Data Obfuscation

Bimal Kumar Roy

December 17, 2015

Problem description (informal)

- Owner with large database.
- Lends the database for public use – user is allowed to run restriced set of queries on data items.
- Goal is to prevent the user from deriving any other information from database than what is derivable from the allowed queries.
  Or
  Whatever information the user can efficiently derive can be efficiently derived from the allowed queries
- Data Obfuscation is a type of data masking where some useful information about the complete dataset remains even after hiding the individual sensitive information

Some techniques

- Substitution- substitute an entry with a randomly chosen entry from the same domain.
- Example:
  - Database contains column with names from different nationalities.
  - User is interested in statistic on nationality.
  - Original names in the database can be substituted with randomly chosen names from the same nationality.

  $\rightarrow$ **Substitution has to be sufficiently random.**
- Shuffling- shuffle entries from same domain.

  $\rightarrow$ **Shuffling has to be sufficiently random**

Data Perturbation

- **Example**
  - Column contains employee salaries.
  - User wants average of the salaries
  - Goal is to prevent the user from getting information about individual salaries.
  - For each entry add a randomly sampled number from suitable distribution with average zero.
  - For a sufficiently large database the average is very close to the original average with high confidence.

$\rightarrow$ **Need large amount of random data.**

- Deletion - delete columns that the user does not require.

$\rightarrow$ **Sometimes it is not possible to alter the structure of the database by deleting columns. Also, if there are no copies of the original database, information will be permanently lost.**

Theoretical Example (Narayanan et al.)

- Database consits of tuples of the form $< x, y >$.
- It is required that y be accessible only if x is known . (Prevents against mass-harvesting of data)
- Keep tuples as $< hash(x||r1), hash(x||r2) + y, r1, r2 >$

Application of Cryptography

Consider the following scenario.

- Size of the database is huge, so duplication is not possible.
- User requires original database to test application.
- Owner requires privacy of certain columns (attributes).

Solution:

- Encrypt data of the private columns. It requires a short (128 bit, 256 bit etc.) random key which remains secret with the owner.
- Problem with traditional encryption modes:
  - They are not format preserving.
  - AADHAAR number 4580 5000 8000 encrypts to **** under 256-bit AES ECB mode – user application accessing AADHAAR field has check and validation for 12-digit AADHAAR number, and aborts on finding long "unformatted" encryption.

Format Preserving Encryption

- Mode of encryption where "format" of ciphertext is same as plaintext – it is a "random-like" permutation of the domain of the plaintext.
  - 12-digit AADHAAR number maps to 12-digit AADHAAR number.
  - 16-digit credit card number maps to 16-digit credit card number.
- FFX mode of operation of AES proposed by Bellare et al. supports format preserving encryption and is under consideration by NIST for approval.

Objectives of Data Obfuscation

- Minimize risk of disclosure resulting from providing access to the data.
- Maximize the analytical usefulness of the data.

Methods for Obfuscating Data

- *Topcoding*
- *Grouping*
- *Adding or Multiplying noise*
- *Rank Swapping*

Additive Noise Model

- $\{X_i, 1 \leq i \leq n\}$: True data set $\sim G(.)$
- $\{Y_i, 1 \leq i \leq n\}$: Obfuscating data set $\sim F(.)$ (known)
- $\{Z_i, 1 \leq i \leq n\}$: Obfuscated data set $\sim H(.)$

$$Z_i = X_i + Y_i \text{ is released.}$$

AIM: To estimate all the quantiles from the masked data Z and obfuscating distribution f.

Assumptions

- $G(.)$ and $F(.)$ and hence $H(.)$ are continuous functions.
- Y_i is independent of X_i $\forall i$.

Derivation of Equation

For $x \in R$,

$$\begin{aligned} H(x) &= P(Z \leq x) \\ &= P(X + Y \leq x) \\ &= \int_{-\infty}^{\infty} P(X \leq x - y | Y = y) f(y) \partial y \\ &= \int_{-\infty}^{\infty} G(x - y) f(y) \partial y \\ &= \int_{-\infty}^{\infty} f(x - y) G(y) \partial y \end{aligned}$$

Hence we have an integral equation to be solved for. Here $G(.)$ is the unknown to be solved for, $f(.)$ is the known Kernel and $H(.)$ is to be estimated from the sample.

Uniform Error

Here, $F(\cdot) \sim Unif(0, a),\ ,\ a \in R^+$

$$H(z) = \frac{1}{a} \cdot \int_0^a G(z-y)\partial y$$

Now differentiating w.r.t z we have,

$$h(z) = \frac{1}{a} \cdot \{G(z) - G(z-a)\} \text{ which again gives}$$

$$G(z) = a \cdot h(z) + G(z-a)$$

Now successively putting the values $G(z-ma)$ for $m = 1, 2, \ldots$ we have

$$G(x) = a \cdot h(x) + a \cdot h(x-a) + a \cdot h(x-2a) \cdots\ ,\ \forall x \in R$$

In our problem, $h(.)$ is unknown; so instead we can use the kernel density estimate of $h(.)$ to get an estimate of $G(x)$ for all $x \in R$. Then, equating $\hat{G}(x) = \alpha$ for $0 < \alpha < 1$ we get the α^{th} quantile of X.

Normal Error

Here, $F(\cdot) \sim \Phi(\cdot, 0, \sigma^2)$ σ*known*.
[Note that, the mean is taken to be 0 w.l.g. since if the mean is non-zero, say, $\mu \neq 0$,$Z - \mu = X + (Y - \mu)$, $Y - \mu \sim N(0, \sigma^2)$. Hence we can work with $Z - \mu$ instead of Z and we will end up with the same X and $Y \sim N(0, \sigma^2)$].
So, our equation becomes:

$$H(x) = \int_{-\infty}^{\infty} G(x - y)\phi_\sigma(y)\partial y = \int_{-\infty}^{\infty} \phi_\sigma(x - y)G(y)\partial y$$

This is a **Fredholm equation of first kind with infinite limits.**
To solve this, we use the following result.

Result

Consider the equation,

$$f(x) = \int_{-\infty}^{\infty} K(x-t)y(t)\partial t,\ -\infty < x < \infty.$$

Then if

- $f(.), y(.) \in L^2(-\infty, \infty)$
- $K(.) \in L^1(-\infty, \infty)$
- $\frac{\tilde{f}(.)}{\tilde{K}(.)} \in L^2(-\infty, \infty)$, where $\tilde{f}(.)$,$\tilde{K}(.)$ are Fourier transforms of $f(.)$ and $K(.)$ respectively.

The exact solution of $y(x)$ is given by,

$$y(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\tilde{f}(u)}{\tilde{K}(u)} e^{iux} \partial u$$

Estimation of the distribution function

In our equation, note that $H(.)$, $G(.)$ are not L^2. So we take a derivative of the equation w.r.t x to get

$$h(x) = \int_{-\infty}^{\infty} g(x-y)\phi_\sigma(y)\partial y$$

Now note that here both $h(.)$ and $g(.)$ are p.d.f's and hence L^2 functions. Moreover $K(.)$ is L^1 since a p.d.f.

Hence an inversion can be done.

Moreover putting $\hat{h}(.)$ instead of $h(.)$, where $\hat{h}(.)$ is the Kernel Density estimate of $h(.)$ using *Gaussian Kernel and standard bandwidth, which is,*

$$b = 1.06 \cdot n^{-1/5} \cdot A, \text{ where } A = \min[\hat{\sigma}(Z), IQR(Z)/1.34]$$

We have,

$$\tilde{\hat{h}}(k) = constant \times \sum_{j=1}^{n} e^{iZ_j k - \frac{k^2 b^2}{2}}$$

$$\tilde{\phi}_\sigma(k) = constant \times e^{-\frac{k^2\sigma^2}{2}}$$

Continuation

Hence

$$\frac{\tilde{\hat{h}}(k)}{\tilde{\phi}_\sigma(k)} = constant \times e^{iZ_j k - \frac{k^2}{2}(b^2-\sigma^2)}$$

is L^2 if $b > \sigma$. In such cases, an estimate of $G(.)$ is given by,

$$\hat{G}(x) = 1/n \sum_{j=1}^{n} \Phi(x - Z_j, 0, \sqrt{b^2 - \sigma^2}).$$

Now, to find the α^{th} quantile, we solve $\hat{G}(x) = \alpha$ for some given α by some iterative method like Newton-Raphson or finding root within an interval.
But, in case $b < \sigma$, the solution does not exist as the integral does not exist. Moreover, note that if σ is too small compared to the range of the data, obfuscation will have no effect.

Intuition to the solution of the problem

- To find a solution to the above problem first lets check why this occurs !
- To check, we see it occurs because in the equation,

$$\frac{\tilde{\hat{h}}(k)}{\tilde{\phi}_\sigma(k)} = constant \times e^{iZ_j k - \frac{k^2}{2}(b^2 - \sigma^2)}$$

  the numerator always has the form $constant \times e^{iZ_j k - \frac{k^2}{2}(b^2)}$
- And the denominator has the form $constant \times e^{\frac{k^2}{2}(\sigma^2)}$
- Now if $b < \sigma$ the coefficient of k^2 becomes positive making the whole function non-L^2.
- But if we could choose a ϕ which would involve a weaker function of k then we could vary σ
- Now, we know, fourier transform of Laplace is,

$$\tilde{L}_\sigma(k) = \frac{1}{k^2 + \sigma^2}$$

  which is much weaker to the exponential function in

Solution to the problem

In case of Laplace Error our equation will be:

$$h(x) = \int_{-\infty}^{\infty} g(x-y) f_\sigma(y) \partial y$$

where

$$f_\sigma(x) = \frac{1}{2\sigma} \cdot e^{-\frac{|x|}{\sigma}} \text{ for } -\infty < x < \infty$$

Hence applying the same result as before and same technique to estimate the density, we have

$$\frac{\tilde{\hat{h}}(k)}{\tilde{f}_\sigma(k)} = (1 + k^2\sigma^2) e^{iZ_j k - \frac{k^2 b^2}{2}}$$

which will be L^2 for all values of σ. The form of $G(.)$ in this case will be given by

$$\hat{G}(x) = 1/n \sum_{j=1}^{n} \{(1 + \frac{\sigma^2}{b^2}) \Phi(x, Z_j, b) - \frac{\sigma^2}{b^2} \int_{-\infty}^{(x-Z_j)/b} u^2 \phi(u) \partial u\}$$

Simulation Results

A sample is drawn from some non-normal distribution with high variation($\frac{IQR}{1.34} \approx 1000$). Sample size is taken to be $n = 2000$. For a comparatively small σ (say, $\sigma = \mathbf{200}$), we add a Uniform,Normal and a Laplace Error with scale σ and check how the form of $G(x)$ works to estimate the empirical c.d.f of X.

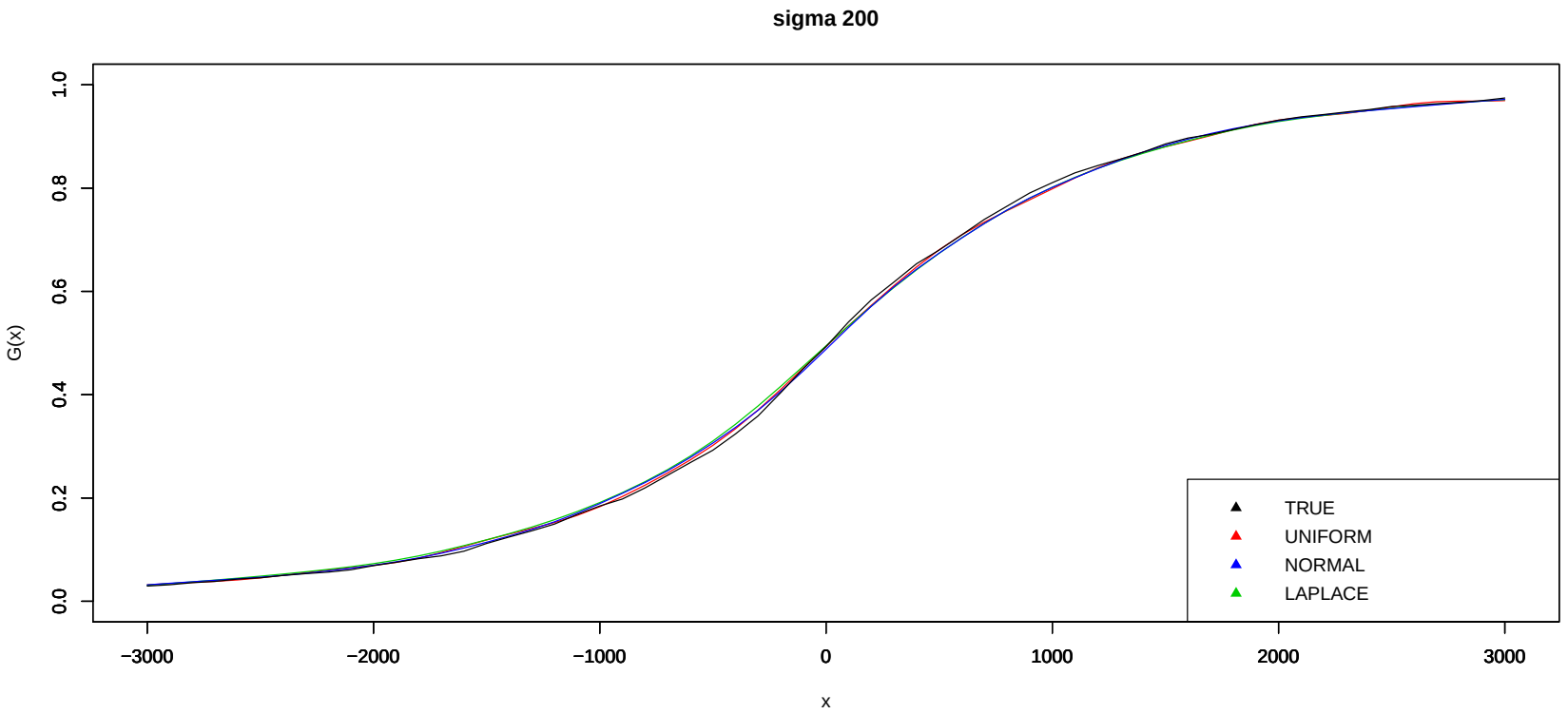

Figure: Checking the Performance of Estimation

Simulation Results

The following table will show the true and estimated values of the quantiles of X.

"Alpha"	"TRUE"	"Uniform"	"Normal"	"Laplace"
"0.1"	-1576.72	-1652.403	-1673.045	-1630.072
"0.2"	-0892.373	-0913.834	-0953.046	-0944.601
"0.3"	-0480.524	-0506.302	-0533.804	-0522.084
"0.4"	-0209.441	-0225.224	-0242.204	-0216.965
"0.5"	0019.652	0016.732	0014.146	0028.594
"0.6"	0239.942	0270.666	0278.556	0275.719
"0.7"	0556.685	0565.444	0586.859	0587.931
"0.8"	0935.518	1006.787	0992.126	0992.953
"0.9"	1663.415	1685.224	1674.487	1646.105

Simulation Results

As we increase σ (say, $\sigma = (\mathbf{200}, \mathbf{400}, \mathbf{600}, \mathbf{800})$), we see what happens to the estimation of $G(x)$

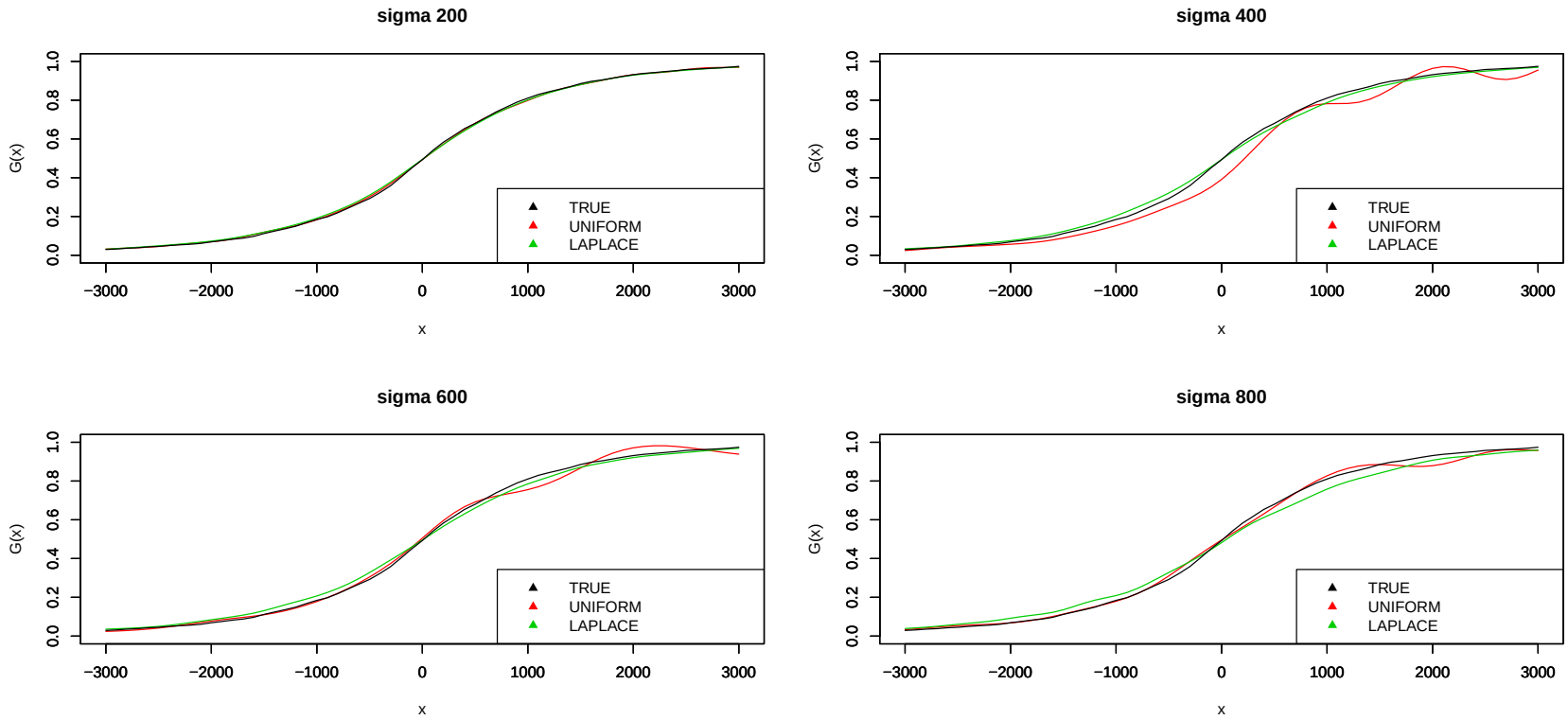

Figure: Checking the Performance of Estimation for increasing σ

Remarks

- It is clear from the graph that the error in estimation increases as σ increases. And this is quite intuitive as a small σ means approximately no obfuscation at all.
- Hence σ should be wisely chosen (not too small to degrade obfuscation, not too large to harm the extraction).
- Note that, the performance of such estimation depends hugely on the kernel density estimation. If the density estimation is weak, its performance becomes weak simultaneously.

Significance

After obfuscation, one will get a C.I. for each observation. Suppose for each X_i we want a $100(1-\delta)\%$ C.I. to be $(Z_i - \epsilon, Z_i + \epsilon)$ for fixed ϵ.

We know,

$$\begin{aligned} P(X_i \in (Z_i - \epsilon, Z_i + \epsilon)) &= P(|Z_i - X_i| < \epsilon) \\ &= P(|Y_i| < \epsilon) \\ &= 2F_\sigma(\epsilon) - 1 \end{aligned}$$

Now,

$$\begin{aligned} 2F_\sigma(\epsilon) - 1 &= 1 - \delta \\ \Rightarrow 2F_\sigma(\epsilon) &= 2 - \delta \\ \Rightarrow F_\sigma(\epsilon) &= 1 - \delta/2 \end{aligned}$$

Hence for fixed ϵ and δ we find a value of σ from the equation

$F_\sigma(\epsilon) = 1 - \delta/2$, where $F(.)$ is either Laplace or Normal c.d.f.

SOME REFERENCES

- Fuller Wayne A., *Masking Procedures for Microdata Disclosure Limitation, Journal of Official Statistics, Vol. 9, 1993, pp383-406*
- *Mukherjee Sumitra, Duncan George T.,* Disclosure Limitation through Additive Noise Data Masking: Analysis of Skewed Sensitive Data,IEEE 1997
- Kim Hang J., Karr Alan F. *The Effect of Statistical Disclosure Limitation on Parameter Estimation for a Finite Population, October 2013*
- *Kotsireas Ilias S.,* A Survey on Solution Methods for Integral Equations, June 2008
- Polyanin Andrei D., Manzhirov Alexander V., *Handbook of Integral Equations*