CS 499/599: Machine Learning Security
02.21: Privacy

Mon/Wed 12:00 – 1:50 pm

Sanghyun Hong

sanghyun.hong@oregonstate.edu

Oregon State University

SAIL
Secure AI Systems Lab

Checkpoint Presentation II (Akshith and Matt)

Notice

- Due dates
  - Written paper critique (21st)

- Sign-up (on Canvas)
  - Scribe lecture note [3 slots remain]
  - In-class paper presentation / discussion [2 slots remain]

- Notice
  - You can receive 50% of the total credits if you submit HW1 – 4 by **03.16**
  - Grading scheme (total: 143 pts = 120 pts + 23 pts)
    - **A:** 108 <= total <= 143
    - **B:** 96 <= total < 108
    - **C:** 84 <= total < 96
    - **D:** 72 <= total < 84
    - **F:** total < 72

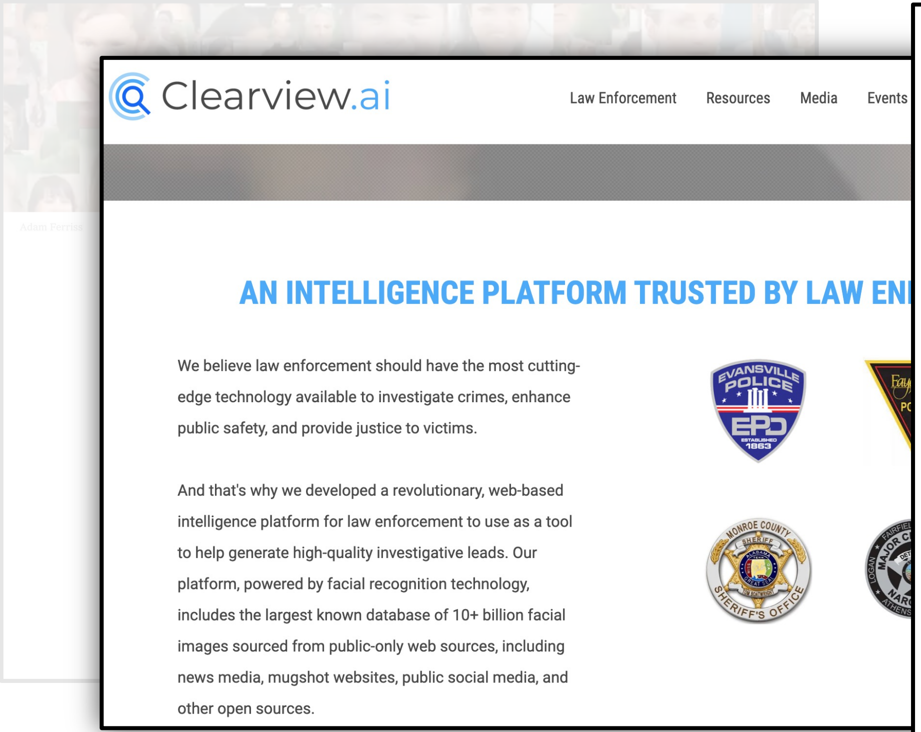
Topics for Today

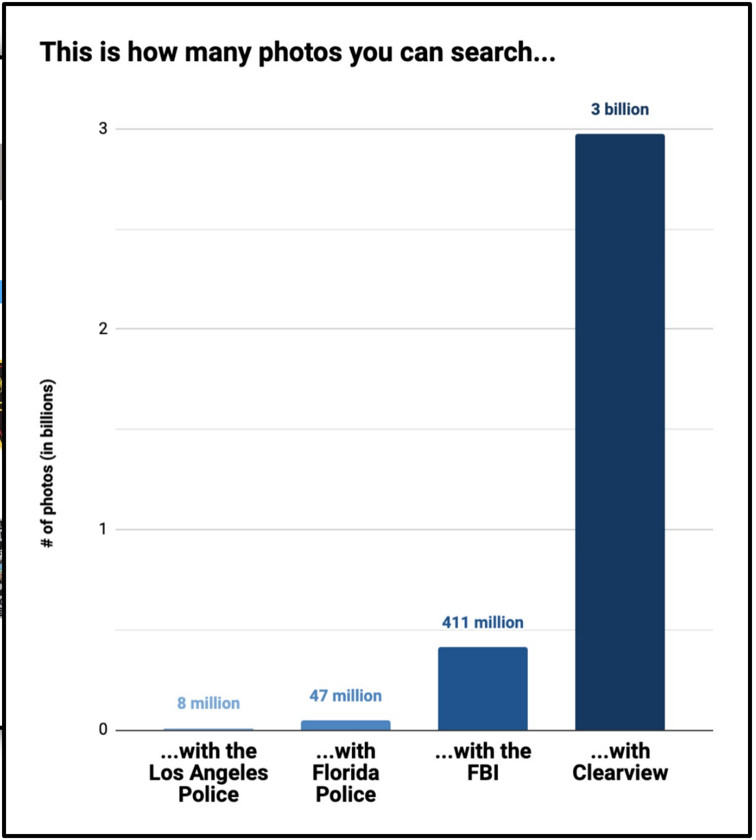
- Privacy
  - Warm-boot
  - Threat Models
    - Reconstruction attack
    - Tracing attack
    - Model extraction [controversial]
  - Differential privacy (DP)

- Privacy Attacks and Defenses
  - Non-ML
    - Data anonymization

Dwork *et al.*, Exposed! A Survey of Attacks on Private Data

Your Data Is Very Privately Managed!

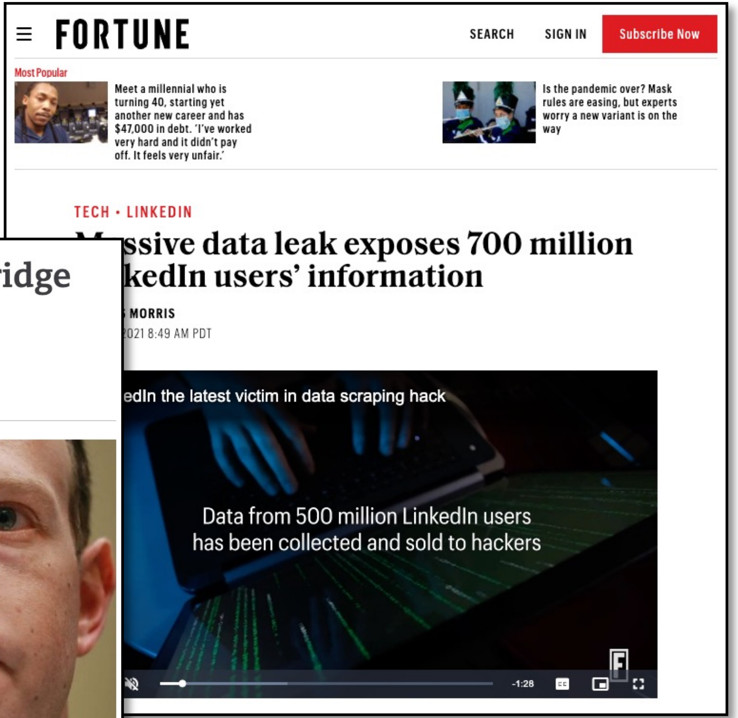




[1]https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html
[2]https://www.muckrock.com/news/archives/2020/jan/18/clearview-ai-facial-recogniton-records/

Privacy, Privacy, Privacy

- Let's do some discussions
  - What is privacy?
  - What does privacy matter?
  - How is it different from security?


FORTUNE

SEARCH SIGN IN Subscribe Now

Most Popular

Meet a millennial who is turning 40, starting yet another new career and has $47,000 in debt. 'I've worked very hard and it didn't pay off. It feels very unfair.'

Is the pandemic over? Mask rules are easing, but experts worry a new variant is on the way

TECH · LINKEDIN

M ssive data leak exposes 700 million kedIn users' information

MORRIS

021 8:49 AM PDT

edIn the latest victim in data scraping hack

Data from 500 million LinkedIn users has been collected and sold to hackers

-1:28


Facebook agrees to pay Cambridge Analytica fine to UK

30 October 2019

Facebook has agreed to pay a £500,000 fine imposed by the UK's data protection watchdog for its role in the Cambridge Analytica scandal.

Let's Talk A Threat Model to Study Privacy Risks!

Threat Model

- Goal
  - **Attacker:** extract some sensitive information about you (*e.g.*, data analyst in insurance firm)
  - **Victim** : minimize the leakage of such information (*e.g.*, your driving habits)
- Knowledge of the attacker
  - Additional (or auxiliary information) about the dataset D_{tr}
    - **Ex. :** Your friends on Facebook have 90% chances to drive recklessly
- Capability of the attacker
  - Query your data with some **mechanisms**
    - **Def:** a randomized algorithm M mapping datasets to an arbitrary set of outputs q
    - **Ex. :** how many times you were pulled over by police?
  - Perform post-processing computations on q (outputs)

Threat Model

- Privacy Attacks
  - **Re-identification**
    - **Goal:** de-identify anonymized datasets
    - **Ex.** : in an election poll, is this vote for President candidate A from you?
  - **Reconstructions**
    - **Goal:** reconstruct all the properties of a target instance in the dataset
    - **Ex.** : in the Census dataset, what are the attribute values associated with you?
  - **Tracing**
    - **Goal:** identify whether some instances are in the dataset or not
    - **Ex.** : do you participate in a clinical trial?
  - **[Note]**
    - Extract well-known facts or highly-correlated information is not the attacker's goal

Reconstruction Attack

- Setup
  - **Victim:**
    - For each i-th instance, the victim has (x_i, s_i) information
    - $x_i \in \{0, 1\}^d$: public info. accessible by an adversary and s_i: is the one-bit secret
  - **Attacker:**
    - Perform an attack A that reconstructs s_i by exploiting query outputs $\hat{q}$ and the public information $A\big(x, M(x, s)\big)$, where the attacker knows $k > 1$ public attributes
  - **Formally**

Reconstruction Attack

$x_i = [x_i(1), \dots, x_i(d), s_i]$

- Setup
  - **Victim:**
    - For each i-th instance, the victim has (x_i, s_i) information
    - $x_i \in \{0, 1\}^d$: public info. accessible by an adversary and s_i: is the one-bit secret
  - **Attacker:**
    - Perform an attack A that reconstructs s_i by exploiting query outputs $\hat{q}$ and the public information $A(x, M(x, s))$, where the attacker knows $k > 1$ public attributes

0.3 ← 0 : Trump
1 : Biden

q_i	s_i	pub. $[x_i(j) \dots x_i(j+k)]$	
0.3	1	0 1 0 1 0 0 1	x_1
0.5	0	1 0 0 1 0 0 1	x_2
0.6			
1.0			
0.1			

s_i \ $x_i(j)$	0	1
0	0	0
1	0	1

Q: predict s_i

Reconstruction Attack

- Setup
  - **Victim:**
    - For each i-th instance, the victim has (x_i, s_i) information
    - $x_i \in \{0,1\}^d$: public info. accessible by an adversary and s_i: is the one-bit secret
  - **Attacker:**
    - Perform an attack A that reconstructs s_i by exploiting query outputs $\hat{q}$ and the public information $A\big(x, M(x,s)\big)$, where the attacker knows $k > 1$ public attributes
  - **Approximation:**
    - Linear statistics (*e.g.*, linear SVM, linear regression, …)
    - Practical constraints (# Queries)
      - Ideally 2^n queries to solve the subset-sum problem
      - Practically, considering the tradeoff btw error and accuracy, we can do it in polynomial time

Tracing (less strong) Attack

- Setup
  - **Victim:**
    - Has a dataset x = $\{x_1, \ldots, x_n\}$ with n-i.i.d samples where each x_i is drawn from P over $\{\pm 1\}^d$
    - For each query M, the victim returns the sample mean q over given sample x_i's
  - **Attacker:**
    - Perform an attack $A(y, q, z)$ that identify whether a target instance $y \in \{\pm 1\}^d$ **IN** the dataset x or not (**OUT**) with m-i.i.d reference samples z = $\{z_1, \ldots, z_n\}$ and the sample mean q
  - **Procedure:**

Tracing (less strong) Attack

- Setup
  - **Victim:**
    - Has a dataset $x = \{x_1, \ldots, x_n\}$ with n-i.i.d samples where each x_i is drawn from P over $\{\pm 1\}^d$
    - For each query M, the victim returns the sample mean q over given sample x_i's
  - **Attacker:**
    - Perform an attack $A(y, q, z)$ that identify whether a target instance $y \in \{\pm 1\}^d$ **IN** the dataset x or not (**OUT**) with m-i.i.d reference samples $z = \{z_1, \ldots, z_n\}$ and the sample mean q
  - **Procedure:**

$\in x \quad \notin x$

$[y, z_1, z_2, \ldots, z_m]$

$[y, z_1, z_2] \longrightarrow \hat{q}_1$

$[z_1, z_2, z_3] \longrightarrow q_1$

$\vdots$

Topics for Today

- Privacy
  - Warm-boot
  - Privacy attacks:
    - Reconstruction attack
    - Tracing attack
    - Model extraction [controversial]
  - Defense: differential privacy (DP)

- Privacy Attacks and Defenses
  - Non-ML:
    - Data anonymization

Proposing Defenses

- Challenges
  - How can we define a privacy guarantee?
    - **Problem:** Adversaries may *break* some heuristic defenses (arms-race)
    - **Example:** A defense and its pitfall:
      - In DB query responses, a defender can randomly drop k rows ($k \ll r$, r: # rows in resp.)
      - One can submit the same query multiple times, and then they compares responses
  - What if we apply the strongest privacy guarantee?
    - **Problem:**
      - Well, if you do not share, you do not leak any information
      - But it is *NOT* what we want (the end of arms-race)
  - How can we offer an upper-bound of privacy leakage?
    - **Problem:** It is hard to define what is the leakage of private information
    - **Example:** Many definitions are feasible (*e.g.*, certain attributes, specific samples, etc…)

Proposing Defenses: Differential Privacy

- Differential Privacy (DP)
  - How can we offer an upper-bound of privacy leakage?
    - Focus on the **smallest** perturbations on a dataset we protect: **a single instance**
    - Make the outputs of **any** algorithms (*e.g.*, query processing) compute on datasets with a single item difference **cannot be different** from each other **with ε probability**
  - Formally,
    - An algorithm (or a mechanism) M satisfies **ε-differential privacy** if, for any datasets x and y differing only on the data of a single instance and any potential outcome $\hat{q}$,

$$\mathbb{P}\left[\mathcal{M}(x) = \hat{q}\right] \leq e^{\varepsilon} \cdot \mathbb{P}\left[\mathcal{M}(y) = \hat{q}\right].$$

Proposing Defenses: Differential Privacy

- Differential Privacy (DP)
  - How can we offer an upper-bound of privacy leakage?
    - Focus on the **smallest** perturbations on a dataset we protect: **a single instance**
    - Make the outputs of **any** algorithms (*e.g.*, query processing) compute on datasets with a single item difference **cannot be different** from each other **with ε probability**
  - Formally,
    - An algorithm (or a mechanism) M satisfies **ε-differential privacy** if, for any datasets x and y differing only on the data of a single instance and any potential outcome $\hat{q}$,

$$\mathbb{P}[\mathcal{M}(x) = \hat{q}] \leq e^{\varepsilon} \cdot \mathbb{P}[\mathcal{M}(y) = \hat{q}].$$

$E = -\log \frac{1}{P[x]}$

$\ln P[Mx = \hat{q}] \leq \varepsilon + \ln P[My = \hat{q}]$

$E_x \quad -\ln \underbrace{\frac{1}{P[M(x) = \hat{q}]}} \leq \varepsilon - \ln \underbrace{\frac{1}{P[M(y) = \hat{q}]}} \quad E_y$

$E.$

$E_x - E_y \leq \varepsilon.$

Proposing Defenses: Differential Privacy

- 3 Important Properties of DP
  - DP-Definition
    - An algorithm (or a mechanism) M satisfies ε-differential privacy if, for any datasets x and y differing only on the data of a single instance and any potential outcome $\hat{q}$,

$$\mathbb{P}\left[\mathcal{M}(x)=\hat{q}\right] \leq e^{\varepsilon} \cdot \mathbb{P}\left[\mathcal{M}(y)=\hat{q}\right].$$

  - **Post-processing**
    - Any **post-processing** of differentially-private data **won't change the DP guarantee**
  - **Composition**
    - If the **same instance in multiple datasets** (where each satisfies ε-DP), **the combination** of those releases also satisfies **$k\varepsilon$-DP** (*i.e.*, the guarantees will degrade by k)
  - **Group-privacy**
    - If we want **to protect k instances**, instead of a single item, we require **$k\varepsilon$-DP** guarantee

Proposing Defenses: Differential Privacy

- Implementation
  - DP-Definition
    - An algorithm (or a mechanism) M satisfies ε-differential privacy if, for any datasets x and y differing only on the data of a single instance and any potential outcome $\hat{q}$,

$$\mathbb{P}\left[\mathcal{M}(x) = \hat{q}\right] \leq e^{\varepsilon} \cdot \mathbb{P}\left[\mathcal{M}(y) = \hat{q}\right].$$

  - **Gaussian mechanism**-Definition
    - **Formally:** Suppose properties $q = (q_1, \ldots, q_k)$, the Gaussian mechanism M_{q,σ^2} takes x as input and releases $\hat{q} = (\widehat{q_1}, \ldots, \widehat{q_k})$ where each $\widehat{q_i}$ is independent sample from $N(q_i(x), \sigma^2)$, for an appropriate variance σ^2
    - **Easy-way:** I will **add Gaussian noise** with a variance σ^2 **to the output $\hat{q}$**, such that the output **satisfies ε-differential privacy** guarantee

Recap!

- Privacy
  - Warm-boot
  - Privacy attacks:
    - Reconstruction attack
    - Tracing attack
    - Model extraction [controversial]
  - Defense: differential privacy (DP)

- Privacy Attacks and Defenses
  - Non-ML:
    - Data anonymization **[Left for you]**

Thank You!

Mon/Wed 12:00 – 1:50 pm

Sanghyun Hong

https://secure-ai.systems/courses/MLSec/W22

Oregon State University

SAIL
Secure AI Systems Lab