Differential Privacy
And Data Analysis

Aaron Roth
April 5, 2017
Protecting Privacy is Important

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2008

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the user's anonymity.

Class action lawsuit accuses AOL of violating the Electronic Communications Privacy Act, seeks $5,000 in damages per user. AOL’s director of research is fired.
Could a new Netflix contest put private customer data at risk?

By Matthew Shaer / September 22, 2009

Class action lawsuit (Doe v. Netflix) accuses Netflix of violating the Video Privacy Protection Act, seeks $2,000 in compensation for each of Netflix’s 2,000,000 subscribers. Settled for undisclosed sum, 2nd Netflix Challenge is cancelled.
The National Human Genome Research Institute (NHGRI) immediately restricted pooled genomic data that had previously been publically available.
But what is “privacy”?
But what is “privacy” not?

- Privacy is not hiding “personally identifiable information” (name, zip code, age, etc...)

A Face Is Exposed for AOL Searcher No. 4417749

Published: August 9, 2006

Buried in a list of 8 million Web search queries collected by AOL and recently released on the Internet is case No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

No. 4417749 contained hundreds of searches over a three-month period on topics ranging from “touch fingers” to “60 single men” to “dog that urinates on everything.”

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are questions for “landscapers in Lithuna, Ga.” several people with the last name Arnold and “horses sold in Shawnee” and a dogehr in the Homesteads of prominent country people.

It did not take much investigating to follow that trail to Thaddeus Arnold, a 60-year-old widower who lives in Lithuna, Ga., frequently researches her friends’ medical ailments and loves her three dogs. “Those are my searches,” she said, after a reporter read part of the list to her.

Could a new Netflix contest put private customer data at risk?

By Matthew Ewan / November 22, 2009

Back in 2006, Netflix announced it would give $1 million to the first team that could develop a predictive recommendations algorithm more accurate than the one currently used by Netflix. Long story short, this algorithm is the thing that suggests new DVDs for you to order. Based on your past viewing preferences. On Monday, the rental company shut out the cash to a multinational team of engineers calling themselves Netflix’s Pragmatic Chaos.

The contest was seen by many web analysts to be a free test of crowdsourcing—just about anyone could enter, and many hundreds of people did. This Netflix Prize has also been a publicity coup for Netflix, which got plenty of newspaper and blog coverage. But it was a new algorithm to boot. Now, a second contest proposed by Netflix has drawn fire from Paul Ohm, an Associate Professor of Law at the University of Colorado Law School who writes frequently on privacy issues.

The New York Times describes the competition this way:
But what is “privacy” not?

- Privacy is not releasing only “aggregate” statistics.
So what is privacy?

• Idea: Privacy is about promising people freedom from harm.
  – Attempt 1: “An analysis of a dataset $D$ is private if the data analyst knows no more about Alice after the analysis than he knew about Alice before the analysis.”
So what is privacy?

• Problem: Impossible to achieve with auxiliary information.
  – Suppose an insurance company knows that Alice is a smoker.
  – An analysis that reveals that smoking and lung cancer are correlated might cause them to raise her rates!

• Was her privacy violated?
  • This is exactly the sort of information we want to be able to learn...
    – This is a problem even if Alice was not in the database!
So what is privacy?

• Idea: Privacy is about promising people freedom from harm.
  – Attempt 2: “An analysis of a dataset D is private if the data analyst knows almost no more about Alice after the analysis than he would have known had he conducted the same analysis on an identical database with Alice’s data removed.”
So What is Differential Privacy?
Differential Privacy

[Dwork-McSherry-Nissim-Smith 06]
Differential Privacy

\(X\): The data universe.

\(\mathcal{D} \subseteq X\): The dataset (one element per person)

Definition: Two datasets \(\mathcal{D}, \mathcal{D}' \subseteq X\) are neighbors if they differ in the data of a single individual.
Differential Privacy

$X$: The data *universe*.

$D \subset X$: The dataset (one element per person)

**Definition:** An algorithm $M$ is $\epsilon$-differentially private if for all pairs of neighboring datasets $D, D'$, and for all outputs $x$:

$$\Pr[M(D) = x] \leq (1 + \epsilon) \Pr[M(D') = x]$$
Theorem (Postprocessing): If $M(D)$ is $\epsilon$-private, and $f$ is any (randomized) function, then $f(M(D))$ is $\epsilon$-private.
Definition: An algorithm $M$ is $\epsilon$-differentially private if for all pairs of neighboring datasets $D, D'^\uparrow$, and for all outputs $x$:

$$\Pr[M(D)=x] \leq (1+\epsilon) \Pr[M(D'^\uparrow)=x]$$
So...

Definition: An algorithm $M$ is $\epsilon$-differentially private if for all pairs of neighboring datasets $D, D'$, and for all outputs $x$:

$$\Pr[M(D) = x] \leq (1 + \epsilon) \Pr[M(D')] = x$$
Definition: An algorithm $M$ is $\epsilon$-differentially private if for all pairs of neighboring datasets $D, D'\uparrow$, and for all outputs $x$:

$$\Pr[M(D) = x] \leq (1 + \epsilon) \Pr[M(D') = x]$$
Some Useful Properties

Theorem (Composition): If $M \downarrow 1, \ldots, M \downarrow k$ are $\epsilon$-private, then:

$$M(D) \equiv (M \downarrow 1(D), \ldots, M \downarrow k(D))$$

is $k\epsilon$-private.
So...

You can go about designing algorithms as you normally would. Just access the data using differentially private “subroutines”, and keep track of your “privacy budget” as a resource. Private algorithm design, like regular algorithm design, can be modular.
Some simple operations: Answering Numeric Queries

Def: A numeric function $f$ has sensitivity $c$ if for all neighboring $D, D^{↑'}$:

$$|f(D) - f(D^{↑'})| \leq c$$

Write $s(f) \equiv c$

- e.g. “How many professors are in the building?” has sensitivity 1.
- “What fraction of people in the building are professors?” has sensitivity $1/n$. 
Some simple operations:
Answering Numeric Queries

The Laplace Mechanism:

\[ M\downarrow\text{Lap}(D,f,\epsilon) = f(D) + \text{Lap}(s(f)/\epsilon) \]

Theorem: \( M\downarrow\text{Lap}(\cdot,f,\epsilon) \) is \( \epsilon \)-private.
Some simple operations: Answering Numeric Queries

The Laplace Mechanism:

\[ M \downarrow \text{Lap}(D, f, \epsilon) = f(D) + \text{Lap}(s(f)/\epsilon) \]

**Theorem:** The expected error is \( s(f)/\epsilon \)

(can answer “what fraction of people in the building are professors?” with error 0.2%)
Some simple operations: Answering Non-numeric Queries

“What is the modal eye color in the room?”

\[ R = \{ \text{Blue, Green, Brown, Red} \} \]

• If you can define a function that determines how “good” each outcome is for a fixed input:
  – E.g.
  
  \[ q(D, \text{Red}) = \text{“fraction of people in } D \text{ with red eyes”} \]
Some simple operations: Answering Non-numeric Queries

\[ M\downarrow \text{Exp} (D,R,q, \epsilon): \]
Output \( r \in R \) w.p. \( \propto e^{2\epsilon \cdot q(D,r)} \)

**Theorem:** \( M\downarrow \text{Exp} (D,R,q, \epsilon) \) is \( s(q) \cdot \epsilon \)-private, and outputs \( r \in R \) such that:

\[
E[|q(D,r) - \max_{r^* \in R} q(D,r^*)|] \leq 2s(q)/\epsilon \cdot \ln|R|
\]

(can find a color that has frequency within 0.5% of the modal color in the building)
So what can we do with that?

Empirical Risk Minimization:
*i.e. almost all of supervised learning

Find $\theta$ to minimize:

$$L(\theta) = \sum_{i=1}^{n} \ell(\theta, (x^{\downarrow i}, y^{\downarrow i}))$$
Stochastic Gradient Descent

Let $\theta^1 = 0^d$
For $t=1$ to $T$:
   Pick $i$ at random. Let $g_t \leftarrow \nabla \ell (\theta^t, (x^i, y^i))$
   Let $\theta^{t+1} \leftarrow \theta^t - \eta g_t$

Convergence depends on the fact that at each round: $\mathbb{E}[g_t] = \nabla L (\theta)$
Private Stochastic Gradient Descent

Let $\theta^1 = 0^d$
For $t=1$ to $T$:
   Pick $i$ at random. Let $g \downarrow t \leftarrow \nabla \ell (\theta^t, (x \downarrow i, y \downarrow i)) + \text{Lap} (\sigma)^d$
   Let $\theta^{t+1} \leftarrow \theta^t - \eta \cdot g \downarrow t$

Still have: $\mathbb{E}[g \downarrow t] = \nabla \mathcal{L}(\theta)!$  
(Can still prove convergence theorems, and run the algorithm...)

Privacy guarantees can be computed from:
1) The privacy of the Laplace mechanism
2) Preservation of privacy under post-processing, and
3) Composition of privacy guarantees.
What else can we do?

• Statistical Estimation
• Graph Analysis
• Combinatorial Optimization
• Spectral Analysis of Matrices
• Anomaly Detection/Analysis of Data Streams
• Convex Optimization
• Equilibrium computation
• Computation of optimal 1-sided and 2-sided matchings
• Pareto Optimal Exchanges
• ...
Differential Privacy \Rightarrow Learning

Theorem*: An $\epsilon$-differentially private algorithm cannot overfit its training set by more than $\epsilon$.

*Lots of interesting details missing!
Choosing a Formalism: Statistical Queries

- A data universe $X$
- A distribution $P \in \Delta X$
- A dataset $D \subseteq X$ consisting of $n$ points $x \in X$ sampled i.i.d. from $P$. 
Choosing a Formalism: Statistical Queries

• A statistical query is defined by a predicate
  \[ \phi : X \rightarrow [0,1] \].

• The value of a statistical query is
  \[ \phi(P) = \mathbb{E}_{x \sim P} [\phi(x)] \]

• A statistical estimator is an algorithm for estimating statistical query:
  \[ A \downarrow D (\phi) \rightarrow [0,1] \]
Choosing a Formalism: Statistical Queries

Loses little generality. Captures, e.g.

• Means, variances, correlations, etc.
• Risk of a hypothesis:
  \[ R(h) = \mathbb{E}_{(x,y) \sim P} [L(h(x), y)] \]
• Gradient of risk of a hypothesis:
  \[ \nabla R(h) = \mathbb{E}_{(x,y) \sim P} [\nabla L(h(x), y)] \]
• Almost* all of PAC learning

*Except Parity functions
Choosing a Formalism: Statistical Queries

• Adaptively Chosen Queries:
Choosing a Formalism: Statistical Queries

- Adaptively Chosen Queries:
Choosing a Formalism: Statistical Queries

• Adaptively Chosen Queries:

A statistical estimator $A$ is $(\epsilon, \delta)$-accurate for sequences of $k$ adaptively chosen queries $\phi^\downarrow 1, ..., \phi^\downarrow k$ if for $\phi^\downarrow k$ and $P$, with probability $1 - \delta$:

$$\max_{i} |A^\downarrow D (\phi^\downarrow i) - \phi^\downarrow i (P)| \leq \epsilon.$$
A Baseline

• Non-Adaptive Queries:

The “empirical average mechanism”: $A\downarrow D (\phi) = \phi(D) := \frac{1}{n} \sum_{x \in D} \phi(x)$ can answer $k$ non-adaptive queries with $(0.01,0.01)$-accuracy where:

$$k = e^{\Theta(n)}$$
A Baseline

• Non-Adaptive Queries:

  • The “empirical average mechanism”: \( A \downarrow D (\phi) = \phi(D) := \frac{1}{n} \sum_{x \in D} \phi(x) \) can answer \( k \) adaptive queries with \((0.01,0.01)\)-accuracy where:

  \[ k = O(n) \]
Differential Privacy $\Rightarrow$ Learning

**Theorem:** [DFHPRR’15, BNSSSU’16]:

Let $A$ be a statistical estimator for adaptively chosen statistical queries. Let $P$ be any distribution, and let $D \sim P^n$. If:

1. $A$ is $(\epsilon, \epsilon \cdot \delta)$-differentially private, and
2. $A$ is $(\epsilon, \epsilon \cdot \delta)$-accurate with respect to the sample $D$, then:
   
   $A$ is $(O(\epsilon), O(\delta))$-accurate with respect to the distribution $P$. 
Applications

Using Independent Gaussian Perturbation

**Theorem**: There exists a simple, computationally efficient statistical estimator that can answer $k$ *adaptive* queries to non-trivial accuracy where:

$$k = \Theta (n^{1/2})$$

A *quadratic* improvement over the empirical average mechanism!
Applications

Using State of the Art Differentially Private Mechanisms

**Theorem:** There exists a statistical estimator that can answer $k$ adaptive queries to non-trivial accuracy where:

$$k = e^{\Theta \left( \frac{n}{\log|X|} \right)}$$

An exponential improvement if the data universe $X$ is finite and $n \gg \log|X|$.
Applications

Data → training data

Reusable holdout

unrestricted access

can be used *many* times adaptively

valid estimate every time you use the holdout
from numpy import *

def Thresholdout(sample, holdout, q, sigma, threshold):
    sample_mean = mean([q(x) for x in sample])
    holdout_mean = mean([q(x) for x in holdout])
    if abs(sample_mean - holdout_mean) < random.normal(threshold, sigma):
        # q does not overfit
        return sample_mean
    else:
        # q overfits
        return holdout_mean + random.normal(0, sigma)
Reusable holdout example

• Data set with $2n = 20,000$ rows and $d = 10,000$ variables. Class labels in $\{-1,1\}$

• Analyst performs **stepwise variable selection**:
  1. Split data into training/holdout of size $n$
  2. Select “best” $k$ variables on training data
  3. Only use variables also good on holdout
  4. Build linear predictor out of $k$ variables
  5. Find best $k = 10,20,30,\ldots$
Classification after feature selection

**No signal:** data are random gaussians
labels are drawn *independently* at random from \{-1,1\}

Thresholdout correctly detects overfitting!
Classification after feature selection

**Strong signal:** 20 features are mildly correlated with target remaining attributes are uncorrelated

Thresholdout correctly detects right model size!
So...

- Differential privacy provides:
  - A rigorous, provable guarantee with a strong privacy semantics.
  - A set of tools and composition theorems that allow for modular, easy design of privacy preserving algorithms.
  - *Protection against overfitting* even when privacy is not a concern.
Thanks!

To learn more:

• Our textbook on differential privacy:
  – Available for free on my website: http://www.cis.upenn.edu/~aaroth

• Connections between Privacy and Overfitting: