Differential Privacy and Statistical Inference

Vishesh Karwa

Affiliation: Harvard University and Carnegie Mellon University

Privacy Tools for Sharing Research Data

A National Science Foundation Secure and Trustworthy Cyberspace Project

with additional support from the Sloan Foundation and Google, Inc.

Sharing Social Network Data

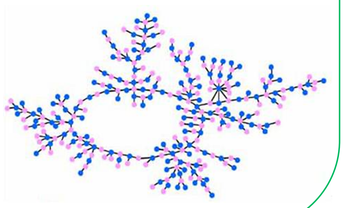
Goal: Enable sharing of social network data under rigorous privacy guarantees and maintain data utility for statistical inference.

Motivating Example - Epidemiological studies on sexual networks

→ Survey number of partners : Degree Sequence

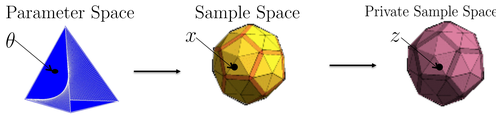
→ Estimate parameters, reconstruct typical networks, test hypothesis

Privacy Release "noisy" data
Utility Inference using "noisy" data



Network Example: Bearman, Moody, Stovel (2004). *American J. Sciology*.

Statistical Inference under Privacy Constraints



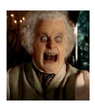
Statistical Model $\mathbb{P}(X;\theta)$ — Privacy Mechanism $\mathbb{P}(Z|X,\gamma)$

Statistician:

Estimation: Learning about θ

Inference: Hypothesis tests , e.g. $H_0: \theta = 0$

Adversary:



Learning about $X = x$

Learning if edge exists between i and j in $X = x$

Inference should be based on the Private Likelihood

$$P(Z;\theta,\gamma) = \sum_{x\in\mathcal{X}} P(Z|X=x,\gamma)P(X;\theta)$$

$$P(X=z;\theta) \neq P(Z=z;\theta,\gamma)$$

Exponential Random Graph Models (ERGM)

Any model-class for a network X can be parametrized in the form:

$$P_\theta(X=x) = \frac{\exp\{\theta\cdot g(x)\}}{c(\theta,\mathcal{X})}, \quad x\in\mathcal{X}$$

Besag (1974), Frank and Strauss (1986)

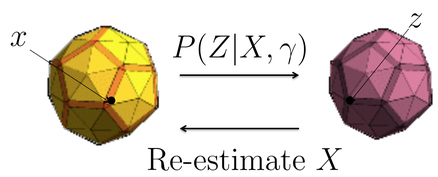
→ $\theta \in R^q$ a q-vector of parameters

→ $g(x)$ a q-vector of sufficient statistics

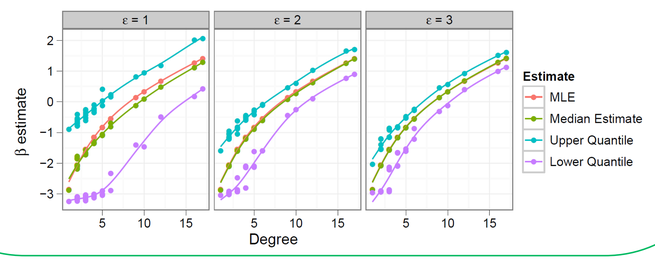
→ $c(\theta,\mathcal{X})$ distribution normalizing constant

The Degree Sequence ERGM

Theorem (Informal). *If $g(X)$ is the degree sequence, there exists an efficient differentially private estimator of θ that is also asymptotically optimal, in particular, consistent and asymptotically normal.*



Likoma	n=250, m = 248	Sexual network of individuals on Likoma Island
Karate	n = 34, m = 78	Network of Members of Karate club



Beyond Degree Sequences

→ What about other sufficient statistics?

→ What if we dont know what set of sufficient statistics are needed?

Randomized Response

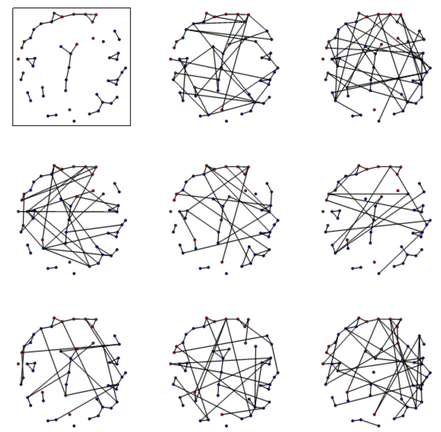
Old Wine in new Bottle

e.g. flip each edge randomly with probability γ

Differentially Private Estimator:

$$\hat{\theta}^{\epsilon}_{mle} = \text{argmax} \sum_{x\in\mathcal{X}} P(Z=z|X=x,\gamma)P(X;\theta)$$

Theorem (Informal). *The above differentially private estimator for any Exponential Random Graph model runs at only twice the computational cost of the non-private estimator.*



Conclusions

→ Sharing relationship data for reproducability and new scientific discoveries.

→ Design differentially private algorithms for statistical inference.

→ Privacy preseving inference (in some cases optimal) for random graph models.

→ Extensive experiments show that approach can work in practise.

References

→ Karwa, Krivistsky and Slavkovic, *Sharing Social Networks using Edge Differential privacy*, Forthcoming.

→ Karwa and Slavkovic, 2015, *Inference using noisy Degrees - Differentially private β model and synthetic graphs*, The Annals of Statistics.

→ Karwa, Slavkovic and Krivistsky, 2014, *Differentially Private Exponential Random Graph Models, Privacy in Statistical Databases.*

Vishesh Karwa
vishesh@seas.Harvard.edu
vishesh@cmu.edu

CRCS Center for Research on Computation and Society at Harvard School of Engineering and Applied Sciences

BERKMAN CENTER FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

The Institute for Quantitative Social Science at Harvard University

DATA PRIVACY LAB