Chapter 15
Data Privacy: Outline

Ágoston Reguly and Zsigmond Pálvölgyi

- **Introduction**
  - Rise of data privacy in modern research and policy analysis. (legal mandates, respondent trust, and reproducibility pressures in official statistics and applied economics)
  - Overview of challenges when dealing with sensitive data. (linkage risk, auxiliary information, and the risk–utility tradeoff)
  - The two problem of confidentiality:
    1. *Revealable*: a legal/ethical obligation that microdata about persons/firms are used only for modeling purposes and not disclosed in identifiable form. In this case data is precisely measured, but limited for sharing.
    2. *Unrevealable*: persons/firms willingness to reveal sensitive information (e.g., income or health) for other parties (e.g., questionnaire). In this case data is not measured directly only through some process which allows some degree of anonymity.
  - Two different solutions for the two types:
    1. Data is available but sensitive – how to protect anonymity, while preserving statistical properties
       · Examples: data centres containing sensitive variables such as income, medical records, financial transactions, demographic data, etc. and want to use these for econometric modeling.
    2. Data is sensitive and cannot be directly observed – how to estimate econometric model parameters?
       · Examples: surveys and imprecise measuring variables such as measurement error or observing through intervals or (likert) scales.

Ágoston Reguly ✉
Corvinus University of Budapest, Budapest, Hungary and Georgia Institute of Technology, Atlanta, Georgia, USA, e-mail: agoston.reguly@uni-corvinus.hu

Zsigmond Pálvölgyi
Corvinus University of Budapest, Budapest, Hungary, e-mail: zsigmond.palvolgyi@uni-corvinus.hu

- **Data Available but Sensitive**
  - History of Privacy-Preserving Techniques
    - 1970s foundation: early problems with data confidentiality in statistical offices and early protocol: SDC (Statistical Disclosure Control). Earliest paper is Dalenius (1977), which articulates that releases should not increase what can be learned about any individual.
    - 1990s–2008: Lessons from early anonymization methods
      - After multiple data leakage (due to linkage attacks or quasi-identifiers), lesson learned that traditional methods fail as data growths, especially when multiple data tables are available from different sources.
      - Systematic critique of the early methods during the reidentification era: Sweeney's k-anonymity and empirical attacks; Netflix de-anonymization shows linkage defeats 'release-and-forget' methods.
    - 1990-2010s: Synthetic approach
      - New method which replace confidential values with (random) draws from statistical model fit and create synthetic releases. Analysts then can use standard econometric models with potential adjustments.
      - Common approaches include sequential regression or more complicated nonlinear models (tree-based methods, Bayesian/nonparametric schemes, etc.).
      - Lessons and criticism: the synthetic method is limited by providing verification for privacy measures. Another criticism is the synthetic approach is limited by the way the synthetic data is generated, and it assumes that the original statistical model is not misspecified.
    - 2006-nowadays: Differential Privacy
      - Concept and types of differential privacy (global vs local)
      - Different implementations of privatization.
      - Trade-offs: privacy vs. statistical accuracy.
      - Lessons: Why most of the differential privacy methods are not usable in econometrics – lack of consistency or problems in comparing different specifications.
- **Data Sensitive and Not Directly Observable**
  - 1950s–1975: Ordered choice models appear and spread.
    - The ordered probit/logit framework is introduced for ranked/categorical outcomes (Aitchison & Silvey, 1957; generalized in political and social sciences by McKelvey & Zavoina, 1975), establishing tools widely used in survey-based research.
  - 1970s–1990s: Recognition of measurement error/latency and its inferential consequences.
    - Work in econometrics and official statistics emphasizes that censoring, missingness, and latent constructs challenge point identification and standard inference. First approaches to treat the problem as a special measurement error.

  - 1980s–2000s: Maximum likelihood methods.
    - Extension of classical likelihood theory to incomplete information such as interval censored data. Following Turnbull (1976) nonparametric MLE, Groeneboom and Wellner (1992) and Finkelstein (1986) developed asymptotic theory and practical algorithms, including EM-based approaches. This further progressed ordered choice models and other models using interval censored data. Lessons: how to relax distributional assumptions.
  - 2000s: Breakthroughs in *partial identification*.
    - Manski & Tamer (2002) formalize inference when outcomes or regressors are observed only in intervals, deriving nonparametric bounds and set estimators; methods are informative but often yield wide sets without strong structure, which limits routine use. Manski's book (2003) consolidates partial identification; Chernozhukov–Hong–Tamer (2007) develop set estimators/inference for moment (in)equalities; Andrews–Soares (2010) and Andrews–Shi (2013) extend to conditional moments and generalized moment selection, power improves, but computation and communication of 'sets' remain barriers to wider adoption.
      Lessons: These breakthroughs are followed by disappearance of measurement-error approach and maximum likelihood based methods using interval censored data.
  - 2000s–2010s: Other approaches mature (simulation-based, Bayesian).
    - Simulation/incomplete-model analyses show how bounds can be informative without full structural specification; Bayesian methods for partially identified models and moment inequalities provide alternative inferential lenses. These methods gets popular as it uses external information to reduce the size of the identified set for the parameter(s) of interest, which is typically too wide in many application.
  - 2010s–2020s: Limits getting clearer, but...
    - Ordered/latent-index models require scale normalizations and distributional assumptions that can shift the estimand; set-ID methods may need large samples for economically tight bounds. These constraints started to be documented in methodological surveys or software notes, that explain why some tools remain specialized. However, there is still many competing solutions that are actually referring to different quantities.

- Future directions of not directly observable data
  - Collecting more and more data
  - Measurement error or rounding bias will be always present
  - When this really matter in empirics...

- Comparative Discussion
  - Strengths and weaknesses of both approaches.
  - When to use privacy-preserving techniques vs. latent variable models.
  - Modelling with sensitive variables – split sampling approach which combines both.

– Challenges with replication when used data in confidential.
– Future directions in privacy-aware econometrics.

Details

Introduction

Over the past two decades, privacy concerns have moved from a peripheral constraint to a central design principle in empirical research and policy analysis. Legal mandates and agency confidentiality pledges set hard limits on the disclosure of microdata, while respondent trust and ethical norms require that sensitive information be used only for legitimate analytic purposes. At the same time, pressure for data-owners to satisfy data privacy requirements and for applied economics to ensure econometrically consistent and asymptotically normal parameter estimates is increasing. These forces collide with modern data realities: linkage risk grows as auxiliary information proliferates across public and private sources, and every protection choice involves a risk–consistency trade-off, tightening privacy often reduces statistical accuracy, whereas loosening it threatens confidentiality.

This chapter distinguishes two core confidentiality problems and aligns methodological solutions to each. In the revealable confidentiality case, data are precisely measured but legally or ethically restricted from identifiable release; the task is to preserve anonymity while maintaining econometric validity, as in secure data centers analyzing income, medical records, transactions, or demographics with privacy-preserving procedures (e.g., noise addition, synthetic data, or guarded summaries). In the unrevealable confidentiality case, sensitive quantities cannot be observed directly as individuals disclose through ranges, Likert scales, or other masked mechanisms—so inference proceeds via models tailored to incomplete information (ordered choice, interval regression, and partial-identification tools that accommodate measurement error). The remainder of the chapter discusses lessons from these two tracks: what were the main tools, and if they flawed, why they have not worked properly. What are the current solutions and possible challenges. We emphasize two angles: compromising privacy and econometric rigour.

Data Available but Sensitive

Concerns about confidentiality in official statistics first crystallized in the 1970s, when statistical agencies adopted Statistical Disclosure Control (SDC) protocols to ensure that public releases would not increase what could be learned about any single person or firm. Those protocols—suppression, top-coding, cell perturbation, and record swapping—worked tolerably well when datasets were modest and external linkages rare. As data volume, dimensionality, and auxiliary sources exploded through the

1990s and 2000s, however, quasi-identifiers enabled linkage attacks that defeated 'release-and-forget' anonymization. The empirical re-identification episodes of that era made clear that coarsening and removal of direct identifiers are insufficient once adversaries can combine datasets at scale.

A natural response was the synthetic-data paradigm: fit statistical or machine-learning models to confidential microdata and then release draws from the fitted model in place of the original values. Synthetic releases preserve familiar workflows—regressions, tabulations, prediction—while mitigating direct disclosure. Yet two limitations proved central. First, privacy is only as strong as the modeling assumptions; with rich adversarial side information, synthesizers can leak structure. Second, misspecification risks propagate to downstream inference, often requiring multiple synthetic datasets plus diagnostics to stabilize estimates and communicate uncertainty.

Differential Privacy (DP) reframed the problem by providing formal guarantees: the presence or absence of any single record has a tightly controlled effect on the released output. In practice, DP forces an explicit trade-off between privacy budgets and statistical accuracy. For econometrics, this trade-off is consequential: privatization noise can undermine consistency, complicate model comparison across specifications, and require adjustments to standard errors and test statistics. The emerging remedy is not to abandon econometrics, but to adapt it, design releases around statistics with known sensitivity, articulate privacy budgets, and account for noise in estimation and inference.

Data Sensitive and Not Directly Observable

When variables are coarsened or masked—common for income, health, or sensitive behaviors—econometricians confront inference with incomplete information. Ordered choice models provided early tools by linking observed categories to a latent index via threshold crossing, enabling estimation of relationships in ranked outcomes. For interval-censored quantities, maximum-likelihood and nonparametric approaches developed throughout the late twentieth century supplied algorithms and asymptotic theory for estimation without exact values.

A decisive shift arrived with the partial identification program. Rather than forcing point identification through strong assumptions, researchers derived informative bounds under minimal structure for outcomes or regressors observed only in ranges. Set estimators and inference for moment inequalities broadened the toolbox, extending identification to complex settings and improving power via generalized moment selection. In practice, this movement redirected attention from conventional measurement-error fixes toward transparent, assumption-lean analyses.

Limits nevertheless persist. Latent-index models depend on scale normalizations and distributional choices that can change estimands, and set-identified procedures often require large samples to tighten economically meaningful bounds. Computation can be demanding, while interpretation is challenging. Complementary

strategies—simulation-based incomplete-model analyses and Bayesian treatments of partially identified parameters—use external information to shrink sets, but must be carefully justified to avoid importing unwarranted certainty.

Comparative Discussion and Future Directions

Privacy-preserving releases and latent-variable/partial-identification models address different confidentiality constraints and offer complementary strengths. When microdata are accurately measured, but cannot be disclosed, mechanisms such as differential privacy (DP), guarded summaries, or synthetic data provide formal protections and scalable workflows at the cost of injected noise that can complicate consistency, model comparison, and classical inference. When sensitive quantities are not directly observable (intervals, Likert scales, masked values), latent-index models, interval regression, and set-identified procedures respect the coarsened nature of the data, making assumptions explicit and avoiding overpromised precision, yet they may yield wide ranges without additional structure and can be computationally demanding. A practical bridge is split sampling: collect the most sensitive variables with multiple discretization scheme, while observing other covariates directly, and integrate them with joint models that propagate both privatization and coarsening uncertainty.

Confidential workflows also reshape eventual replications and future research priorities. Exact numerical reproduction is frequently not feasible once privacy budgets, randomized mechanisms, or access rules vary across time and teams; replication should emphasize documented pipelines, explicit privacy parameters, public synthetic companions with validation reports, and privacy-protected sufficient statistics that allow independent checks. Looking ahead, ‘privacy-aware’ econometrics will likely to benefit from optimal allocation of privacy budgets across competing specifications; hybrids that combine DP with partial identification to deliver safe yet informative bounds; systematic utility benchmarking across model classes; and reporting standards where privacy parameters and identification diagnostics accompany conventional outputs. Lessons from DP (sensitivity-aware design, explicit trade-offs) and latent-data models (assumption transparency, bound reporting) point to a unified practice that treats privacy and identification as design constraints. Open questions include how to compare models under heterogeneous privacy noise, how to fuse auxiliary information without eroding guarantees, and how to communicate uncertainty when both confidentiality and observability limit what can be learned.