Integrating Differential Privacy with Statistical Theory

Adam Smith
Computer Science & Engineering Department
Penn State

NCHS/CDC Workshop on Data Confidentiality
May 1, 2008
Differential Privacy

- Definition of privacy in statistical databases
  - Imposes restrictions on algorithm A generating output
- If A satisfies restrictions, then output provides privacy no matter what user/intruder knows ahead of time
- **Question**: how **useful** are algorithms that satisfy differential privacy?
This talk: Useful Statistical Inference

- Two situations where differential privacy compatible with statistical methodology
- In both cases: construct differentially private algorithm with same asymptotic error as best non-private algorithm

- **Parametric**: for any* parametric model, there exists a private, efficient estimator (i.e. minimal variance)

- **Nonparametric**: for any* distribution on [0,1], there is a private histogram estimator with same convergence rate as best (non-private) fixed-width estimator
Main Idea for both cases

• Add noise to carefully modified estimator
  ➢ Several ways to design differentially private algorithms
  ➢ Adding noise is the simplest

• Prove that required noise is less than inherent variability due to sampling
Bigger Goal

• Understanding how rigorous notions of privacy relate to statistical inference
  ➢ (Also: crossing disciplinary boundaries requires understanding, and working with, other communities’ language)

• First step: basic asymptotic theory
  ➢ Cornerstone of statistical techniques
  ➢ Qualitative statements
    • asymptotic regime allows for clean statements
    • highlights where techniques breakdown
  ➢ Intuition for messier real settings
Reminder: differential privacy

• Intuition:

  - Changes to my data not noticeable by users
  - Output is “independent” of my data
Defining Privacy \([\text{DiNi, DwNi, BDMN, DMNS}]\)

- **Data set** \(x = (x_1, ..., x_n) \in D^n\)
  - Domain \(D\) can be numbers, categories, tax forms
  - Think of \(x\) as **fixed** (not random)
- **A** = **randomized** procedure run by the agency
  - \(A(x)\) is a random variable distributed over possible outputs
  - Randomness might come from adding noise, resampling, etc.
Defining Privacy [DiNi,DwNi,BDMN,DMNS]

\[ x_1, x_2, \ldots, x_n \]

A

A(x)

\[ A(x') \]

local random coins

x’ is a neighbor of x if they differ in one data point
Defining Privacy [DiNi,DwNi,BDMN,DMNS]

Local random coins $A(x)$

$x'$ is a neighbor of $x$ if they differ in one data point

Neighboring databases induce close distributions on outputs
Defining Privacy [DiNi, DwNi, BDMN, DMNS]

\[ x' \text{ is a neighbor of } x \]
\[ \text{if they differ in one data point} \]

**Definition:** A is \( \epsilon \)-differentially private if, for all neighbors \( x, x' \), for all subsets \( S \) of outputs

\[
\Pr( A(x) \in S ) \leq e^\epsilon \cdot \Pr( A(x') \in S )
\]
**Defining Privacy** [DiNi,DwNi,BDMN,DMNS]

- $\epsilon$ cannot be too small (think $\frac{1}{10}$, not $\frac{1}{2^{50}}$)
- This is a condition on the **algorithm** (process) $A$
  - Saying “this output is safe” doesn’t take into account how it was computed
- Meaningful semantics no matter what user knows ahead of time

**Definition**: $A$ is $\epsilon$-differentially private if, for all neighbors $x, x'$, for all subsets $S$ of outputs

$$\Pr(A(x) \in S) \leq e^{\epsilon} \cdot \Pr(A(x') \in S)$$

Neighboring databases induce close distributions on outputs.
Example: Perturbing the Average

\[ A(x) = \bar{x} + \text{noise} \]

\[ x_i \in \{0, 1\} \]

\[ \bar{x} = \frac{1}{n} \sum_i x_i \]
Example: Perturbing the Average

- Data points are binary responses $x_i \in \{0, 1\}$
- Server wants to release sample mean $\bar{x} = \frac{1}{n} \sum_i x_i$
Example: Perturbing the Average

- Data points are binary responses \( x_i \in \{0, 1\} \)
- Server wants to release sample mean \( \bar{x} = \frac{1}{n} \sum_i x_i \)
- **Claim:** If noise \( \sim \text{Lap}(\frac{1}{\epsilon n}) \) then \( A \) is \( \epsilon \)-differentially private
Example: Perturbing the Average

- Data points are binary responses \( x_i \in \{0, 1\} \)
- Server wants to release sample mean \( \bar{x} = \frac{1}{n} \sum_i x_i \)
- **Claim:** If noise \( \sim \text{Lap}\left(\frac{1}{\epsilon n}\right) \) then A is \( \epsilon \)-differentially private

\[
A(x) = \bar{x} + \text{noise} \\
\approx \bar{x} \pm \frac{1}{\epsilon n}
\]

If \( x \) is a random sample from an underlying population, then get sampling noise \( \approx \frac{1}{\sqrt{n}} \).
Example: Perturbing the Average

- Data points are binary responses $x_i \in \{0, 1\}$
- Server wants to release sample mean $\bar{x} = \frac{1}{n} \sum_i x_i$
- **Claim:** If noise $\sim \text{Lap}(\frac{1}{\epsilon n})$ then $A$ is $\epsilon$-differentially private

- Laplace distribution $\text{Lap}(\lambda)$ has density $h(y) \propto e^{-|y|/\lambda}$
- Sliding property: $\frac{h(y)}{h(y+\delta)} \leq e^{\delta/\lambda}$

If $x$ is a random sample from an underlying population, then get sampling noise $\approx \frac{1}{\sqrt{n}}$. 
Example: Perturbing the Average

- Data points are binary responses $x_i \in \{0, 1\}$
- Server wants to release sample mean $\bar{x} = \frac{1}{n} \sum_i x_i$
- **Claim:** If noise $\sim \text{Lap}(\frac{1}{\epsilon n})$ then $A$ is $\epsilon$-differentially private
  
  - Laplace distribution $\text{Lap}(\lambda)$ has density $h(y) \propto e^{-|y|/\lambda}$
  - Sliding property: $\frac{h(y)}{h(y+\delta)} \leq e^{\delta/\lambda}$

If $x$ is a random sample from an underlying population, then get sampling noise $\approx \frac{1}{\sqrt{n}}$. 
Example: Perturbing the Average

- Data points are binary responses \( x_i \in \{0, 1\} \)
- Server wants to release sample mean \( \bar{x} = \frac{1}{n} \sum_i x_i \)
- **Claim:** If noise \( \sim \text{Lap}(\frac{1}{\epsilon n}) \) then A is \( \epsilon \)-differentially private

- Laplace distribution \( \text{Lap}(\lambda) \) has density \( h(y) \propto e^{-|y|/\lambda} \)
- Sliding property: \( \frac{h(y)}{h(y+\delta)} \leq e^{\delta/\lambda} \)
- \( A(x) = \text{blue curve}, \quad A(x') = \text{red curve} \)
- \( \delta = |\bar{x} - \bar{x}'| \leq \frac{1}{n} \implies \text{blue curve} \leq e^\epsilon \leq \text{red curve} \)
What can we compute privately?

• “Privacy” = change in one input leads to small change in output distribution

What computational tasks can we achieve privately?

• Research so far

  ➢ Function approximation \([\text{DN, DN, BDMN, DMNS, NRS, BCDKMT, BLR}]\)
  ➢ Mechanism Design \([\text{MT}]\)
  ➢ Learning \([\text{BDMN, KLNRS}]\)
  ➢ Statistical estimation \([\text{S}]\)
  ➢ Synthetic Data \([\text{MKAGV}]\)
  ➢ Distributed protocols \([\text{DKMMN, BNO}]\)
  ➢ Impossibility results / lower bounds \([\text{DiNi, DMNS, DMT}]\)
When Does Noise **Not** Matter?

- **Average:** $A(x) = \bar{x} + \text{Lap}\left(\frac{1}{\epsilon n}\right)$

  - Suppose $X_1, X_2, X_3, ..., X_n$ are i.i.d. random variables
  
  - $\bar{X}$ is a random variable, and $\sqrt{n} \cdot (\bar{X} - \mu) \xrightarrow{D} \text{Normal}$
  
  - $\frac{A(X) - \bar{X}}{\text{StdDev}(\bar{X})} \xrightarrow{P} 0$ if $\epsilon \gg \frac{1}{\sqrt{n}}$

- **No “cost” to privacy:**
  
  - $A(X)$ is “as good as” $\bar{X}$ for statistical inference*

![Graph](https://via.placeholder.com/150)
When Does Noise *Not* Matter?

\[ \sqrt{d} \]
When Does Noise Not Matter?

• Mean example generalizes to other statistics

• **Theorem:** For any* exponential family, can release “approximately sufficient” statistics

  ➢ Suff. stats $T(X)$ are sums, add noise $\frac{d}{\epsilon n}$ for dimension $d$

  ➢ $\frac{A(X) - T(X)}{\text{StdDev}(T(X))} \xrightarrow{P} 0$

\[ \sqrt{d} \]
When Does Noise **Not** Matter?

- **Mean example generalizes to other statistics**

- **Theorem:** For any* exponential family, can release “approximately sufficient” statistics
  
  - Suff. stats $T(X)$ are sums, add noise $\frac{d}{\epsilon n}$ for dimension $d$
  
  - \[
  \frac{A(X) - T(X)}{\text{StdDev}(T(X))} \overset{P}{\rightarrow} 0
  \]

- **Asymptotic result:** Indicates that useful analysis possible
  
  - Requires more sophisticated processing for small $n$

\[\sqrt{d}\]
When Does Noise **Not** Matter?

- Mean example generalizes to other statistics

- **Theorem:** For any* exponential family, can release “approximately sufficient” statistics
  - Suff. stats $T(X)$ are sums, add noise $\frac{d}{\epsilon n}$ for dimension $d$
  - $\frac{A(X) - T(X)}{\text{StdDev}(T(X))} \xrightarrow{P} 0$

- Asymptotic result: Indicates that useful analysis possible
  - Requires more sophisticated processing for small $n$

- Noise degrades with dimension (can get noise $\sim \sqrt{d}$)
  - More information $\Rightarrow$ less privacy
  - Research question: Is this necessary?
Two More Examples

- **Theorem:** For any well-behaved parametric family, one can construct a private **efficient** estimator \( A \), if \( \epsilon \sqrt[4]{n} \rightarrow \infty \)
  - \( A(X) \) converges to MLE

- For any smooth density \( h \), if \( X_i \) i.i.d. \( \sim h \), noisy histogram converges to \( h \)
  - Expected L2 error \( O\left(\frac{1}{\sqrt[3]{n}}\right) \) if \( \epsilon \geq \frac{1}{\sqrt[3]{n}} \)
• Histogram Density Estimation
  ➢ Calibrating noise to sensitivity

• Maximum Likelihood Estimator
  ➢ Sub-sample and aggregate
Output Perturbation, more generally

- May be interactive
  - Non-interactive: release pre-defined summary stats + noise
  - Interactive: respond to user requests
- May be repeated many times
  - Composition: $q$ releases are jointly $q\varepsilon$-differentially private
- How much noise is enough? (How much is too much?)
Global Sensitivity [DMNS06]

- Intuition: $f(x)$ can be released accurately when $f$ is insensitive to individual entries $x_1, x_2, \ldots, x_n$

- Global Sensitivity: $GS_f = \max_{\text{neighbors } x, x'} \| f(x) - f(x') \|_1$

- Example: $GS_{\text{average}} = \frac{1}{n}$
Global Sensitivity [DMNS06]

- Intuition: $f(x)$ can be released accurately when $f$ is insensitive to individual entries $x_1, x_2, \ldots, x_n$.

- Global Sensitivity:
  \[
  GS_f = \max_{\text{neighbors } x, x'} \| f(x) - f(x') \|_1
  \]

- Example: $GS_{\text{average}} = \frac{1}{n}$

Theorem: If $A(x) = f(x) + \text{Lap} \left( \frac{GS_f}{\epsilon} \right)$, then $A$ is $\epsilon$-differentially private.
Example: Histograms

\[ f(x) = (n_1, n_2, ..., n_d) \text{ where } n_j = \# \{ i : x_i \text{ in } j\text{-th interval} \} \]

\[ \text{Lap}(1/\epsilon) \]
Example: Histograms

- Say $x_1, x_2, ..., x_n$ in $[0, 1]$
  - Partition $[0, 1]$ into $d$ intervals of equal size
  - $f(x) = (n_1, n_2, ..., n_d)$ where $n_j = \#\{ i : x_i \text{ in } j\text{-th interval}\}$
  - $\text{GS}_f = 2$
  - Sufficient to add noise $\text{Lap}(1/\epsilon)$ to each count
    - Independent of the dimension
Example: Histograms

- Say $x_1, x_2, \ldots, x_n$ in $[0, 1]$
  - Partition $[0, 1]$ into $d$ intervals of equal size
  - $f(x) = (n_1, n_2, \ldots, n_d)$ where $n_j = \# \{ i : x_i \text{ in } j\text{-th interval} \}$
  - $G_{S_f} = 2$
  - Sufficient to add noise $Lap(1/\epsilon)$ to each count
    - Independent of the dimension

- For any smooth density $h$, if $X_i$ i.i.d. $\sim h$, noisy histogram converges to $h$
  - Expected $L_2$ error $O\left(\frac{1}{\sqrt[3]{n}}\right)$ if $\epsilon \geq \frac{1}{\sqrt[3]{n}}$
  - Same as non-private estimator
Example: Histograms

- Say $x_1, x_2, ..., x_n$ in $[0, 1]$ arbitrary domain $D$
  - Partition $[0, 1]$ into $d$ intervals of equal size into $d$ disjoint "bins"
    - $f(x) = (n_1, n_2, ..., n_d)$ where $n_j = \#\{i: x_i \text{ in } j\text{-th interval}\}$
    - $\text{GS}_f = 2$
    - Sufficient to add noise $\text{Lap}(1/\epsilon)$ to each count
      - Independent of the dimension

- For any smooth density $h$, if $X_i$ i.i.d. $\sim h$, noisy histogram converges to $h$
  - Expected $L_2$ error $O\left(\frac{1}{\sqrt{d}n}\right)$ if $\epsilon \geq \frac{1}{\sqrt{d}n}$
  - Same as non-private estimator

\[ f(x) = (n_1, n_2, ..., n_d) \text{ where } n_j = \#\{i: x_i \text{ in } j\text{-th interval}\} \]
More detail

- This actually shows that for any given bin width, can find noisy estimator that is close to non-noisy estimator

- Does not address how to choose bin width
  - Subject to extensive research
  - Common “bandwidth selection” criteria can be approximated privately
  - Two-stage process
• Histogram Density Estimation
  ➢ Calibrating noise to sensitivity

• Maximum Likelihood Estimator
  ➢ Sub-sample and aggregate
High Global Sensitivity: Median

Example 1: median of $x_1, \ldots, x_n \in [0, 1]$

$x = \underbrace{0 \cdots 0}_n \underbrace{1 \cdots 1}_n$

$\frac{n-1}{2} \frac{n-1}{2}$

$\text{median}(x) = 0$

$x' = \underbrace{0 \cdots 0}_n \underbrace{1 \cdots 1}_n$

$\frac{n-1}{2} \frac{n-1}{2}$

$\text{median}(x') = 1$

$\text{GS}_{\text{median}} = 1$

- Noise magnitude: $\frac{1}{\varepsilon}$. Too much noise!

- But for most neighbor databases $x, x'$,

  $|\text{median}(x) - \text{median}(x')|$ is small.

- Can we add less noise on ”good” instances?
What about MLE?

• Sometimes MLE is well-behaved,
  ➢ e.g. observed proportion for binomial

• Sometimes we have no idea
  ➢ e.g. no closed form expression for mildly complex loglinear models
  ➢ Can have arbitrarily bad sensitivity
  ➢ NB: Similar problems faced by robust statistics
Getting Around Global Sensitivity

- **Local sensitivity** measures variability in neighborhood of specific data set [Nissim-Raskhodnikova-S, STOC 2007]
  - Connections to robust statistics
    - Bounded influence function implies expected local sensitivity is small
  - Local sensitivity needs to be smoothed
    - Interesting algorithmic/geometric problems
  - Not this talk

- Instead: Generic framework for smoothing functions so they have low sensitivity
  - No need to “understand” structure of function
**Intuition:** Replace $f$ with a less sensitive function $\tilde{f}$.

$$\tilde{f}(x) = g(f(\text{sample}_1), f(\text{sample}_2), \ldots, f(\text{sample}_s))$$
Example: Efficient Point Estimates

- Given a parametric model \( \{ f_\theta : \theta \in \Theta \} \)
- MLE = \( \text{argmax}_\theta (f_\theta(x)) \)
- Converges to Normal
  - Bias(MLE) = \( O(1/n) \)
  - Can be corrected so that bias(\( \hat{\theta} \)) = \( O(n^{-2}) \)

**Theorem:** If model is well-behaved, then sample-aggregate using \( \hat{\theta} \) gives **efficient estimator** if \( \epsilon n^{1/4} \rightarrow \infty \)

- Question: What is the best private estimator?
  - Error bounds degrade with dimension...
Conclusions

• Define privacy in terms of my effect on output
  - Meaningful despite arbitrary external information
  - I should participate if I get benefit

• What can we compute privately?
  - This talk: statistical estimators that are “as good” as optimal non-private estimators
  - New aspect to “curse” of dimensionality

• Data privacy is now (even) more challenging than in past
  - Data vastly more varied and valuable
  - External information more available
  - How can we reason rigorously about data privacy?