Notes from wrap-up session

IPAM Workshop on Statistical and Learning-Theoretic Challenges in Data Privacy

February 26, 2010

John Abowd

- Goal:
  - Allow study of what are called "natural experiments" in applied economics

$$y_i = f(\beta, x_i, z_i)$$

    - i is the correct view (individual, other entity, search, job, *etc.*)
    - y is the correct outcome; x are data derived from the internal (relational) database; z are data linked from external sources that define the natural experiment (*e.g.*, special UI benefit eligibility; one-time subsidy for job search)
  - Given relational tables linked via several keys
  - Provide a set of primitives for analyzing tables privately, *e.g.*:
    - Analyst provides a model together with external variables linked to some of the same keys
    - Gets back approximate posterior predictive distribution for the model or posterior of model parameters given actual data, not synthetic data (or sampling distribution corrected for privacy including confidence bands)
  - Complete work environment would include error detection, edit, imputation, diagnostics
    - Did the external data integrate correctly?
    - Are there missing data in the system? Options for imputation
    - Does the model fit? Options for adjustment
  - The challenge is to do all of this within a formal privacy system without forcing the analyst to spend his entire privacy budget on data preparation

Katrina Ligett & Aaron Roth

- Can we use game theory as a tool for privacy?
  - Can I penalize people for improperly releasing/handling the data I give them?
    - See Golle, Mironov, McSherry, *Data Collection with Self-Enforcing Privacy*, *CCS* '06.
  - Can I incentivize correct answers?
  - [Frank M.: Can we price information leakage, e.g. dollars/epsilon used by PInQ? Frank keeps 10%.]
  - More mechanism design via diffe.p.? Pointer to Kobbi's talk.
    - Empirically, it seems easier to design diffe.p. algorithms than actually truthful ones

Kobbi Nissim

- Privacy budget
  - Needed: methodology for setting the privacy parameter
  - With simple composition, privacy parameters get eaten fast. What do we do? Better composition-style results?
    - Continual observation
    - Charge less for queries we already know answers to? (e.g. Roth-Roughgarden)
- Can we use crypto to increase our functionality?
  - Crypto "inside" the functionality to give better functionality privacy?
    - (non-)example: lattice hardness to release better subset sums?

Salil Vadhan

Worst- vs Average-case analysis

- Relaxing privacy seems tricky
  - Crypto history has shown that adversaries attack systems in unexpected ways
  - Can be risky to assume that adversary's uncertainty about database fits some model
- Relaxing utility much more natural, seems be to going already in works on learning & statistics
  - A good way to get around hardness results?

Theory vs Practice

- "differential privacy" ≠ "known differentially private algorithms"
- dialogue is valuable

Semantics of definitions (comment on Adam's remarks)

- When is protecting "local info" (as in diffe.p.) enough, and when is even "global info" too sensitive?
- Non-row-structured data, *e.g.* edge privacy vs. vertex privacy?
- Meaning of epsilon?

Adam Smith

- This week: the "Computational Lens" at work
  - Several results (McSherry, Nardi) inspired by numerical-analytic, rather than structural, approach
  - Salil Vadhan's talk on hardness
  - Big help in dialogue between stats and CS: express inference process algorithmically
- Can we have "cryptanalysis" for privacy in statistical databases?
  - Systematic study of attacks
  - Nomenclature (help to understand talks?!)
    - e.g. linkage, reconstruction, composition, …
  - Even an incomplete taxonomy is valuable
- Relaxing Definitions: Can we exploit *uncertainty* about the data?
  - Caveat: easy to lose the semantics of diffe.p. (see http://arxiv.org/abs/0803.3946)
  - What properties should be our guides? *e.g.*
    - Composition & resistance to side information
    - post-processing/ convexity (see Dan Kifer's talk)
  - No one size fits all (?)
- Exploiting *sparseness* for better diffe.p. algorithms?
- Optimizing compilers and other *automated* techniques for making better diffe.p. algorithms?
  - See poster by Li, Hay, Rastogi, Miklau, McGregor. http://arxiv.org/abs/0912.4742

Steve Fienberg

- Private analysis of graphs?
  - Consider model with three numbers per node (variant of "P1")
    - in-degree
    - out-degree
    - "reciprocity" (how correlated are in- and out-going edges on a per node basis?)
  - What's the right notion of privacy? Edge? Node? The latter seems more meaningful.
- For large scale databases we need:
  - Impossibility and complexity results to understand the limit of differential privacy.
  - Alternative approaches to differential privacy for such situations.
- Can we do more to to match the privacy protection method to the nature of the statistical problem:
  - This begins by cast utility in statistical terms.
  - Then we may want DP alternatives to Laplace noise that are tailored to the statistical output.

Cynthia Dwork

- GWAS
  - = genome-wide association study
  - It's important
    - Remember: NIH and Wellcome Trust forced statistics from studies they had funded to be taken off the web
  - **Official** people care
  - We can make an impact