March 4-8th 2019

Simons Institute for the Theory of Computing
Data Privacy: From Foundations to Applications

Statistical Disclosure Limitation: Where do we go from here?

Natalie Shlomo, University of Manchester

Natalie.shlomo@manchester.ac.uk

Topics

- Overview of types of disclosure risk in traditional forms of statistical data dissemination
- Common statistical disclosure limitation methods
- Risk-Utility Analysis
- Why Differential Privacy?
- Online flexible table builders and future dissemination strategies
- Discussion

Traditional Statistical Outputs

- Survey Microdata
  - Social surveys (census/register and business survey microdata generally not released)
  - Available from data archives for registered users
- Tabular Data

Frequency Tables	Magnitude Tables
Census/registers (whole population) counts	Business Statistics, eg., total turnover
Weighted sample counts	

Types of Disclosure Risks

Identity Disclosure

Identification is widely referred to in confidentiality pledges, legislation and codes of practice

Individual Attribute Disclosure

Confidential information about a data subject is revealed and can be attributed to the subject (Identity disclosure a necessary pre- condition)

Group Attribute Disclosure

Confidential information is learnt about a group and may cause harm

Common SDL Methods

Social Survey Microdata

Identity Disclosure (assume no response knowledge)- rare categories of identifying variables (population unique)

Recoding/grouping identifying variables, eg. k-anonymity

Suppressing variables such as high level geographies

Sub-sampling, eg. census samples

Attribute disclosure - individual(s) identified and survey target variables learnt, eg. health, income

Top-coding sensitive variables

Recoding / Microaggregation, eg. l-diversity

Common SDL Methods

Frequency Tables (whole population counts)

Identity Disclosure –small cells

Table design, eg. spanning variables and grouped categories

Minimum population thresholds

Attribute disclosure - zeros in row/column and one populated cell

Pre-tabular and/or post-tabular perturbation to introduce ambiguity in zero cells

Nested tables to avoid disclosure by differencing

Common SDL Methods

Magnitude Tables (Business statistics)

Assumptions:

- Intruders are competitors in the cell and can form coalitions
- Businesses in a cell are known
- The ranking of the businesses with respect to their size is known

Table design

Minimum population thresholds

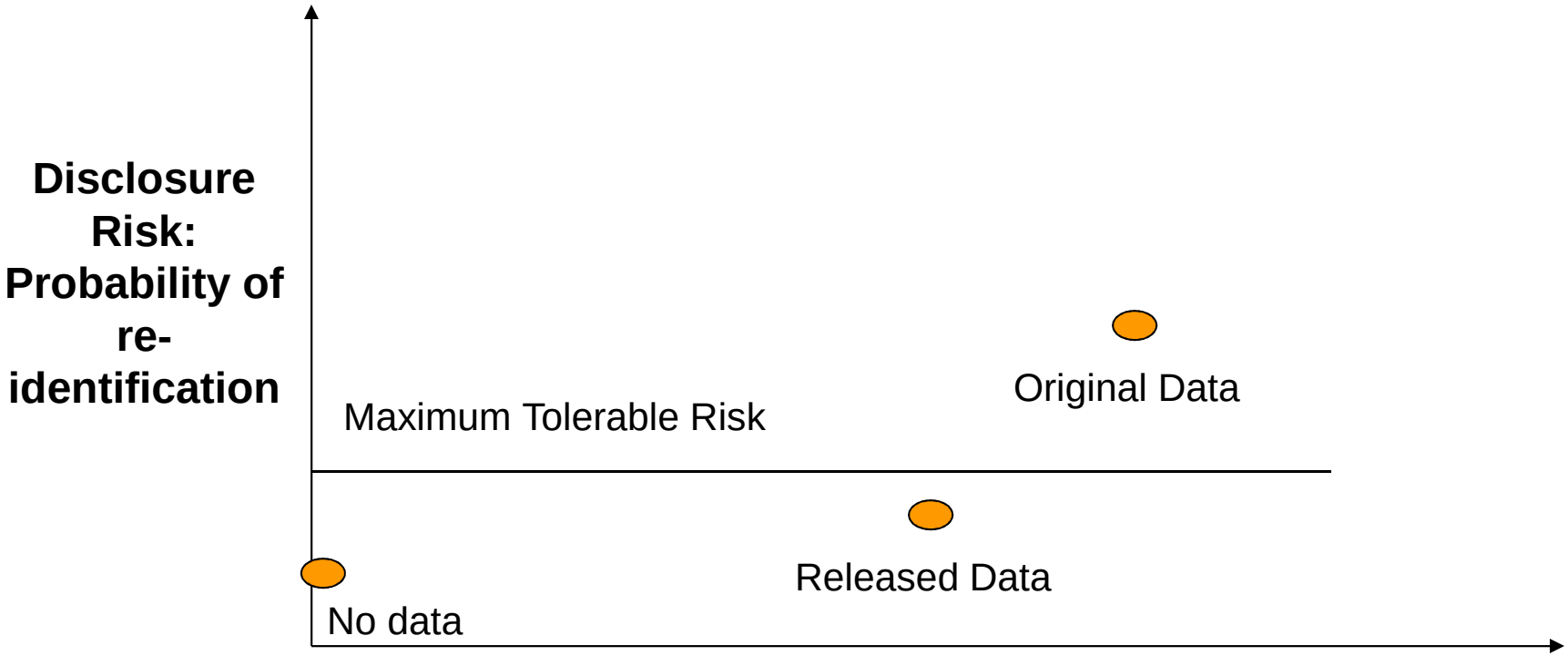
Cell suppression: primary and secondary

(mathematical programming and optimization)

Attribute disclosure - What can a competitor learn with sufficient precision

Disclosure Risk and Data Utility

R-U Confidentiality Map (Duncan, et.al. 2001)



Data Utility: Quantitative measure on the statistical quality

Inferential Disclosure

Confidential information may be revealed exactly or to a close approximation with high confidence from statistical properties of released and combined data

Examples:

Survey microdata – a good prediction model with very high R^2

Census tables – disclosure by differencing and linking tables

This type of disclosure has largely been ignored since it was mitigated through strict control of released data

- Microdata deposited in archives for registered users only and licensed
- Strict control of release of tabular data, eg. review boards for special request tabulations

Why Differential Privacy?

- Traditional forms of statistical data and their confidentiality protection rely heavily on assumptions that may no longer be relevant

Digitalization of all aspects of our society leading to new and linked data sources offering opportunities for research and evidence-based policies

With detailed personal information easily accessible from the internet, traditional SDC no longer sufficient and agencies relying more on restricting and licensing data

- More rigorous data protection mechanisms are needed with stricter privacy guarantees
- Collaborations with computer scientists through scientific programs

Mechanisms in Differential Privacy

Non-interactive Mechanism

Data custodian produces a 'safe' object, such as a synthetic database or collection of summary statistics

After this *release* all post-perturbative analyses are safe (no privacy budget spent after the original object)

Interactive Mechanisms

Data analyst sends queries (functions applied to a database) adaptively, deciding which query to pose next based on observed responses to previous queries

Accuracy will deteriorate with the number of questions asked, and providing accurate answers to all possible questions will be infeasible

To incorporate DP into the SDL toolkit, need to consider non-interactive mechanisms

Online Flexible Table Builder

- Flexible table generation for census tables (ABS, USA, EU)
  - Web-based platform (drop down lists) with restrictions: number of dimensions, population thresholds, no sparse tables
  - SDL on-the-fly: pre-tabular (hypercubes, data swapping – also known as input perturbation) and/or post-tabular methods (noise addition, rounding – also known as output perturbation)
- Risk of inferential disclosure since tables can be manipulated, differenced and linked
- Perturbation matrix **P** where $p_{ij} = p(\text{perturbed cell value is } j \mid \text{original cell value is } i)$
- For each cell count, change (or not change) the value according to probability p_{ij} and the outcome of a random draw

Properties of Table Builders

- Census counts non-negative integers, zeros not perturbed
- Dimension restrictions so if only 3 way tables are allowed out of 10 variables there are $\binom{10}{3} = 120$ possible tables
- Perturbations applied in advance (Fraser and Wooton (2006)):
  - Same participants-same perturbation principle - avoids repeated queries to same group having independent perturbations which could leak information
  - Same perturbation determined by microdata keys
  - In practice, perturbations determined when submitting queries based on a look-up table
- Perturbations capped (eg. to ± 7) and are unbiased with a constrained variance

Properties of Table Builders

- Margins perturbed separately and IPF for preserving additivity
- Additivity can be preserved in expectation by property of invariance where perturbation matrix satisfies: T**P** = T and T is vector of marginal totals
- Shlomo et al (2015) discuss SDC approaches and define disclosure risk and data utility measures for Table Builders based on information theory

Example: Risk measure $1-\frac{ent}{\log(K)}$ where K is the number of cells in row/column/table with value of 1 for a degenerate distribution and 0 for uniform distribution

Differential Privacy Algorithm

Rinott, Y., O'Keefe, C., Shlomo, N., and Skinner, C.(2018) Confidentiality and Differential Privacy in the Dissemination of Frequency Tables. *Statistical Sciences*, Vol. 33, No. 3, 358-385.

- List space $a=(a_1,...,a_k)$, eg. internal cells and margins (overlapping individuals) in a non-interactive mechanism
- Consider $M(.)$ such that $M(a)=b=(b_1,...b_k)$ where $p(b_k \mid a_k)$ set of conditional probabilities and cells perturbed independently (assume perturbed list has same structure as original list) and in our case, M is discrete
- Definition: $M(.)$ satisfies (ε, δ)-differential privacy if for all neighbouring lists a , a' differing by one individual:

$$P(M(a)=b) \leq e^{\varepsilon} P(M(a')=b)+\delta$$

and this is true for all potential lists and all possible outcomes

Differential Privacy Algorithm

- Exponential Mechanism (McSherry and Talwar, 2007) defined with respect to utility function *u* which assigns a utility score to possible perturbed values and the mechanism selected that produces values with high utility

- Define two loss functions:

$$l_1 = \sum_{i=1}^{K} | a_k - b_k |$$ (motivated by discretized Laplace)

$$l_2 = \sum_{i=1}^{K} (a_k - b_k)^2$$ (motivated by discretized Normal)

Then define $u_i = -l_i, \quad i = 1,2$

Exponential Mechanism

Exponential mechanism: given a, choose $b \in B$ (*B:* range of *b*) with probability proportional to: $e^{(\varepsilon/2)u/\Delta u}$ where

$$\Delta u = \max_{b \in B} \max_{a \sim a' \in A} |u(a,b) - u(a',b)|$$

Assuming additive loss functions and independent perturbations we can bound the perturbations: $|a_k - b_k| \le m, \forall k$ and this satisfies $DP(\varepsilon, \delta)$

Examples of Laplace perturbation vectors:

$\varepsilon = 1.5, \delta = 0.00002$

-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
0.00002	0.00008	0.00035	0.00157	0.00706	0.03162	0.14172	0.63516	0.14172	0.03162	0.00706	0.00157	0.00035	0.00008	0.00002

$\varepsilon = 0.5, \delta = 0.008$

-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
0.0076	0.0125	0.0206	0.0339	0.0559	0.0922	0.1520	0.2506	0.1520	0.0922	0.0559	0.0339	0.0206	0.0125	0.0076

Exponential Mechanism

Original Value	Range for ε=1.5 and δ=0.00002					Range for ε =0.5 and δ =0.008				
	± 0	± 1	± 2	± 3	± 4	±0	±1	±2	±3	±4
	Laplace m=7					Laplace m=7				
0	0.82	0.96	0.99	1	1	0.63	0.78	0.87	0.93	0.96
1	0.64	0.96	0.99	1	1	0.25	0.78	0.87	0.93	0.96
2	0.64	0.92	0.99	1	1	0.25	0.55	0.87	0.93	0.96
3	0.64	0.92	0.98	1	1	0.25	0.55	0.74	0.93	0.96
4	0.64	0.92	0.98	1	1	0.25	0.55	0.74	0.85	0.96
≥ 5	0.64	0.92	0.98	1	1	0.25	0.55	0.74	0.85	0.92
	Normal m=12					Normal m=10				
0	0.57	0.70	0.81	0.89	0.94	0.54	0.63	0.71	0.78	0.84
1	0.14	0.70	0.81	0.89	0.94	0.09	0.62	0.71	0.78	0.84
2	0.14	0.40	0.81	0.89	0.94	0.09	0.26	0.71	0.78	0.84
3	0.14	0.40	0.62	0.89	0.94	0.09	0.26	0.42	0.78	0.84
4	0.14	0.40	0.62	0.78	0.94	0.09	0.26	0.42	0.57	0.84
≥ 5	0.14	0.40	0.62	0.78	0.88	0.09	0.26	0.42	0.57	0.69

*Negative values to 0

Exponential Mechanism

Implications:

- DP leads to negative values, setting to zero still ensures DP but biased perturbations
- All (non-structural) zeroes must be perturbed
- If list-space has internal cells only $\Delta u = 1$, margins summed from internal cells DP but low utility
- In a *t*-way table all margins, $\Delta u = 2^t - 1$ (not including total) much larger perturbations implying smaller utility
- Margins can be perturbed (with appropriate sensitivity) and prorating to ensure additivity would still be DP

Exponential Mechanism

Implications (cont.):

- Same participants-same perturbation cannot be DP since differencing two tables that differ by one individual will reveal a true value
- Rule to be expanded to include domain totals

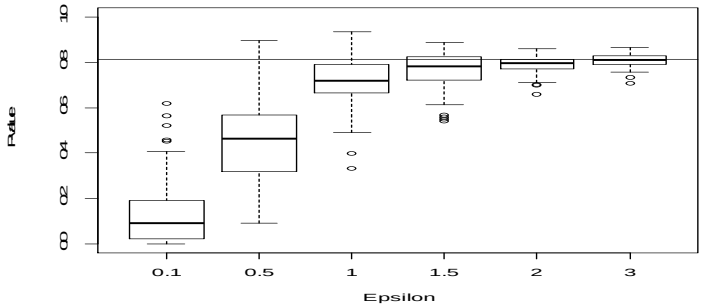
Application from UK Census and Simulation for independent tables:

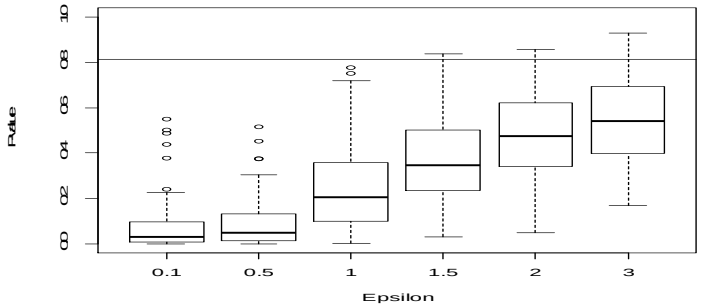
- Two tables (N = 10,000): large (1,000 cells - average cell size of 10) and small (100 cells - average cell size of 100)
- Tables have independent attributes
- 100 perturbations for Laplace and Normal mechanisms under different ε and cap set depending on same δ

Generated independent table, N=10000, K=100 (average cell size=100)

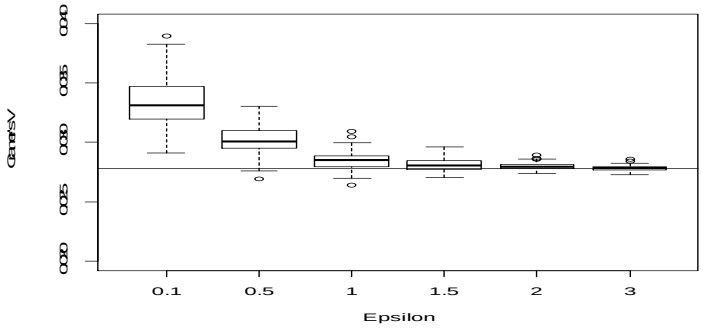
Laplace Perturbations **Normal Perturbations**

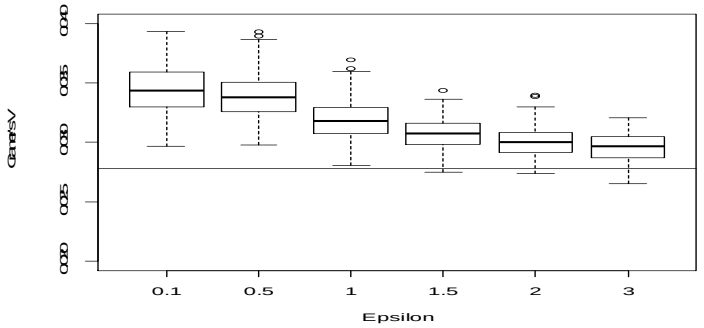
P-Value



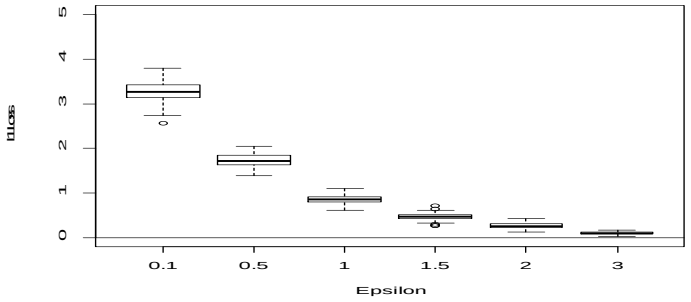


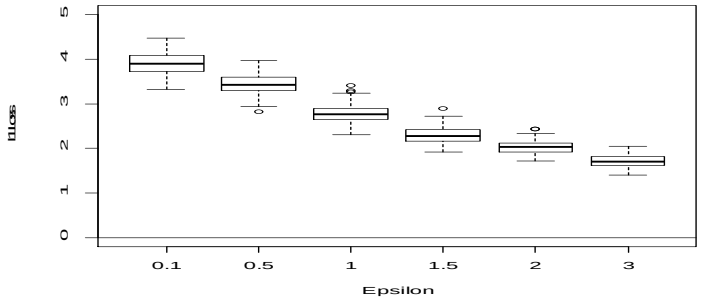
Cramer's V





***l_1* Loss Function**

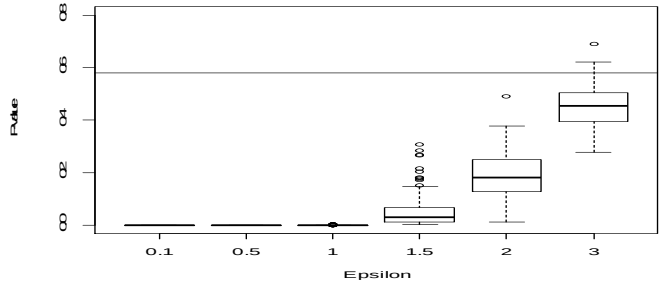


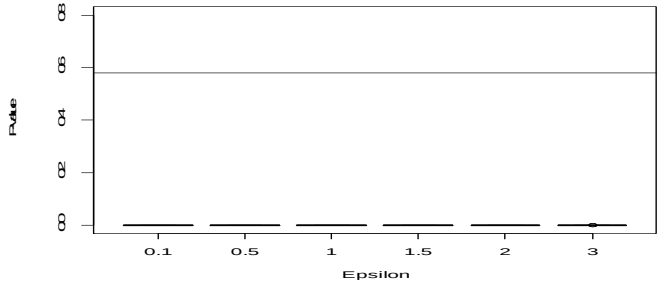


Generated independent table, N=10000, K=1000 (average cell size=10)

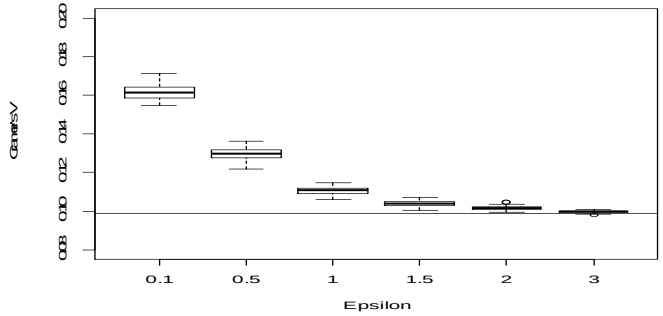
Laplace Perturbations **Normal Perturbations**

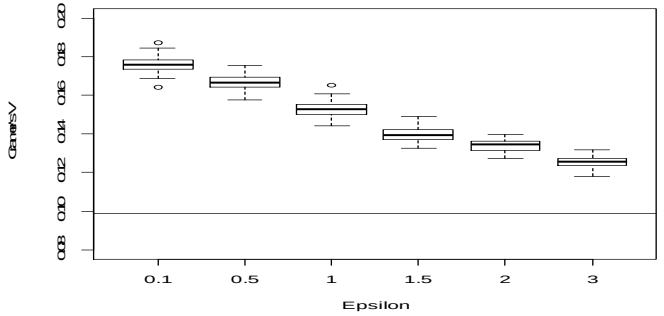
P-Value



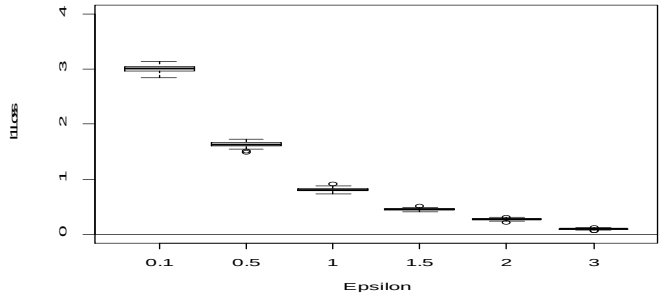


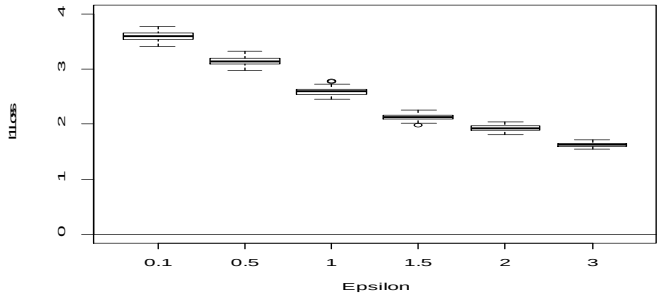
Cramer's V





***l_1* Loss Function**

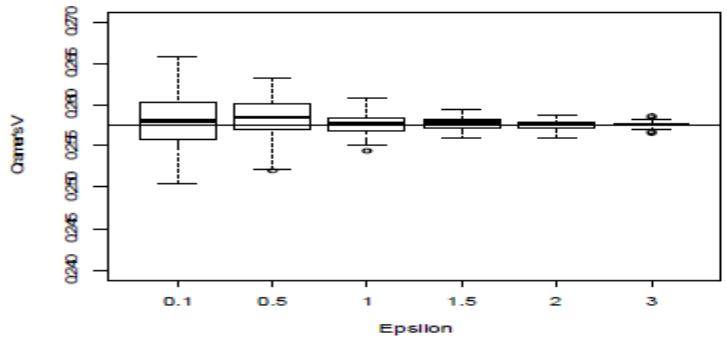


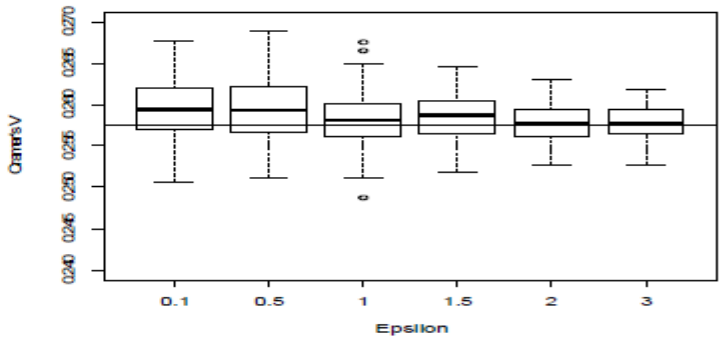


Real (dependent) Table from UK Census Data

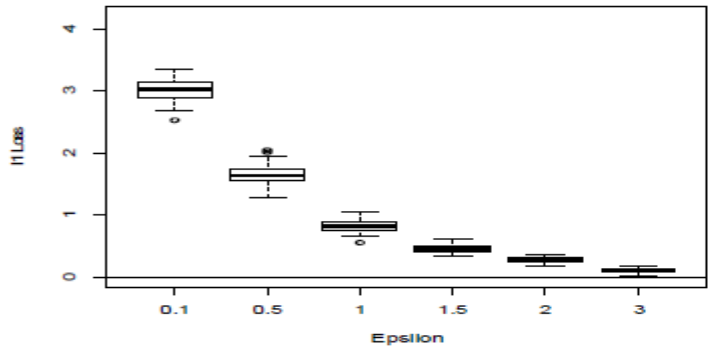
Laplace Perturbations | **Normal Perturbations**

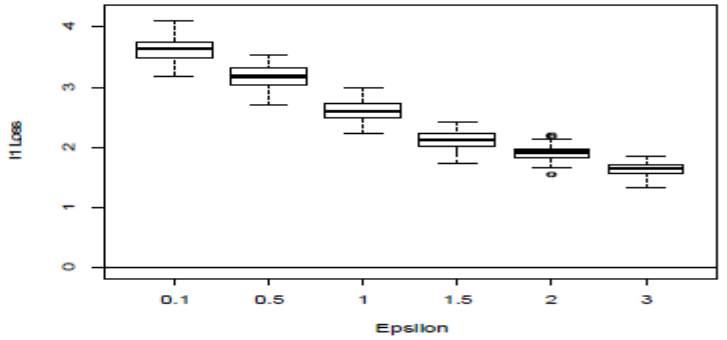
Cramer's V



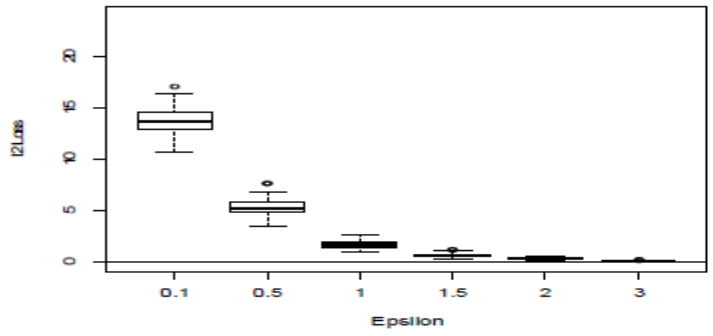


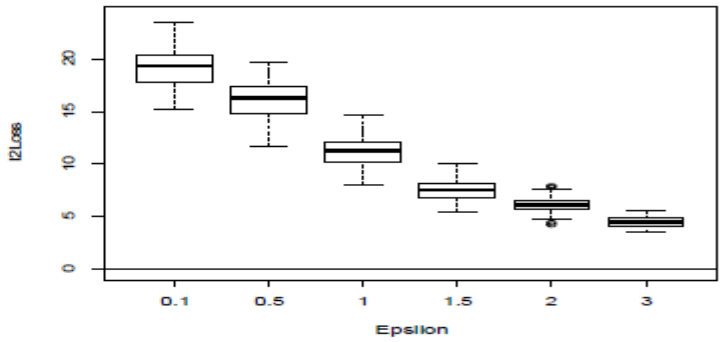
l_1 Loss Function





l_2 Loss Function





Complex Lists with Overlapping Cells

Composition Theorem:

Given M_i independent mechanisms under DP(ε_i, δ_i) then $(M_1 \ldots M_p)$ is DP $(\sum_{i=1}^{p} \varepsilon_i , \sum_{i=1}^{p} \delta_i)$

- Agencies have flexibility in deciding on the lists so amount of perturbation on different parts of the table with different ε's depending on interest in margins

Parameters of Differential Privacy not secret and can be used to adjust statistical analysis

Analysis under DP

- Testing for independence (Rinott, et al. 2018)

$$H_0 : \log(\mu_{ij}) = \eta + \alpha_i + \beta_j$$

- Adjust likelihood ratio test taking into account perturbation

$$L_X(\{\mu_{ij}\}) = \prod_{ij} \sum_{l_{ij}=-m}^{\min\{x_{ij},m\}} P_{\mu_{ij}}(x_{ij} - l_{ij}) e^{-\varepsilon|l_{ij}|}$$

where $\max_{\mu_{ij}} L_X(\{\mu_{ij}\}) / \max_{\log\mu_{ij}=\eta+\alpha_i+\beta_j} L_X(\{\mu_{ij}\})$ and $P_{\mu_{ij}} = e^{-\mu}\frac{\mu^x}{x!}$

- In the numerator, the max is over all μ_{ij} and in the denominator, the max over parameters η, α_i, β_j and $\alpha_1 = \beta_1 = 0$

Simulation: 10 x 10 tables (independent and dependent) with average cell size of about 50 (19 parameters to estimate in the denominator), perturb tables, repeat 1000 times

Analysis under DP

Table Type		Para-meters	% p-value ≤ 0.05	Mean (S.E.)		Para-meters	% p-value ≤ 0.05	Mean (S.E.)	
				Test Statistic	p-value			Test Statistic	p-value
Indepen-dent Attributes	Original	ε =0.1, m=10, δ=0.0283	5.0	81.6 (0.395)	0.487 (0.009)	ε=0.1, m=7, δ=0.0470	6.0	81.7 (0.423)	0.487 (0.009)
	Naive		86.7	124.5 (0.616)	0.027 (0.002)		53.3	105.1 (0.524)	0.123 (0.006)
	LR test		3.0	78.6 (0.388)	0.555 (0.009)		4.0	80.1 (0.400)	0.521 (0.009)
Dependent Attributes	Original		79.3	118.8 (0.587)	0.044 (0.003)		87.3	124.8 (0.620)	0.027 (0.003)
	Naive		99.6	162.1 (0.792)	0.001 (0.000)		98.3	149.2 (0.742)	0.004 (0.001)
	LR test		51.0	103.6 (0.526)	0.140 (0.006)		73.3	114.5 (0.583)	0.066 (0.004)
Indepen-dent Attributes	Original	ε=0.5, m=10, δ=0.0017	5.8	81.7 (0.414)	0.485 (0.009)	ε=0.5, m=7, δ=0.0076	4.7	81.4 (0.395)	0.485 (0.009)
	Naive		25.4	92.9 (0.476)	0.274 (0.008)		18.7	90.8 (0.435)	0.299 (0.008)
	LR test		6.9	82.5 (0.419)	0.467 (0.009)		5.3	82.0 (0.391)	0.473 (0.009)
Dependent Attributes	Original		82.1	118.3 (0.551)	0.041 (0.003)		81.3	119.6 (0.591)	0.040 (0.003)
	Naive		91.3	129.3 (0.601)	0.017 (0.002)		91.0	128.6 (0.627)	0.018 (0.002)
	LR test		76.3	114.8 (0.532)	0.054 (0.004)		76.9	116.2 (0.567)	0.051 (0.003)

Table Builder for Survey Weighted Counts

- Inferential disclosure from multiple data products: tables from original data and public-use files
  - $l_1 = \sum_{i=1}^{K} | a_k - b_k |$ a_k, b_k weighted survey counts and Δu is maximum weight
  - For survey weights with little variation (CV<20%) consider replacing weights by the average weight so that the exponential mechanism: $e^{(\frac{\frac{\varepsilon}{2}\bar{w}u}{\bar{w}})} = e^{(\frac{\varepsilon}{2}u)}$
  - Perturb sample counts, eg. add/subtract k from the count and adjust weighted count by k$\bar{w}$

Future for Differential Privacy

- Synthetic data

Ongoing Research:

- Bayesian Modeling with differentially private priors (On the Map, Abowd and Vilhuber, 2008)
- Adding noise to estimating equations (Chipperfield and O'keefe, 2014) and use them in Sequential (Ridge) Regression modeling in a multiple imputation framework (Ragunathan, et al. 2001)
- Reproducing microdata from differentially private counts (plans for US Census)
- Remote Analysis Servers

Future for Differential Privacy

Is differential privacy useful in the SDL toolkit used at NSIs?

Allows statistical agencies to consider new and open ways of disseminating data

Should be incorporated into common SDL practices, eg. coarsening, suppression, sampling as these methods will influence the privacy budget

Additive noise perturbation of DP for (large) counts can provide more utility than current SDL additive noise perturbation

It provides a formal privacy guarantee for inferential disclosure and parameters can be made public to correct inferences

Challenges for Differential Privacy

Focus must be on non-interactive mechanisms where privacy budget is set at time of perturbation and all subsequent analyses are safe

How to set privacy budget, particularly to differentiate census and sample data and protection provided by sampling and other forms of SDL

More research needed for other forms of data dissemination, eg. synthetic data

Training a new generation of social researchers to analyse 'noisy' perturbed data

Thank you for your attention