privacy, quantitative methods in the social sciences, and social values in data governance. He is the recipient of the U.S Census Bureau Dissertation Fellowship and the ASA Pride Scholarship. Prior to joining Penn State, Jeremy completed his BS in Physics and MS in Statistics at the University of Chicago, where he was a research fellow at the Center for Data Science and Public Policy.