PETs, POTs, and Pitfalls

Rethinking the Protection of Users against Machine Learning

Carmela Troncoso
@carmelatroncoso
Security and Privacy Engineering Lab
The machine learning revolution
The machine learning tsunami

Privacy

Social Justice
The ML tsunami on privacy

Introducing the idea
(see notes for details)

Privacy


Detecting neurodegenerative disorders from web search signals
Ryan M. White, P. Noorali Durrani & A. R. M. Norwitz
my Digital Medicine, Article number: 6 (2016) | Download Citation

Abstract
Facebook Filed A Patent To Predict Your Household’s Demographics Based On Family Photos
Facebook's proposed technology would analyze your Instagram tags, shared IP addresses, and photos to predict who you live with.

Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States
Trent Gruber, Jonathan Krause, Yun Cong, Duyun Chen, Jia Diao, Ezie Lieberman Adian, and Li Fei-Fei
Published December 12, 2017 114 (5) 1319-1311; published ahead of print November 28, 2017

On the Feasibility of Internet-Scale Author Identification
Arvind Narayanan, Hristo Paskov, Neil Zhengjiang Gong, John Bethencourt, Eui Chul Richard Shin, Emily Stanford, and Daon Song

Abstract—We study techniques for identifying an anonymous author via linguistic stochasticity, i.e., comparing the writing style against a corpus of texts of known authorship. We experimentally demonstrate the effectiveness of our techniques with as many as 180,000 candidate authors. Given the increasing availability of writing samples online, our result has serious implications for anonymity and free speech — an anonymous blogger or whistle-blower may be unmasked unless they take steps to obfuscate their writing style.

While there is a line of work on literature on authorship
Yet a right to anonymity is meaningless if an anonymous author’s identity can be unmasked by adversaries. There have been many attempts to legally force service providers and other intermediaries to reveal the identity of anonymous users. While sometimes successful [5, 8], in most cases they have upheld a right to anonymous speech [7]. All of these efforts have relied on the author writing their name or IP address to a service provider, who may in turn reveal their

Combining AI and Location Intelligence to predict Market Demand
Cindy Elliott Contributor
esri Contributor Group

AI can predict your future tweets by looking at your friends’ accounts
A new study shows how machine-learning methods could examine your friends’ past tweets to accurately predict your future behavior online.
Attacks are not new... but the adversary is

Privacy Enhancing Technologies

PETs
Attacks are not new... but the adversary is

PETs??
The goal is not to understand, it is to beat!
The goal is not to understand, it is to beat!
Adversarial examples are only adversarial when you are the owner of the algorithm!
Adversarial examples are only 

**adversarial** when you are the owner of the algorithm!

**PETS!!**

**Ally**

**Enemy**

**ML models**
Wait! Why do we need adversarial examples if we have privacy-preserving ML!!
Machine learning as a privacy adversary

ML Privacy-oriented Literature

Actively (maybe not willingly) provide data. Solutions like Differential privacy and Encryption are suitable

Avoid that learns about data
Machine learning as a privacy adversary

ML Privacy-oriented Literature

Actively (maybe not willingly) provide data. Solutions like Differential privacy and Encryption are suitable

Avoid that

learns about data

Data

Service

In this talk

No active sharing!

Cannot count on
Use ML adversarial example techniques to transform data!
Adversarial examples as privacy defenses

Can this solve all privacy problems?

Use ML adversarial example techniques to transform data!
Can this solve all privacy problems?

Protect web searches from inferences

Protect tweets from inferences

Protect traffic patterns

???

???

???
Can this solve all privacy problems?

In **privacy problems** adversarial examples belong to a **DISCRETE** and **CONSTRAINED** domain.

- **Feasibility**
- **Cost**

Protect web searches from inferences

Protect tweets from inferences

Protect traffic patterns
AttriGuard: A Practical Defense Against Attribute Inference Attacks via Adversarial Machine Learning

Jinyuan Jia
ECE Department, Iowa State University
jinyuan@iastate.edu

Neil Zhengjiang Gong
ECE Department, Iowa State University
neilgong@iastate.edu

Abstract

Users in various web and mobile applications are vulnerable to attribute inference attacks, in which an attacker leverages a machine learning classifier to infer a target user’s private attributes (e.g., location, sexual orientation, political view) from its public data (e.g., rating scores, mobile platforms [10, 11]). In an attribute inference attack, an attacker aims to infer a user’s private attributes (e.g., location, gender, sexual orientation, and/or political view) via leveraging its public data. For instance, in social media, a user’s public data could be the list of pages that the user liked on Facebook. Given these pages, an attacker can use a machine learning classifier to

Nobody has thought of this?

Usenix Security Symposium - 2018

Modify social network attributes to avoid inferences

Use adversarial examples (evasion attacks) to keep utility

Use a version of Jacobian-based Saliency Map Attack (JSMA) “aware of policies” = only do feasible transformations
Nobody has thought of this?

**AttriGuard: A Practical Defense Against Attribute Inference Attacks via Adversarial Machine Learning**

Jinyuan Jia  
ECE Department, Iowa State University  
jinyuan@iastate.edu

Neil Zhengiang Gong  
ECE Department, Iowa State University  
eilgong@iastate.edu

**Abstract**  
Users in various web and mobile applications are vulnerable to attribute inference attacks, in which an attacker leverages a machine learning classifier to infer a target user’s private attributes (e.g., location, sexual orientation, political view) from its public data (e.g., rating scores, mobile platforms [10, 11]). In an attribute inference attack, an attacker aims to infer a user’s private attributes (e.g., location, gender, sexual orientation, and/or political view) via leveraging its public data. For instance, in social media, a user’s public data could be the list of pages that the user liked on Facebook. Given these page likes, an attacker can use a machine learning classifier to

**Usenix Security Symposium - 2018**

Modify social network attributes to avoid inferences

Use adversarial examples (evasion attacks) to keep utility

Use a version of Jacobian-based Saliency Map Attack (JSMA) “aware of policies” = only do feasible transformations

**PoPETS - 2019**

Modify Twitter line to avoid inferences

Add, remove, replace tweets

Greedy search by importance for classifier

“Because… I was told… so much”: Linguistic Indicators of Mental Health Status on Twitter

Abstract: Recent studies have shown that machine learning can identify individuals with mental illnesses by analyzing their social media posts. Topics and words related to mental health are some of the top predictors. These findings have implications for early detection of mental illnesses. However, they also raise numerous privacy concerns. To fully evaluate the implications for privacy, we analyze the performance of different machine learning models in the absence of tweets that talk about mental illnesses. Our results show that machine learning can be used to make predictions even if the users deviate from normal language use, and that these deviations can be used as a diagnostic tool. While early studies analyzed this relationship via patient essays and interview transcripts, recent studies have shown that similar changes in language usage can also be detected in social media posts. Moreover, more recent studies have shown that machine learning can predict the mental status of individuals through the content of their social media posts [17].

The 2015 ACL Workshop on Computational Linguistics and Clinical Psychology built a dataset con-
 Nobody has thought of this?

**Non-privacy constrained applications**

**Text:**
Goal: change classification (positive to negative sentiment, change inferred topic for a post)

**Malware:**
Goal: change classification (from malicious to benign)
Nobody has thought of this?

Non-privacy constrained applications

Text:
Goal: change classification (positive to negative sentiment, change inferred topic for a post)

Malware:
Goal: change classification (from malicious to benign)

Repeated patterns:
- Model transformation
- Find new search algorithm e.g., Hill climbing, beam search
- Evaluate & compare performance

But NO systematic design method ☹️
Our proposal: Evasion as a graph

Protecting users from demographic inferences

**Goal** change Twitter line classification regarding age

**Transformations**
- Use synonyms
- Introduce typos
- Change punctuation

**Cost**
- Keep the meaning!
Our proposal: Evasion as a graph

Cost: keep meaning

I love Justin Bieber!
Our proposal: Evasion as a graph
Cost: keep meaning

I love Justin Bieber!

1
I love Justin Bieber.

2
I like Justin Bieber

20
I loath Justin Bieber

Tweets from an account

ML
Our proposal: Evasion as a graph

Cost: keep meaning

Tweets from an account

I love Justin Bieber!

1

I love Justin Bieber.

2

I like Justin Bieber

20

I loath Justin Bieber

20

I love Justin Trudeau.

cost = 1 + 20 = 21

20

I love Justin Timberlake.

cost = 1 + 5 = 6
Our proposal: Evasion as a graph

Cost: keep meaning

In privacy problems examples belong to a DISCRETE and CONSTRAINED domain

Feasibility ✓
Cost ✓

I love Justin Bieber.

I love Justin Trudeau.

I love Justin Timberlake.

Cost: keep meaning

cost = 1 + 20 = 21

cost = 1 + 5 = 6
The graph approach comes with advantages

✅ Enables the use of graph theory to
   **EFFICIENTLY** find adversarial examples (A*, beam search, hill climbing, etc)
   **CAPTURES** most attacks in the literature! (comparison base)

✅ Finds provable **MINIMAL COST** adversarial examples (A*) if

   - The discrete domain is a subset of $\mathbb{R}^m$
     For example, categorical one-hot encoded features: [0 1 0 0]

   - Cost of each single transformation is $L^p$
     For example, $L^\infty([0 1 0 0], [1 0 0 0]) = 1$

   - We can compute pointwise robustness for the target classifier over $\mathbb{R}^m$
Finding minimal cost adv. examples: the concept

\[ x^* = \arg \min_{x' \in \mathbb{X}} C(x, x') \text{ s.t. } \text{goal}(x') = \top, \]

\[ \text{goal}(x') = \begin{cases} 
\top, & t = 1 \text{ and } \sigma(f(x')) > l \\
\top, & t = 0 \text{ and } \sigma(f(x')) \leq 1 - l \\
\bot, & \text{otherwise}
\end{cases} \]

Confidence of the example
Adversarial examples for privacy

✓ Provide privacy in settings where the ML model is adversarial and not cooperative

✓ Privacy is CONSTRAINED, a graphical approach can be used to EFFICIENTLY find FEASIBLE adversarial examples find MINIMAL COST adversarial examples

✓ Even if they cannot be deployed in practice, this approach provides a BASELINE to compare defenses’ efficiency
**Bonus: applicable to security problems!**

**MINIMAL COST** adversarial examples can become security metrics!

Cost can be associated with **RISK**

Cannot stop attacks, but can we ensure they are expensive?

Constrained domains security

Continuous-domains approaches can be very conservative!
Only privacy is at stake?

Privacy breaches
Only privacy is at stake?

Privacy breaches

Data used to optimize ...
Only privacy is at stake?

Prevalent use of optimization algorithms to extract maximum economic value from the manipulation of people's activities and their environment.

Privacy breaches

Data used to optimize...

Advertisement (e.g., Facebook ads)

Routing (e.g., Waze)

Credit scoring (e.g., FICO)
The ML tsunami on Social Justice

Data Scores as Governance: Investigating uses of citizen scoring in public services

Navigation Apps Are Turning Quiet Neighborhoods Into Traffic Nightmares

Social Sorting as a Tool for Surveillance

The female body is constantly under surveillance - in private spaces as well as in public. Surveillance is about power. It is not just about a violation of privacy, but also an issue of social sorting.

The ML tsunami on Social Justice
Optimization Systems

interact with

Optimization system

Algorithm

Benefit
Optimization Systems
Optimization Systems

Optimization system interact with affect
Optimization Systems

Interact with Optimization system

Affect
Optimization Systems

interact with

Optimization system

Algorithm
Algorithm

affect
Optimization Systems

interact with

Optimization system

Algorithm ... Algorithm

affect
Optimization Systems

interact with

Optimization system

Algorithm

... Algorithm

affect
Optimization Systems

non-users  users

interact with

Optimization system

Algorithm

Algorithm

affect

users  non-users
Optimization Systems

How do we avoid negative effects caused by the Optimization System? (direct and externalities)
We have fairness research!!

“We’re creating algorithms that cause harms, so we need to fix the algorithms”
We have fairness research!!

“We’re creating algorithms that cause harms, so we need to fix the algorithms”

Limited to algorithmic bias within a system

Assumes ML owners have the incentives and the means

Decontextualized from the system’s goal

Ignores other harms

A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry. Os Keyes, Jevan Hutson, and Meredith Durbin
Wait! But we have fairness research!!

A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry. Os Keyes, Jevan Hutson, and Meredith Durbin


Limited to algorithmic bias within a system

Decontextualized from the system’s goal

Assumes ML owners have the incentives and the means

Ignores other harms

Fairness vs. Optimization Systems harms

- disregard non-users and environmental impact
- benefit a few
- fairness
- distributional shift
- distribution of errors
- exploration risks
- reward hacking
- mass data collection
- all while potentially optimizing for asocial behavior or negative environmental outcomes
Protective Optimization Technologies (POTs)

Technologies aimed at mitigating externalities of optimization system’s
Credit scoring

Potential risk posed by lending money to consumers and to mitigate losses due to bad debt

**Biased training data** → Underlying algorithms can:
- discriminate applicants on protected attributes like gender or ethnicity
- cause feedback loops for populations disadvantaged by the financial system

Credit bureaus have little incentive to change
Fairness techniques are incipient and hard to deploy
POTs for Credit scoring

- Enable users to help others get loans
POTs for Credit scoring

- Enable users to help others get loans
- Enable discriminated users to get loans

Poisoning
POTs for Credit scoring

- Enable users to help others get loans
- Enable discriminated users to get loans

Poisoning

Take loans & repay

Bureau

Adversarial examples

DISCRETE AND CONSTRAINED!
Adversarial machine learning for social justice

✅ There is a need to protect individuals beyond preserving their privacy

✅ Protective Optimization Technologies can be deployed to help individuals and groups to counter externalities

✅ POTs are also CONSTRAINED so the graphical approach can also be used as technique to EFFICIENTLY find MINIMAL COST adversarial examples
A challenge ahead
Disparate vulnerability

• Machine learning models inherit biases in the training

• Two Key implications
  • ML-based attacks are unfair
    (like any ML-based model...)

Table 3: Classifier Performance

<table>
<thead>
<tr>
<th>Sample</th>
<th>Gender</th>
<th>Precision</th>
<th>Recall</th>
<th>AUC</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Male</td>
<td>0.817</td>
<td>0.754</td>
<td>0.784</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>0.667</td>
<td>0.744</td>
<td>0.76</td>
<td>0.750</td>
</tr>
<tr>
<td>2</td>
<td>Male</td>
<td>0.727</td>
<td>0.615</td>
<td>0.681</td>
<td>0.666</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>0.528</td>
<td>0.651</td>
<td>0.666</td>
<td>0.630</td>
</tr>
<tr>
<td>3</td>
<td>Male</td>
<td>0.849</td>
<td>0.692</td>
<td>0.802</td>
<td>0.756</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>0.636</td>
<td>0.814</td>
<td>0.741</td>
<td>0.714</td>
</tr>
<tr>
<td>4</td>
<td>Male</td>
<td>0.733</td>
<td>0.704</td>
<td>0.776</td>
<td>0.728</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>0.596</td>
<td>0.791</td>
<td>0.704</td>
<td>0.704</td>
</tr>
<tr>
<td>5</td>
<td>Male</td>
<td>0.704</td>
<td>0.769</td>
<td>0.674</td>
<td>0.667</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>0.595</td>
<td>0.512</td>
<td>0.709</td>
<td>0.667</td>
</tr>
</tbody>
</table>
Disparate vulnerability

- Machine learning models inherit biases in the training

- Two Key implications
  - ML-based attacks are unfair
  - Attacks on ML-models are unfair!
Disparate vulnerability

• Is increased when defending ML models from other shortcomings
Disparate vulnerability

• Is increased when defending ML models from other shortcomings
Disparate vulnerability

• And blanket defenses have disparate impact on utility!
Universal design for protection technologies

We need to take into account attack’s fairness when designing protections

• Is it possible to have secure accurate models with fair privacy?
  • Security vs. privacy trade-off?
  • More importantly: fair privacy at the cost of privacy?

• Are adversarial learning-based defenses immune to this issue?
  • If so, should they be our only way forward?

• Should fairness be a bullet in privacy by design beyond ML?
Takeaways

- Adversarial machine learning is hard to defend from: a great opportunity!

  Adversarial machine learning as protective technologies for privacy (PETs) and social justice (POTs)

- New graphical framework to approach the search of adversarial examples
  ... we can use of graph theory to improve efficiency and provide guarantees

- The fairness problems of machine learning will become a hurdle for protection!
EPFL

http://carmelatroncoso.com/
https://spring.epfl.ch/en

https://github.com/spring-epfl/

Bogdan Kulynych
Mohammad Yaghini
Seda Guerses
Rebekah Overdorf
Ero Balsa
Jamie Hayes
Nikita Samarin