Discussion of Talks on Bayesian Methods in Data Privacy

Jerry Reiter
Department of Statistical Science
Duke University

Remarks on Kazan's talk

- Formal privacy and the Census application
- Just because a release is DP does not mean an agency should disregard disclosure risk assessments
  - Large ε like those used in practical applications could come with disclosure risks
  - Two DP algorithms with the same ε could have different disclosure risk properties for different kinds of attacks
- We need statistical disclosure risk measures that account for formal privacy protections
  - Bayesian approaches like those presented by Kazan offer paths forward. But much work to be done….

Remarks on Kazan's talk, continued

- How to scale up to larger dimensions?
  - Need some approximate computational trick (and maybe parallel processing)
- How to account for features of Census application?
  - Post-processing and invariants complicate computation of posterior probabilities
- What does the work tell us about the effective ε?
  - ε = 17 is huge, but actual risks likely lower than implied

Remarks on Bowen's talk

- Generating synthetic data with formal privacy guarantees is appealing
- But, is DP the right criterion for synthetic data?
  - In one sense, yes
    - The community has not developed a better formal privacy metric
  - In another sense, no:
    - Complexities that DP does not yet know how to solve: imputation of missing values, correcting faulty values, survey weights
    - If agencies do not make every release formally private, do we really need features of DP like composability?
    - Do we need to protect the fact that a unit is in the sample?

Remarks on Bowen's talk, continued

- What does "statistically valid" mean?
  - Having no bias seems impossible, for any synthesis method
  - Is a better notion tied to fitness for use?
- Releasing multiple DP synthetic datasets
  - Enables multiple imputation type inferences, but uses privacy budget
  - Better to release multiple copies with reduced privacy budget for each, or to release a single copy?

Remarks on Quick's talk

- Use publically available data as prior information to enhance usefulness of DP algorithms
  - Great idea when possible
- How does one get valid inferences from the synthetic data when public data are used in synthesizer?
  - Some engine for posterior inference would be beneficial
- What does fact that information is public mean for disclosure risk?
  - Individual-level information may be public, too

Remarks on Quick's talk, continued

- Disclosure risk assessments often focus on risks that individuals can be re-identified/participated in a database
- For many government surveys and censuses, why should participation be confidential?
  - Does someone knowing that I participated in the ACS have any implication for my well being?
- Laws written in terms of re-identification, but perhaps those should be revisited…
  - An intruder knowing if someone is a member of a dataset may matter for some databases, but need it be the default?
- Perhaps agencies should focus more on attribute disclosures than re-identification disclosures

Final thoughts on presentations

I congratulate the presenters on their thought-provoking work, and I look forward to seeing their work develop further.

Thank you to the audience for attending our session.