Protection of Data Privacy via Differentially Private Multiple Synthesis

FANG LIU, UNIVERSITY OF NOTRE DAME

Cybersecurity Research Acceleration Workshop and Showcase, Indianapolis, IN
October 11, 2017

Differentially Private Data Synthesis Methods

Fang Liu
Applied and Computational Mathematics and Statistics
University of Notre Dame

Cybersecurity Research Acceleration
Workshop and Showcase
Indianapolis, IN
Oct 11, 2017

Supported by NSF 1546373 and NSF 1717417

Cybersecurity Research Acceleration Workshop and Showcase
October 11, 2017 | Indianapolis, IN

Protection of Data Privacy via Differentially Private Multiple Synthesis

challenge

Seek better ways to protect individual privacy in big data without compromising the accuracy of population-level information for research and public use

approaches and scientific merits

- bring a statistician's viewpoint to the utility question in differential privacy (DP).
- Develop innovative techniques and tools to create synthetic "surrogate datasets" that
  - have the same structure and similar statistical properties as the original dataset, but satisfying differential privacy.
  - are amenable to and accommodate various statistical analysis in real-life data
- Evaluates against both simulated data and real life studies Develops and release as open source tools for dataset creation.
- Leverages differential privacy and establishes an original framework to integrate DP with statistical modelling and inferences.

NSF BIGDATA: F: #1546373
Univeristy of Notre Dame
PI: Fang Liu (fang.liu.131@nd.edu)
Team: Claire Bowen, Evercita Eugenio, Yinan Li, Ashley Ahimbisibwe

broader impacts

- Helps to increase the efficiency of data collection and dissemination cycle, without concerns on individual information breach, leading to better decision making and more transformative discoveries based on data of higher quality
- Promote awareness of data privacy in the general public & stimulates interests in STEM careers among K-12 students.

research results and future plan

- Have developed
  - modips (model-based differentially private data synthesis /dips)
  - mwas (dips via multiplicative weighting with adaptive selection of queries)
  - SAFE (Statistical allocation of Epsilon)
  - GGM (generalized Gaussian mechanism)
  - Noninformative bounding
- Have compared some of the dips methods in data utility via empirical studies
- Have applied some of the methods in the Current Population Survey and the American Housing Survey
- Future: methods for better utility in released data, development of practical tools/software

Statistical Disclosure Limitation

- A collection of statistical approaches to protect data privacy
- Aims to provide protection for individual sensitive information when releasing data for research and public use.
- Data synthesis (DS) releases pseudo individual-level data
  - To reflect the uncertainty introduced during the synthesis process, multiple sets of synthetic data are often released (Rubin, 1993; Little, 1993)
  - Inferential methods are available to combine information from multiple synthetic data sets to yield valid inferences (Raghunathan et al., 2003; Reiter, 2003).
- Most statistical disclosure risk assessment approaches reply on strong and ad-hoc assumptions on the background knowledge and behaviors of data intruders

Differential privacy (DP)

A algorithm $\mathcal{A}$ is ϵ-differentially private if for all data sets (x, x') that is $\Delta(x, x') = 1$ and all possible result subsets Q to all queries q (Dwork et al; 2006)

$$\left| \log \left(\frac{\Pr(\mathcal{A}(q(\mathbf{x})) \in Q)}{\Pr(\mathcal{A}(q(\mathbf{x}')) \in Q)} \right) \right| \leq \epsilon$$

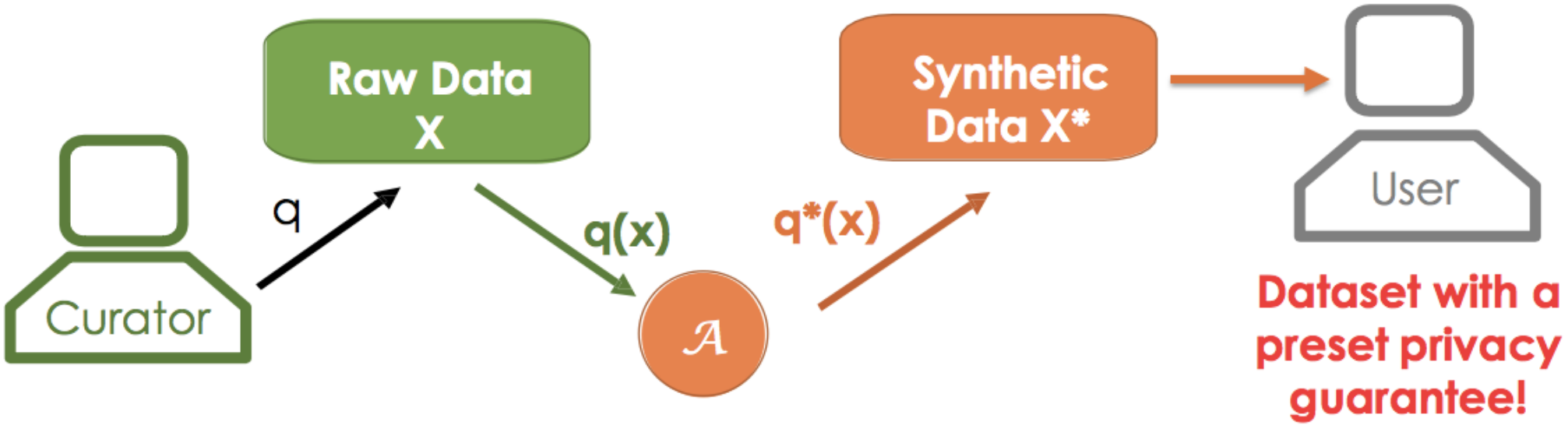
There are several versions of the relaxation of the "pure" -DP , such as the probabilistic DP (Machanavajjhala et al., 2008), approximate DP (Dwork et al., 2006), concentrated DP (Dwork and Rothblum, 2016).

Common differentially private mechanisms:

- Laplace mechanism
- Exponential mechanism
- Gaussian mechanism
- Others

Differentially Private Data Synthesis (DIPS)

Aims to integrate the DP concept into traditional multiple synthesis



1. How to choose q?
2. How to sanitize $q(X)$ with minimal info loss?
3. How to generate individual-level data X^* from $q^*(X)$?
4. How to account for uncertainty from the sanitization process to ensure valid inference?

DIPS methods we have developed

- modips (Bayesian model-based differentially private data synthesis) (Liu, 2016; Bowen and Liu, 2016)

  A Bayesian approach to sanitize sufficient statistics in the posterior distribution of model parameters to generate differentially private predictive posterior distributions from which differentially private individual-level data are generated.

- mwas (dips via multiplicative weighting with adaptive selection of queries) (Liu and Eugenio, 2017)

  An iterative approach to generate differentially private individual-level data by selecting the most deviant queries to update the weight for each individual in each iteration until convergence

- SAFE (Statistical allocation of Epsilon) (Bowen and Liu, 2017)

  The concept of optimizing the allocation of overall privacy budget among multiple queries via statistical tools to maximize information preservation in released data

- GGM (generalized Gaussian mechanism) (Liu, 2017)

  An extension of the Laplace mechanism and Gaussian mechanism

- Rules to combine info across multiple synthetic sets to yield valid inference (Liu, 2016)

Results

- We have compared different dips methods (nonparametric and modips) in several empirical data sets of various data types and showed no dips methods is universally the best for all data on all analysis.
- We have applied the SAFE to sequentially allocate ϵ to synthesize the the Current Population Survey data (2000 to 2012)
- We have applied to mwas method the American Housing Survey 2015 and showed mwas is better than the multiplicative weighting with the Exponential mechanism.

Next step

- Development of better and more dips methods that minimizes the information loss in synthetic data due to sanitization
- Robust and effective selection of easy-to-perturb queries
- Development of feasible practical tools and software