On Differential Privacy and Data Utility in SDC

J. Soria-Comas J. Domingo-Ferrer

Department of Computer Engineering and Mathematics
Universitat Rovira i Virgili

UNESCO Chair in Data Privacy

UNECE/Eurostat Work Session on SDC, 2011

Outline

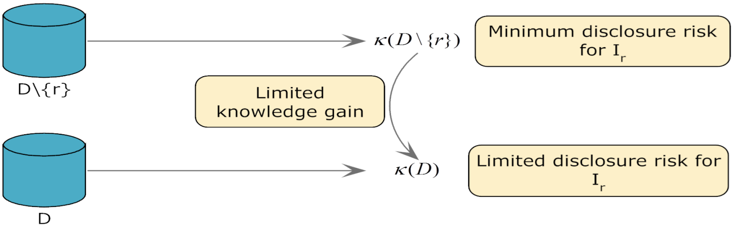
Differential Privacy

- Limit the knowledge gain achievable by performing a query over data sets that differ in one individual (a.k.a. neighbor data sets)

ε-differential privacy

A randomized function gives ε-differential privacy if for all neighbor data sets D and D', and all $S \subset Range(\kappa)$

$$P(\kappa(D) \in S) \leq e^{\varepsilon} P(\kappa(D') \in S)$$



Types of Noises

- Data-Independent Noise
  - Distribution of data-independent noise is constant across data sets
  - The required amount of noise depends on the maximum change in the query function between neighbor data sets
  - Commom procedure: Add independent Laplace distributed noise with zero mean and and ${}^{\Delta f}/_{\varepsilon}$ scale, to each component of the query response
- Data-Dependent Noise
  - The distribution of a data-dependent noise is adjusted to the sensitivity of the query function local to each data set
  - Eligible distributions must be heavy tailed.
  - The proposal is to use $\frac{4\delta \times S^*_{f,\beta}}{\varepsilon} Z$, where Z is a random noise with density function proportional to $\frac{1}{1+|x|^\delta}$.

Outline

Optimal Noise Distribution

- Several criteria are commonly used: variance, expectation of the L_1 norm, size of a confidence region, etc.
- The essence is to take smaller noise values with greater probability.

Definition

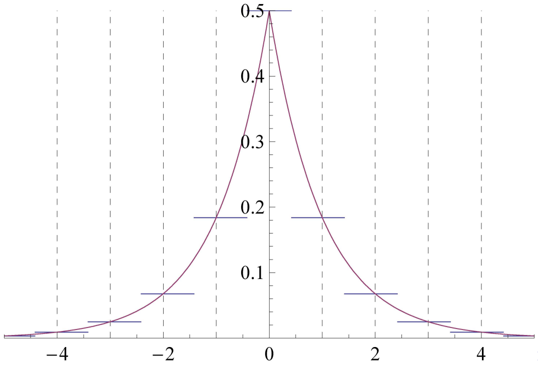
Let N_1 and N_2 be random noises.

- N_1 is smaller than N_2, $N_1 \leq N_2$, if for all α, $P(|N_1| \leq \alpha) \geq P(|N_2| \leq \alpha)$
- N_1 is strictly smaller than N_2, $N_1 < N_2$, if the above inequality is strict
- $N_1 \in \mathscr{C}$ is optimal within $\mathscr{C}$, if for any $N_2 \in \mathscr{C}$ it holds $N_2 \not< N_1$

- A family of optimal distributions exists. Another criterion may be used to further refine the search.

Laplace is not Optimal

- It is possible to modify the Laplace density function in such a way that:
  - ε-differential privacy still holds
  - The probability mass is more concentrated towards the zero.
- Idea: Split the range into disjoint intervals and redistribute the probability mass inside each interval.

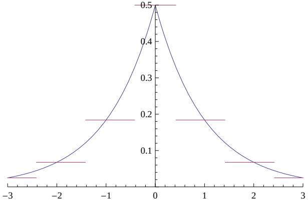


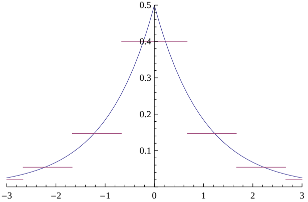
Optimal A. C. Data-Independent Noise

- Idea: Apply to a generic distribution a procedure similar to the one applied to the Laplace distribution.
- The density of an optimal a.c. data-independent distribution has the form

$$pdf(x) = \begin{cases} M & |x| \in [0, d] \\ Me^{-i\varepsilon} & |x| \in [d + i\Delta f, d + (i+1)\Delta f] \end{cases}$$

for some values M and d such that $d \in [0, \Delta f]$ and the total probability mass equals one.





Comparison (I): Single Query

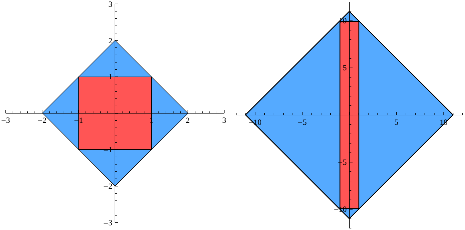
- The table compares the variance of Laplace to the minimum variance achievable with a data-independent noise.

Laplace Optimal		ε = 0.1	ε = 1	ε = 10
Δf	0.1	2 1.999	0.02 0.0192	2×10^{-4} 8.47×10^{-6}
	1	200 199.9	2 1.92	0.02 8.47×10^{-4}
	10	20000 19991	200 191.8	2 8.47×10^{-2}

- For the case of a single query function, the improvement is relatively small. Only for large values of ε the improvement is relatively significant, but the disclosure risk for such ε is large.
- If Laplace does not provide the desired data quality, there is not much we can do with a data-independent noise.

Comparison (II): Multiple Queries

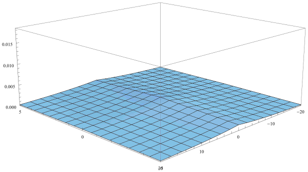
- With Laplace all the points with the same L_1-norm have the same density.

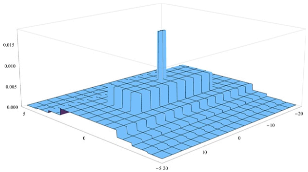


- The density function is similar to the one for a single query: it is a stepwise function that reaches its maximum in a set that contains zero.

Comparison (III): Multiple Queries

- Sample density functions when using Laplace and one optimal a.c. distribution .





Conf. Level	Laplace	Optimal
0.99	10663	1790
0.95	5445	916

Outline

Data-Independent vs Data-Dependent Noise

- Data-dependent noise makes sense only if the smooth sensitivity at the actual data set is small compared to the L_1-sensitivity.
- By comparing the variances we may come up with a rule of thumb to choose between data-independent and data-dependent noise:

$$V_{Laplace} = 2 \times (\Delta f/\varepsilon)^2$$
$$V_{Dependent} = 14\delta^2 \sin(\pi/\delta)/\sin(3\pi/\delta) \times (S^*_{f,\beta}(D)/\varepsilon)^2$$

- The smooth sensitivity at the actual data set must be at least 10.96 times smaller than the L_1-sensitivity.

Outline

Adding/Removing vs Modifying Records

- There are two main definition of what neighbor data sets are:
  1. Two data sets D and D' are said to be neighbors if one can be obtained from the other by adding or removing a record (Neighborhood 1)
  2. Two data sets D and D' are said to be neighbors if one can be obtained from the other by modifying a record (Neighborhood 2)
- Modifying a record does not change the cardinality of the data set
  - With Neighborhood 2 we may restrict the comparison to data sets with the same cardinality as the actual data set D
- It may seem that Neighborhood 2 may provide more accurate results when the query function has reduced sensitivity over the set of data sets with equal size

Example: The Relative Frequency

- Let f be a query function that returns the relative frequency of some property
- Let Δ_i be the sensitivity under Neighborhood i
- When querying the whole data set we have:

$$\Delta_1 f = 1/2$$
$$\Delta_2 f = 1/|D|$$

- When querying some subset we have:

$$\Delta_1 f = 1/2$$
$$\Delta_2 f = 1/2$$

- To get some benefit from Neighborhood 2, we must query the whole data set
- Neighborhood 2 may lead to higher sensitivity than Neighborhood 1 with multiple queries
- We only consider Neighborhood 1 in what follows

Absolute Frequency

- The local sensitivity is constant and equal to the L_1-sensitivity. Using data-dependent noise makes no sense.

	$\varepsilon = 0.1$	$\varepsilon = 1$
Confidence intervals at 95%		
Laplace	$[m-29.9, m+29.9]$	$[m-2.99, m+2.99]$
Smooth Sensitivity $\delta = 4.37$	$[m-285, m+285]$	$[m-28.5, m+28.5]$
Variance		
Laplace	200	2
Smooth Sensitivity $\delta = 4.52$	24045	240

- The utility of the result depends on the actual value m of the absolute frequency. The greater m, the less relative error introduced.

Relative Frequency (I)

- The local sensitivity depends on both the size of the data set n, and on the number of records satisfying the property m.

$$\begin{aligned} \Delta f &= 1/2 \\ LS_f(D) &= \max\{\tfrac{n-m}{n(n-1)}, \tfrac{m}{n(n-1)}\} < 1/n-1 \end{aligned}$$

- With data-independent Laplace distributed noise we have

Laplace	$\varepsilon = 0.1$	$\varepsilon = 1$
Confidence intervals at 95%	$[\frac{m}{n} - 15, \frac{m}{n} + 15]$	$[\frac{m}{n} - 1.5, \frac{m}{n} + 1.5]$
Variance	50	5

$\Rightarrow$ Data-independent noise is not usable for the relative frequency

Relative Frequency (II)

- With the corresponding data-dependent noise that minimizes the size of the confidence interval and the variance, we have

Conf.Int. Variance		n = 100	n = 1000	n = 10000
m	0	$m/n \pm 8.34$ 30.2	$m/n \pm 0.285$ 0.024	$m/n \pm 0.0285$ 0.00024
	$0.5n$	$m/n \pm 8.60$ 32.0	$m/n \pm 0.143$ 0.006	$m/n \pm 0.0143$ 0.00006

- Data-dependent noise improves a lot over data-independent noise; however the size of the data set needs to be quite big for the results to be acceptable.

Maximum/Minimum Queries

- Let f return the maximum value in a field with range $[0,1]$
- The L_1-sensitivity equals the size of the range of the function: data-independent noise is not usable.

Laplace	$\varepsilon = 0.1$	$\varepsilon = 1$
Confidence intervals at 95%	$f(D) \pm 15$	$f(D) \pm 1.5$
Variance	50	5

- The smooth sensitivity depends on the actual values in the data set. A systematic approach is not possible.
- We simulate data set values following a uniform distribution in $[0,1]$, and a beta distribution with $\alpha = 2$ and $\beta = 5$. Results are only good for very large n.

Confidence intervals at 95%		$\mathscr{U}[0,1]$	$Be(2,5)$
n	100	$f(D) \pm 39.4$	$f(D) \pm 92.3$
	1000	$f(D) \pm 3.99$	$f(D) \pm 55.4$
	10000	$f(D) \pm 0.399$	$f(D) \pm 34.1$

Thank you