On the Impossible Security of Very Large Foundation Models

Lê Nguyên Hoang,
Calicarpa & Tournesol,
FLAIM, IHP, November 2022

Section 1

Impossibility theorems in ML security

The framework

Signed data

We consider a set $[N] = \{1, 2, \ldots, N\}$ of data sources (users).
Each source $n \in [N]$ provides a *signed* dataset $\mathcal{D}_n$.
We denote $\overrightarrow{\mathcal{D}} = (\mathcal{D}_1, \ldots, \mathcal{D}_N)$ the tuple of source's datasets.

The framework

Signed data

We consider a set $[N] = \{1, 2, \ldots, N\}$ of data sources (users).
Each source $n \in [N]$ provides a *signed* dataset $\mathcal{D}_n$.
We denote $\overrightarrow{\mathcal{D}} = (\mathcal{D}_1, \ldots, \mathcal{D}_N)$ the tuple of source's datasets.

Performance measure

Minimize $\textsc{Loss}(\theta|\overrightarrow{\mathcal{D}}) \triangleq \sum_{n \in [N]} \mathcal{L}(\theta|\mathcal{D}_n) + \mathcal{R}(\theta)$.

The framework

Signed data

We consider a set $[N] = \{1, 2, \dots, N\}$ of data sources (users).
Each source $n \in [N]$ provides a *signed* dataset $\mathcal{D}_n$.
We denote $\overrightarrow{\mathcal{D}} = (\mathcal{D}_1, \dots, \mathcal{D}_N)$ the tuple of source's datasets.

Performance measure

Minimize $\textsc{Loss}(\theta|\overrightarrow{\mathcal{D}}) \triangleq \sum_{n \in [N]} \mathcal{L}(\theta|\mathcal{D}_n) + \mathcal{R}(\theta)$.

Privacy constraints

User-level differential privacy: $\mathbb{P}[\textsc{Learn}(\overrightarrow{\mathcal{D}}) \in S] \leq e^{\varepsilon} \mathbb{P}[\textsc{Learn}(\overrightarrow{\mathcal{D}}_{-n}) \in S] + \delta$.

The framework

Signed data

We consider a set $[N] = \{1, 2, \dots, N\}$ of data sources (users).
Each source $n \in [N]$ provides a *signed* dataset $\mathcal{D}_n$.
We denote $\overrightarrow{\mathcal{D}} = (\mathcal{D}_1, \dots, \mathcal{D}_N)$ the tuple of source's datasets.

Performance measure

Minimize $\textsc{Loss}(\theta|\overrightarrow{\mathcal{D}}) \triangleq \sum_{n\in[N]} \mathcal{L}(\theta|\mathcal{D}_n) + \mathcal{R}(\theta)$.

Privacy constraints

User-level differential privacy: $\mathbb{P}[\textsc{Learn}(\overrightarrow{\mathcal{D}}) \in S] \leq e^{\varepsilon}\mathbb{P}[\textsc{Learn}(\overrightarrow{\mathcal{D}}_{-n}) \in S] + \delta$.

Security constraints

Resilience to data poisoning: $\forall H \subset [N],\ \textsc{Loss}(\textsc{Learn}(\overrightarrow{\mathcal{D}})|\overrightarrow{\mathcal{D}}_H)$ small.

An equivalence between data poisoning and gradient attacks

Performance measure

Minimize $\textsc{Loss}(\theta|\vec{\mathcal{D}}) \triangleq \sum_{n\in[N]} \mathcal{L}(\theta|\mathcal{D}_n) + \mathcal{R}(\theta)$.

An equivalence between data poisoning and gradient attacks

Performance measure

Minimize $\textsc{Loss}(\theta|\vec{\mathcal{D}}) \triangleq \sum_{n \in [N]} \mathcal{L}(\theta|\mathcal{D}_n) + \mathcal{R}(\theta)$.

Personalized federated learning

Each source n is given a personalized model φ_n:
$\mathcal{L}(\theta|\mathcal{D}_n) \triangleq \inf_{\varphi_n} \left\{ \mathcal{R}_n(\varphi_n, \theta) + \sum_{(y,z) \in \mathcal{D}_n} \ell(f_{\varphi_n}(y), z) \right\}$.

An equivalence between data poisoning and gradient attacks

Performance measure

Minimize $\textsc{Loss}(\theta|\vec{\mathcal{D}}) \triangleq \sum_{n\in[N]} \mathcal{L}(\theta|\mathcal{D}_n) + \mathcal{R}(\theta)$.

Personalized federated learning

Each source n is given a personalized model φ_n:
$\mathcal{L}(\theta|\mathcal{D}_n) \triangleq \inf_{\varphi_n} \left\{ \mathcal{R}_n(\varphi_n, \theta) + \sum_{(y,z)\in\mathcal{D}_n} \ell(f_{\varphi_n}(y), z) \right\}$.

Theorem (Farhadkhani, Guerraoui, H and Villemaud (ICML 2022))

Assume $\mathcal{R}_n$ is ℓ_2^2, ℓ_2 or smooth-ℓ_2, and assume ℓ does logistic or linear regression (and consider $\mathcal{R}$ is convex). Fix $\vec{\mathcal{D}}_{-n}$. Consider any converging (admissible) gradient attack $g_n^t \to g_n^\spadesuit$ by source n, implying a learned model $\theta^\dagger$. Then, for any $\varepsilon > 0$, there exists a poisoning dataset $\mathcal{D}_n^\spadesuit$ such that $||\theta^\dagger - \theta^*(\vec{\mathcal{D}}_{-n}, \mathcal{D}_n^\spadesuit)||_2 \leq \varepsilon$.

Proof sketch

Lemma (easy)

$g_n^{\spadesuit}$ is equivalent to an attack model $\varphi_n^{\spadesuit}$, reconstructible from $\theta^{\spadesuit}$.

Proof sketch

Lemma (easy)

$g_n^\spadesuit$ is equivalent to an attack model $\varphi_n^\spadesuit$, reconstructible from $\theta^\spadesuit$.

Lemma (easy)

Assuming local PAC learning, $\varphi_n^\spadesuit$ is equivalent to a poisoning dataset $\mathcal{D}_n^\spadesuit$, which labels randomly drawn data with model $\varphi_n^\spadesuit$.*

Proof sketch

Lemma (easy)

$g_n^\spadesuit$ is equivalent to an attack model $\varphi_n^\spadesuit$, reconstructible from $\theta^\spadesuit$.

Lemma (easy)

Assuming local PAC learning, $\varphi_n^\spadesuit$ is equivalent to a poisoning dataset $\mathcal{D}_n^\spadesuit$, which labels randomly drawn data with model $\varphi_n^\spadesuit$.*

Lemma (not difficult)

Gradient PAC of ℓ implies local PAC* learning.*

Proof sketch

Lemma (easy)

$g_n^\spadesuit$ is equivalent to an attack model $\varphi_n^\spadesuit$, reconstructible from $\theta^\spadesuit$.

Lemma (easy)

Assuming local PAC learning, $\varphi_n^\spadesuit$ is equivalent to a poisoning dataset $\mathcal{D}_n^\spadesuit$, which labels randomly drawn data with model $\varphi_n^\spadesuit$.*

Lemma (not difficult)

Gradient PAC of ℓ implies local PAC* learning.*

Lemma (not easy)

Logistic and linear regression with spanning random features satisfy gradient PAC.*

Gradient PAC*

Definition

Let $\mathcal{E}(\mathcal{D}, \varphi^\dagger, \mathcal{I}, A, B, \alpha)$ defined by

$$\forall \varphi \in \mathbb{R}^d, (\varphi - \varphi^\dagger)^T \sum_{(y,z)} \nabla \ell(f_\varphi(y), z) \geq A\mathcal{I} \min \left\{ ||\varphi - \varphi^\dagger||_2, ||\varphi - \varphi^\dagger||_2^2 \right\} - B\mathcal{I}^\alpha ||\varphi - \varphi^\dagger||_2.$$

The loss ℓ is gradient-PAC* if, for any $K > 0$, there exists $A_K, B_K > 0$ and $\alpha_K < 1$ such that, for any $\varphi^\dagger \in \mathcal{B}(0, K)$, assuming that the dataset $\mathcal{D}$ is obtained by honestly collecting and labeling $\mathcal{I}$ data points according to the preferred model $\theta^\dagger$, the probability of $\mathcal{E}(\mathcal{D}, \varphi^\dagger, \mathcal{I}, A_K, B_K, \alpha_K)$ goes to 1 as $\mathcal{I} \to \infty$.

Gradient PAC*

Definition

Let $\mathcal{E}(\mathcal{D}, \varphi^\dagger, \mathcal{I}, A, B, \alpha)$ defined by

$$\forall \varphi \in \mathbb{R}^d, (\varphi - \varphi^\dagger)^T \sum_{(y,z)} \nabla \ell(f_\varphi(y), z) \geq A\mathcal{I} \min \left\{ ||\varphi - \varphi^\dagger||_2, ||\varphi - \varphi^\dagger||_2^2 \right\} - B\mathcal{I}^\alpha ||\varphi - \varphi^\dagger||_2.$$

The loss ℓ is gradient-PAC* if, for any $K > 0$, there exists $A_K, B_K > 0$ and $\alpha_K < 1$ such that, for any $\varphi^\dagger \in \mathcal{B}(0, K)$, assuming that the dataset $\mathcal{D}$ is obtained by honestly collecting and labeling $\mathcal{I}$ data points according to the preferred model $\theta^\dagger$, the probability of $\mathcal{E}(\mathcal{D}, \varphi^\dagger, \mathcal{I}, A_K, B_K, \alpha_K)$ goes to 1 as $\mathcal{I} \to \infty$.

Lemma

Gradient PAC of ℓ implies local PAC* learning. Moreover, logistic and linear regression with spanning random features satisfy gradient PAC*.*

Our theoretical equivalence yields practical attacks!

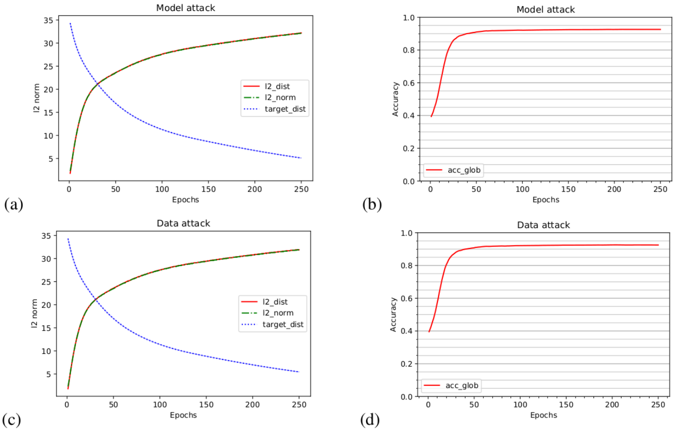


Figure 2. (a) Distance between ρ^t and $\theta_s^\dagger$ (target_dist), under model attack (combining CGA and Proposition 4). (b) Accuracy of ρ^t according to $\theta_s^\dagger$ (which relabels $0 \to 1 \to 2 \to \ldots \to 9 \to 0$), under model attack (combining CGA and Proposition 4). (c) Distance between the global model ρ^t and the target model $\theta_s^\dagger$ (target_dist), under our data poisoning attack. (d) Accuracy of ρ^t according to $\theta_s^\dagger$ (which relabels $0 \to 1 \to 2 \to \ldots \to 9 \to 0$), under our data poisoning attack.

What about neural nets?

- Gradient PAC* does not hold for neural nets.
- But gradient PAC* holds for most last-layer fine-tuning.
- One (minor) challenge is to generate a spanning distribution of embeddings.

Another equivalence in secure ML

Theorem (El-Mhamdi, Farhadkhani, Guerraoui, Guirguis, H & Rouault (NeurIPS 2021))

C-collaborative learning is equivalent to C-averaging.
Roughly, the guarantee on the norm of the true gradient at termination for collaborative learning can only be as good as the guarantee we can have when estimating the average of a set of vectors, assuming that some data source / vector providers are Byzantine.

Another equivalence in secure ML

Theorem (El-Mhamdi, Farhadkhani, Guerraoui, Guirguis, H & Rouault (NeurIPS 2021))

C-collaborative learning is equivalent to C-averaging.
Roughly, the guarantee on the norm of the true gradient at termination for collaborative learning can only be as good as the guarantee we can have when estimating the average of a set of vectors, assuming that some data source / vector providers are Byzantine.

Averaging is a particular case of learning

Averaging corresponds to losses $\mathcal{L}(\theta|\mathcal{D}_n) = ||\theta - \mathcal{D}_n||_2^2$.

Another equivalence in secure ML

Theorem (El-Mhamdi, Farhadkhani, Guerraoui, Guirguis, H & Rouault (NeurIPS 2021))

C-collaborative learning is equivalent to C-averaging.
Roughly, the guarantee on the norm of the true gradient at termination for collaborative learning can only be as good as the guarantee we can have when estimating the average of a set of vectors, assuming that some data source / vector providers are Byzantine.

Averaging is a particular case of learning

Averaging corresponds to losses $\mathcal{L}(\theta|\mathcal{D}_n) = ||\theta - \mathcal{D}_n||_2^2$.

From secure ML to secure vector aggregation

Secure vector averaging contains much of the difficulty of secure ML.

Averaging problem

Averaging problem

Given $x_1, \ldots, x_N \in \mathbb{R}^d$, securely compute y close to the true average $\bar{x}$.

Differential privacy

With the constraint $\mathbb{P}[y \in S | \overrightarrow{x}] \leq e^{\varepsilon} \mathbb{P}[y \in S | \overrightarrow{x}_{-n}] + \delta$.

Byzantine resilience

Where $\bar{x}$ is the average of $\overrightarrow{x}_H$, for $H \subset [N]$.

Heterogeneity is a (privacy) killer

Denote $\mathcal{B}(0, \Delta)$ the ball of $\mathbb{R}^d$ centered on 0, and of radius Δ.

Heterogeneity is a (privacy) killer

Denote $\mathcal{B}(0, \Delta)$ the ball of $\mathbb{R}^d$ centered on 0, and of radius Δ.

Theorem (Kattis & Nikolov (SoCG 2017))

For any (ε, δ)-differentially private estimator y, there exists $\overrightarrow{x} \in \mathcal{B}(0, \Delta)^N$ for which

$$\mathbb{E}||y - \bar{x}||_2^2 \geq \Omega\left(\frac{\sigma(\varepsilon, \delta) d \Delta^2}{N^2 (\log 2d)^4}\right), \tag{1}$$

where σ is a positive and non-increasing function.

Heterogeneity is a (privacy) killer

Denote $\mathcal{B}(0, \Delta)$ the ball of $\mathbb{R}^d$ centered on 0, and of radius Δ.

Theorem (Kattis & Nikolov (SoCG 2017))

For any (ε, δ)-differentially private estimator y, there exists $\overrightarrow{x} \in \mathcal{B}(0, \Delta)^N$ for which

$$\mathbb{E}||y - \bar{x}||_2^2 \geq \Omega\left(\frac{\sigma(\varepsilon, \delta) d \Delta^2}{N^2 (\log 2d)^4}\right), \tag{1}$$

where σ is a positive and non-increasing function.

Corollary

Assume $\Delta = \Theta(\sqrt{d})$. Then $\mathbb{E}||y - \bar{x}||_2^2 \geq \tilde{\Omega}(d^2/N^2)$.

Heterogeneity is a (privacy) killer

Denote $\mathcal{B}(0, \Delta)$ the ball of $\mathbb{R}^d$ centered on 0, and of radius Δ.

Theorem (Kattis & Nikolov (SoCG 2017))

For any (ε, δ)-differentially private estimator y, there exists $\overrightarrow{x} \in \mathcal{B}(0, \Delta)^N$ for which

$$\mathbb{E}||y - \bar{x}||_2^2 \geq \Omega\left(\frac{\sigma(\varepsilon, \delta) d \Delta^2}{N^2 (\log 2d)^4}\right), \tag{1}$$

where σ is a positive and non-increasing function.

Corollary

Assume $\Delta = \Theta(\sqrt{d})$. Then $\mathbb{E}||y - \bar{x}||_2^2 \geq \tilde{\Omega}(d^2/N^2)$.

Corollary (Informal)

If high-accuracy demands $d \gg 10^9$, then it cannot be obtained with differential privacy.

Heterogeneity is a (security) killer

Denote $\mathcal{B}(0, \Delta)$ the ball of $\mathbb{R}^d$ centered on 0, and of radius Δ.

Theorem (Adapted from EFGGHR (NeurIPS 2021))

For any (supposedly Byzantine-resilient) estimator y, there exists $\overrightarrow{x} \in \mathcal{B}(0, \Delta)^N$ and $H \subset [N]$ of cardinal $N - F$, such that

$$||y - \bar{x}_H||_2^2 \geq \frac{F^2}{(N-F)^2} \Delta^2. \tag{2}$$

Heterogeneity is a (security) killer

Denote $\mathcal{B}(0, \Delta)$ the ball of $\mathbb{R}^d$ centered on 0, and of radius Δ.

Theorem (Adapted from EFGGHR (NeurIPS 2021))

For any (supposedly Byzantine-resilient) estimator y, there exists $\overrightarrow{x} \in \mathcal{B}(0, \Delta)^N$ and $H \subset [N]$ of cardinal $N - F$, such that

$$||y - \bar{x}_H||_2^2 \geq \frac{F^2}{(N-F)^2}\Delta^2. \tag{2}$$

Corollary

Assume $\Delta = \Theta(\sqrt{d})$ and $F = \Theta(N)$. Then $||y - \bar{x}||_2^2 \geq \tilde{\Omega}(d^2)$.

Heterogeneity is a (security) killer

Denote $\mathcal{B}(0, \Delta)$ the ball of $\mathbb{R}^d$ centered on 0, and of radius Δ.

Theorem (Adapted from EFGGHR (NeurIPS 2021))

For any (supposedly Byzantine-resilient) estimator y, there exists $\overrightarrow{x} \in \mathcal{B}(0, \Delta)^N$ and $H \subset [N]$ of cardinal $N - F$, such that

$$||y - \bar{x}_H||_2^2 \geq \frac{F^2}{(N-F)^2} \Delta^2. \tag{2}$$

Corollary

Assume $\Delta = \Theta(\sqrt{d})$ and $F = \Theta(N)$. Then $||y - \bar{x}||_2^2 \geq \tilde{\Omega}(d^2)$.

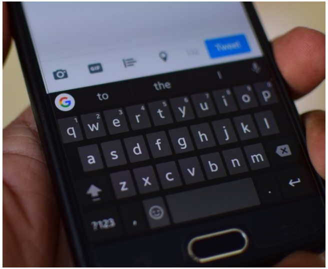
Corollary (Informal)

If high-accuracy demands $d \gg 10^9$, then it cannot be secured against data poisoning.

Section 2

The Alarming Practical Implications

Yet this is used for smart keyboard surveillance!

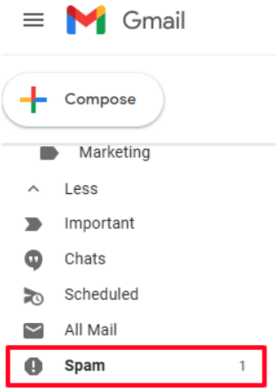


For this study, logs are collected from the English speaking population of Gboard users in the United States. Approximately 7.5 billion sentences are used for training, while the test and evaluation samples each contain 25,000 sentences. The average sentence length in the dataset is 4.1 words. A breakdown of the logs data by app type is provided in Table 1. Chat apps generate the majority of logged text.

Figure: Google has already been deploying high-dimensional language models on billions of phones, without users' informed consent and without an adequate understanding of privacy & security risks (extract from an ArXiV paper by Google authors).

Large-scale lucrative impactful applications are extremely heterogeneous

Google AdSense



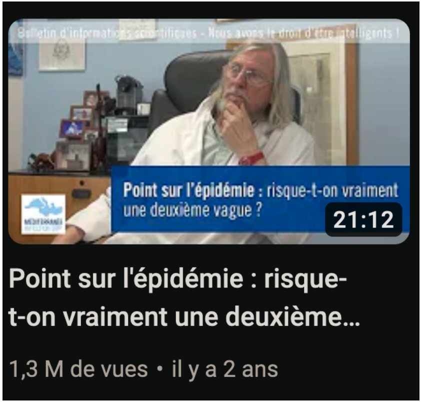


Figure: ML is now ubiquitous.

(Hijacked) recommendations are destroying democracies

Lê Nguyên Hoang 🌻
@le_science4all

Après la page Wikipedia sur les décès mystérieux d'hommes d'affaire Russes en 2022, voici la page sur les crises gouvernementales du Royaume-Uni en 2022 😅

Vous parlez d'une crise... oui... mais... laquelle ?
en.wikipedia.org/wiki/2022_Unit...

Translate Tweet

2022 United Kingdom government crisis

From Wikipedia, the free encyclopedia
(Redirected from 2022 UK government crisis)

2022 United Kingdom government crisis or **crises** may refer to:

- July 2022 United Kingdom government crisis, events culminating in the end of the Second Johnson ministry.
- September 2022 United Kingdom mini-budget, the Truss ministry's attempt to radically alter the British economy to solve the cost of living crisis.
- October 2022 United Kingdom government crisis, events in the aftermath of the September 2022 mini-budget, culminating in Truss' resignation.

*This disambiguation page lists articles associated with the title **2022 United Kingdom government crisis**.*
If an internal link led you here, you may wish to change the link to point directly to the intended article.

10:23 PM · Oct 20, 2022 · Twitter Web App

MARKETING

When Influence Goes Too Far: Social Media's Effect on the Capitol Riots

In this Insights@Questrom Q&A, Barbara Bickart, Senior Associate Dean of Graduate Programs and Associate Professor of Marketing, explains how influencers shape information and ideas on social media. Her insights reveal how persuasive tactics can lead to drastic events such as the Capitol riots.

Published 2 years ago on February 6, 2021
By Barbara Bickart

In this Insights@Questrom Q&A, Barbara Bickart, Senior Associate Dean of Graduate Programs and Associate Professor of Marketing, explains how influencers shape information and ideas on social media. Her insights reveal how persuasive tactics can lead to drastic events such as the riots that took place at the Capitol during President Joe Biden's transition to the presidency.

Question 1: How do influencers, whether that be brands, politicians, or celebrities, guide decision-making on social media?

Effective influencers connect to their audiences and build relationships, which enhances trust. Once trust is established, the influencer's behavior and recommendations are likely to impact the audience's decisions. Establishing trust is not easy. There are several approaches to building trust. First, we tend to like people who are like us and are more persuaded by what they say and do. Therefore, influencers can emphasize their similarity to their audience, disclosing details of their lives or beliefs and opinions to which the audience can relate.

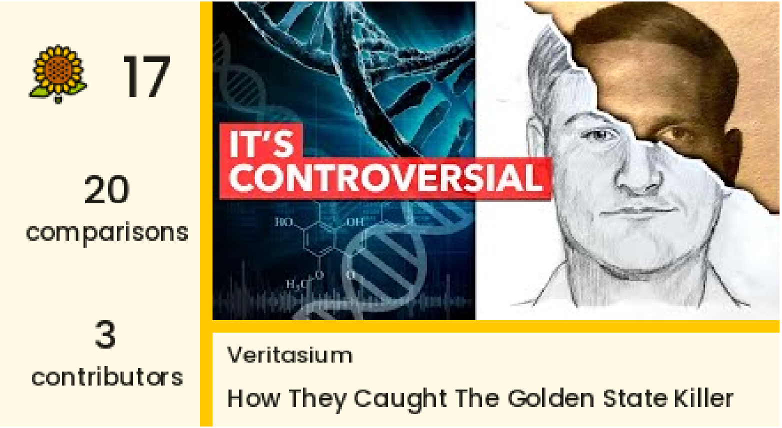
SANTÉ | COMPLOTISME | NEWS
Publié le 11 avril 2022 15:05. Modifié le 09 mai 2022 14:53.

Enlèvement de Christoph Berger, président de la Commission fédérale pour les vaccinations, et mort de son ravisseur antivax

par Annick Chevillot

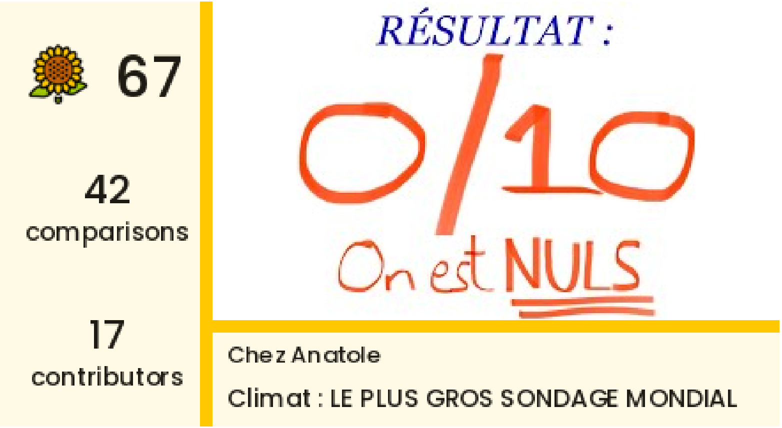
Les faits se sont déroulés en deux temps: l'enlèvement d'un homme le 31 mars, puis la mort de son ravisseur présumé le 6 avril au soir à Wallisellen dans le canton de Zurich. Un fait divers qui va rapidement prendre une dimension nationale et politique: le ravisseur présumé est un Allemand actif dans les milieux complotistes et antivax; et le kidnappé est un médecin mandaté par la Confédération à la tête de la Commission fédérale pour les vaccinations (CFV). Christoph Berger a publié un communiqué personnel le 10 avril dans lequel il explique brièvement ce qu'il lui est arrivé.

And yet our (privacy) math is still flawed!



Personal data (= data associated to a person) is different from *sensitive information* (= information that a person would not want to see spread). Especially for language/DNA data.

And yet our (security) math is still flawed!



Massive amounts of misinformation & hate is shared by (the majority of) authentic persons.

Google's scientific disinformation

RESEARCH-ARTICLE OPEN ACCESS

Practical Secure Aggregation for Privacy-Preserving Machine Learning

Authors: Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, Karn Seth Authors Info & Claims

CCS '17: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security • October 2017 • Pages 1175–1191 • https://doi.org/10.1145/3133956.3133982

Online: 30 October 2017 Publication History

ABSTRACT

We design a novel, communication-efficient, failure-robust protocol for secure aggregation of high-dimensional data. Our protocol allows a server to compute the sum of large, user-held data vectors from mobile devices in a secure manner (i.e. without learning each user's individual contribution), and can be used, for example, in a federated learning setting, to aggregate user-provided model updates for a deep neural network. We prove the security of our protocol in the honest-but-curious and active adversary settings, and show that security is maintained even if an arbitrarily chosen subset of users drop out at any time. We evaluate the efficiency of our protocol and show, by complexity analysis and a concrete implementation, that its runtime and communication overhead remain low even on large data sets and client pools. For 16-bit input values, our protocol offers \$1.73 x communication expansion for 2^{10} users and 2^{20}-dimensional vectors, and 1.98 x expansion for 2^{14} users and 2^{24}-dimensional vectors over sending data in the clear.

EXTENDED-ABSTRACT

Learning Product Rankings Robust to Fake Users

Authors: Negin Golrezaei, Vahideh Manshadi, Jon Schneider, Shreyas Sekar Authors Info & Claims

EC '21: Proceedings of the 22nd ACM Conference on Economics and Computation • July 2021 • Pages 560–561 • https://doi.org/10.1145/3465456.3467580

Online: 18 July 2021 Publication History

ABSTRACT

In many online platforms, customers' decisions are substantially influenced by product rankings as most customers only examine a few top-ranked products. Concurrently, such platforms also use the same data corresponding to customers' actions to learn how these products must be ranked or ordered. These interactions in the underlying learning process, however, may incentivize sellers to artificially inflate their position by employing fake users, as exemplified by the emergence of click farms. Motivated by such fraudulent behavior, we study the ranking problem of a platform that faces a mixture of real and fake users who are indistinguishable from one another. We first show that existing learning algorithms---that are optimal in the absence of fake users---may converge to highly sub-optimal rankings under manipulation by fake users. To overcome this deficiency, we develop efficient learning algorithms under two informational environments: in the first setting, the platform is aware of the number of fake users, and in the second setting, it is agnostic to the number of fake users. For both these environments, we prove that our algorithms converge to the optimal ranking, while being robust to the aforementioned fraudulent behavior; we also present worst-case performance guarantees for our methods, and show that they significantly outperform existing algorithms. At a high level, our work employs several novel approaches to guarantee robustness such as: (i) constructing product-ordering graphs that encode the pairwise relationships between products inferred from the customers' actions; and (ii) implementing multiple levels of learning with a judicious amount of bi-directional cross-learning levels. Overall, our results indicate that online platforms can effectively combat fraudulent users without incurring large costs by designing new learning algorithms that guarantee efficient convergence even when the platform is completely oblivious to the number and identity of the fake users.

Should we trust the central server?

Planting Undetectable Backdoors in Machine Learning Models

Shafi Goldwasser, UC Berkeley
Michael P. Kim, UC Berkeley
Vinod Vaikuntanathan, MIT
Or Zamir, IAS

Abstract

Given the computational cost and technical expertise required to train machine learning models, users may delegate the task of learning to a service provider. Delegation of learning has clear benefits, and at the same time raises *serious concerns of trust*. This work studies possible abuses of power by untrusted learners.

We show how a malicious learner can plant an *undetectable backdoor* into a classifier. On the surface, such a backdoored classifier behaves normally, but in reality, the learner maintains a mechanism for changing the classification of any input, with only a slight perturbation. Importantly, without the appropriate "backdoor key," the mechanism is hidden and cannot be detected by any computationally-bounded observer. We demonstrate two frameworks for planting undetectable backdoors, with incomparable guarantees.

- First, we show how to plant a backdoor in *any model*, using digital signature schemes. The construction guarantees that given query access to the original model and the backdoored version, it is computationally infeasible to find even a single input where they differ. This property implies that the backdoored model has generalization error comparable with the original model. Moreover, even if the distinguisher can request backdoored inputs of its choice, they cannot backdoor a new input—a property we call *non-replicability*.
- Second, we demonstrate how to insert undetectable backdoors in models trained using the Random Fourier Features (RFF) learning paradigm (Rahimi, Recht; NeurIPS 2007). In this construction, undetectability holds against powerful *white-box distinguishers*: given a complete description of the network and the training data, no efficient distinguisher can guess whether the model is "clean" or contains a backdoor. The backdooring algorithm executes the RFF algorithm faithfully on the given training data, tampering only with its random coins. We prove this strong guarantee under the hardness of the Continuous Learning With Errors problem (Bruna, Regev, Song, Tang; STOC 2021). We show a similar white-box undetectable backdoor for random ReLU networks based on the hardness of Sparse PCA (Berthet, Rigollet; COLT 2013).

Check your working hypotheses (and your peers')

The most widespread dangerously unrealistic assumption in ML

"Assume *iid* data..."

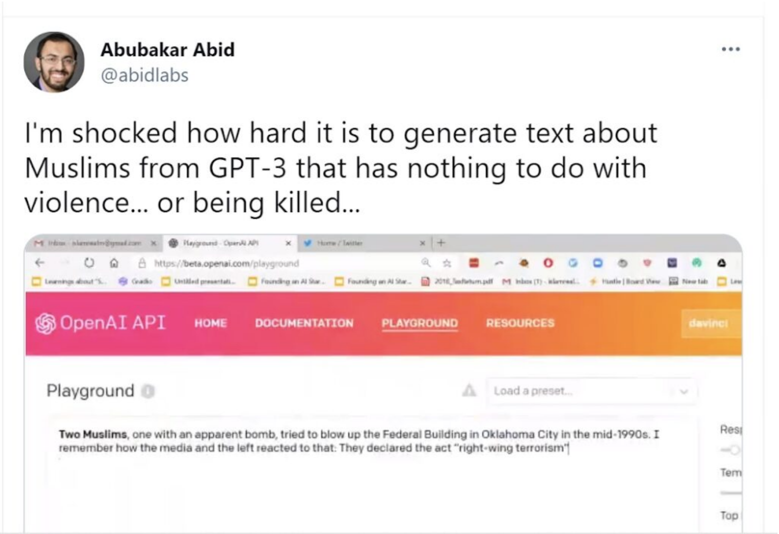
The most widespread politically biased assumption in ML

"We minimize the data-fitting loss..."

Section 3

Towards collaborative and secure governance (Tournesol)

Learning as a voting algorithm



The “ground truth” myth

The most impactful ML applications (language, recommendations, ad targeting...) have no ground truth.

The most impactful ML applications (language, recommendations, ad targeting...) have no ground truth.

Instead, we should (securely) search for (scientific and moral) **consensus** and **compromises**.

Sparse voting is extremely vulnerable

ML's extreme sparsity

If $|\mathcal{D}_n| \ll d$, then each user provides (extremely) *sparse* data.

Sparse voting is extremely vulnerable

ML's extreme sparsity

If $|\mathcal{D}_n| \ll d$, then each user provides (extremely) *sparse* data.

Byzantine vulnerability

Alternatives that no one scored are extremely vulnerable.

Sparse voting is extremely vulnerable

ML's extreme sparsity

If $|\mathcal{D}_n| \ll d$, then each user provides (extremely) *sparse* data.

Byzantine vulnerability

Alternatives that no one scored are extremely vulnerable.

Corollary

Under extreme sparsity, median-based recommendation algorithms are extremely dangerous!

Byzantine resilience revisited

Definition

$\textsc{Alg}$ is W-Byzantine resilient if, for any voting rights $w, w' \in \mathbb{R}_+^N$ and any inputs $x \in X^N$,

$$|\textsc{Alg}(w, x) - \textsc{Alg}(w', x)| \leq \frac{||w - w'||_1}{W}. \tag{3}$$

Byzantine resilience revisited

Definition

$\textsc{Alg}$ is W-Byzantine resilient if, for any voting rights $w, w' \in \mathbb{R}_+^N$ and any inputs $x \in X^N$,

$$\left|\textsc{Alg}(w, x) - \textsc{Alg}(w', x)\right| \leq \frac{||w - w'||_1}{W}. \tag{3}$$

Definition (W-quadratically regularized median)

$$\textsc{QrMed}_W(w, x) \triangleq \arg\min_{m \in \mathbb{R}} \left\{ \frac{1}{2} W m^2 + \sum_{n \in [N]} w_n |x_n - m| \right\}. \tag{4}$$

Byzantine resilience revisited

Definition

$\textsc{Alg}$ is W-Byzantine resilient if, for any voting rights $w, w' \in \mathbb{R}_+^N$ and any inputs $x \in X^N$,

$$\left|\textsc{Alg}(w, x) - \textsc{Alg}(w', x)\right| \leq \frac{||w - w'||_1}{W}. \tag{3}$$

Definition (W-quadratically regularized median)

$$\textsc{QrMed}_W(w, x) \triangleq \arg\min_{m \in \mathbb{R}} \left\{ \frac{1}{2} W m^2 + \sum_{n \in [N]} w_n |x_n - m| \right\}. \tag{4}$$

Theorem

For all $W > 0$, $\textsc{QrMed}_W$ is W-Byzantine resilient.

The French reviewer problem

ML's extreme sparsity

If $|\mathcal{D}_n| \ll d$, then each user provides (extremely) *sparse* data.

The French reviewer problem

ML's extreme sparsity

If $|\mathcal{D}_n| \ll d$, then each user provides (extremely) *sparse* data.

The French reviewer problem

Some alternatives may be scored by systematically unsatisfied reviewers.

The French reviewer problem

ML's extreme sparsity

If $|\mathcal{D}_n| \ll d$, then each user provides (extremely) *sparse* data.

The French reviewer problem

Some alternatives may be scored by systematically unsatisfied reviewers.

The Marseillais reviewer problem

Top alternatives may be those scored by users with extreme judgments.

The French reviewer problem

ML's extreme sparsity

If $|\mathcal{D}_n| \ll d$, then each user provides (extremely) *sparse* data.

The French reviewer problem

Some alternatives may be scored by systematically unsatisfied reviewers.

The Marseillais reviewer problem

Top alternatives may be those scored by users with extreme judgments.

Theorem (Von Neumann - Morgenstern (1944))

VNM utility functions are only defined up to a positive affine transformation.

Robust sparse voting

Definition (Sparse unanimity, informal)

Assuming that

1. all users actually unanimously agree (up to an affine transformation),
2. all alternatives are scored by sufficiently many users, and
3. all pairs of users have scored sufficiently many alternatives in common,

the vote must output the unanimous preference (up to an affine transformation).

Robust sparse voting

Definition (Sparse unanimity, informal)

Assuming that

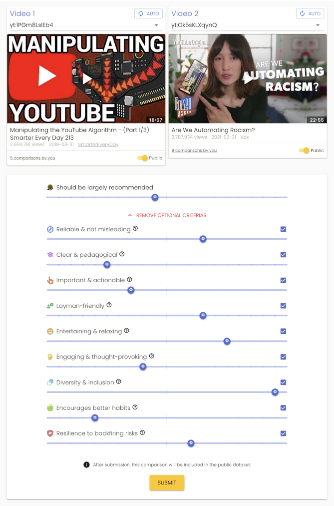
1. all users actually unanimously agree (up to an affine transformation),
2. all alternatives are scored by sufficiently many users, and
3. all pairs of users have scored sufficiently many alternatives in common,

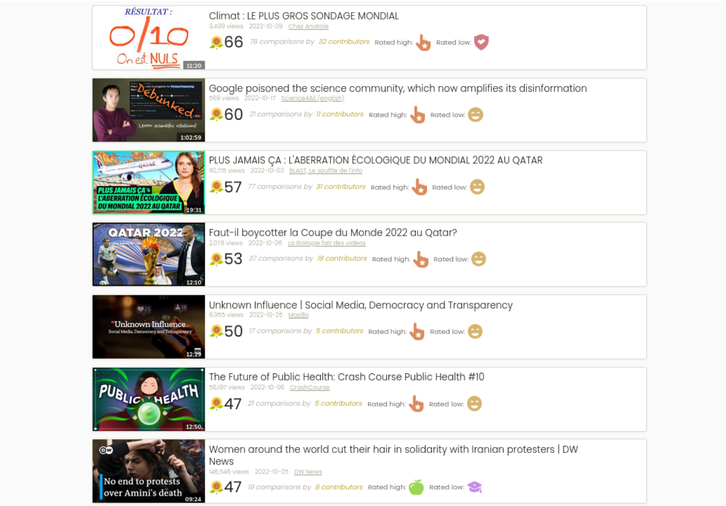
the vote must output the unanimous preference (up to an affine transformation).

Theorem (Allouah, Guerraoui, Hoang & Villemaud (2022))

For all $W > 0$, there is an algorithm (called W-Mehestan) that guarantees both sparse unanimity and W-Byzantine resilience.

Tournesol


Video 1
AUTO
yt:lPGm8LslEb4
MANIPULATING
YOUTUBE
18:57
Manipulating the YouTube Algorithm - (Part 1/3) Smarter Every Day 213
2,666,781 views 2019-03-31 SmarterEveryDay
5 comparisons by you
Public
Video 2
AUTO
yt:Ok5sKLXqynQ
ARE WE AUTOMATING RACISM?
22:55
Are We Automating Racism?
3,787,934 views 2021-03-31 Vox
9 comparisons by you
Public
Should be largely recommended
REMOVE OPTIONAL CRITERIAS
Reliable & not misleading
Clear & pedagogical
Important & actionable
Layman-friendly
Entertaining & relaxing
Engaging & thought-provoking
Diversity & inclusion
Encourages better habits
Resilience to backfiring risks
After submission, this comparison will be included in the public dataset.
SUBMIT


RÉSULTAT : 0/10 On est NULS
11:20
Climat : LE PLUS GROS SONDAGE MONDIAL
3,499 views 2022-10-09 Chez Anatole
66 78 comparisons by 32 contributors Rated high: Rated low:
1:02:59
Google poisoned the science community, which now amplifies its disinformation
569 views 2022-10-17 Science4All (english)
60 21 comparisons by 11 contributors Rated high: Rated low:
PLUS JAMAIS ÇA : L'ABERRATION ÉCOLOGIQUE DU MONDIAL 2022 AU QATAR
19:31
PLUS JAMAIS ÇA : L'ABERRATION ÉCOLOGIQUE DU MONDIAL 2022 AU QATAR
110,715 views 2022-10-03 BLAST, Le souffle de l'info
57 77 comparisons by 31 contributors Rated high: Rated low:
QATAR 2022
12:10
Faut-il boycotter la Coupe du Monde 2022 au Qatar?
2,078 views 2022-10-06 La Biologie fait des vidéos
53 37 comparisons by 18 contributors Rated high: Rated low:
"Unknown Influence
Social Media, Democracy and Transparency
12:29
Unknown Influence | Social Media, Democracy and Transparency
8,955 views 2022-10-25 Mozilla
50 17 comparisons by 5 contributors Rated high: Rated low:
PUBLIC HEALTH
12:50
The Future of Public Health: Crash Course Public Health #10
56,197 views 2022-10-06 CrashCourse
47 21 comparisons by 5 contributors Rated high: Rated low:
DW
No end to protests over Amini's death
09:24
Women around the world cut their hair in solidarity with Iranian protesters | DW News
146,546 views 2022-10-05 DW News
47 19 comparisons by 8 contributors Rated high: Rated low:

Secure collaborative governance

Tournesol: A quest for a large, secure and trustworthy database of reliable human judgments

Lê-Nguyên Hoang[1,2], Louis Faucon[2], Aidan Jungo[2], Sergei Volodin[2], Dalia Papuc[1,2], Orfeas Liossatos[1,2], Ben Crulis[3], Mariame Tighanimine[2,4], Isabela Constantin[2], Anastasiia Kucherenko[1,2], Alexandre Maurer[2,5], Felix Grimberg[1,2], Vlad Nitu[2,6], Chris Vossen[2], Sébastien Rouault[1,2], and El-Mahdi El-Mhamdi[2,7]

[1]IC, EPFL, Switzerland
[2]Tournesol Association, Switzerland
[3]University of Tours, France
[4]LISE, CNAM-CNRS, France
[5]UM6P, Benguerir, Morocco
[6]CNRS, INSA Lyon, France
[7]École Polytechnique, France

Abstract

Today's large-scale algorithms have become immensely influential, as they recommend and moderate the content that billions of humans are exposed to on a daily basis. These algorithms are the de-facto regulators of the information diet of billions of humans, from shaping opinions on public health information to organizing groups for social movements. This creates serious concerns, but also great opportunities to promote quality information [Hoa20, HFE21]. Addressing the concerns and seizing the opportunities is a challenging, enormous and fabulous endeavor [HE19], as intuitively appealing ideas often come with unforeseen unwanted *side effects* [EMH21], and as it requires us to think about what we truly and deeply prefer [Soa15].

To make progress, it is critical to understand how today's large-scale algorithms are built, and to determine what interventions will be most effective. Given that these algorithms rely heavily on *machine learning*, we make the following key observation: *any algorithm trained on uncontrolled data must not be trusted*. Indeed, a malicious entity could take control over the data, poison it with dangerously misleading or manipulative fabricated inputs, and thereby make the trained algorithm extremely unsafe. We thus argue that the first step towards safe and ethical large-scale algorithms must be the collection of a large, secure and trustworthy dataset of reliable human judgments.

To achieve this, we introduce *Tournesol*, an open source platform available at `https://tournesol.app`. Tournesol aims to collect a large database of human judgments on what algorithms ought to widely recommend (and what algorithms ought to stop widely recommending). In this paper, we outline the structure of the Tournesol database, the key features of the Tournesol platform and the main hurdles that must be overcome to make it a successful project. Most importantly, we argue that, if successful, Tournesol may then serve as the essential foundation for any safe and ethical large-scale algorithm.

Tournesol: Permissionless Collaborative Algorithmic Governance with Security Guarantees

Anonymous Author(s)
Submission Id: 833

ABSTRACT

Recommendation algorithms play an increasingly central role in our societies. However, thus far, these algorithms are mostly designed and parameterized in a unilateral manner by private groups or governmental authorities. In this paper, we present an end-to-end permissionless collaborative algorithmic governance method with security guarantees. Our proposed method is deployed as part of an open-source content recommendation platform tournesol.app, whose recommender is collaboratively parameterized by a community of (non-technical) contributors. This algorithmic governance is achieved through three main steps. First, the platform contains a mechanism to assign voting rights to the contributors. Second, the platform uses a comparison-based model to evaluate individual preferences of contributors. Third, the platform aggregates the judgements of all contributors into collective scores for content recommendations. We stress that the first and third steps are vulnerable to attacks from malicious contributors. To guarantee the resilience against fake accounts, the first step combines email authentication, a vouching mechanism, a novel variant of the reputation-based EigenTrust algorithm and an adaptive voting rights assignment for alternatives that are scored by too many untrusted accounts. To provide resilience against malicious authenticated contributors, we adapt MEHESTAN, an algorithm previously proposed for *robust sparse voting*. We believe that these algorithms provide an appealing foundation for a collaborative, effective, scalable, fair, contributor-friendly, interpretable and secure governance. We conclude by highlighting a few key challenges to make our solution applicable to larger-scale settings.

KEYWORDS

Recommendation, vote, security, Sybil, Byzantine, governance.

ACM Reference Format:
Anonymous Author(s). 2023. Tournesol: Permissionless Collaborative Algorithmic Governance with Security Guarantees. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 10 pages.

1 INTRODUCTION

In today's digital information war [15, 55], large-scale algorithms play a *de facto* major political role [28, 30, 33, 62]. Whenever a search engine is given queries like "climate hoax", "vaccination", "vote steal", "Ukraine invasion" or "Xinjiang camps", it must return a ranking, which will inevitably prioritize some views over others. Similarly, chatbots can be asked to discuss these topics, for which "neutrality" may endanger lives and may thus be unsatisfactory.

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

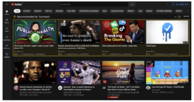


Figure 1: Our browser extension provides Tournesol's recommendations directly on the users' YouTube home page.

Perhaps most importantly, every day, social medias' recommendation algorithms are making billions of content recommendations. Even if a mere 0.1% of the recommendations discuss such important and hotly debated topics, this still represents millions of daily decisions with potential national security implications. In fact, given the central role played by these algorithms in the information market, *not recommending* some content can amount to *silencing their discussion topics*, which can itself be regarded as a disputable political stance, especially when urgent action is required [32, 38, 47].

Unfortunately, building information systems that appropriately prioritize information (and with its societal implications) is arguably largely under-researched, and currently lacks satisfactory solutions. As a result, unsurprisingly, today's algorithms are mostly designed, managed and governed in a relatively unilateral (opaque) manner.

In this paper, we present the algorithms of *Tournesol* [31], an online platform that proposes an end-to-end solution to the above problem. Building upon the more democratic principles of social choice theory [8, 13], the Tournesol recommendation algorithm is constructed as an aggregation of contributors' content recommendation preferences. As such, Tournesol is a proposal for a *collaborative* algorithmic governance mechanism. Making such a system effective, scalable, fair, contributor-friendly, interpretable and secure raises significant challenges that are the focus of this paper.

Tournesol is already deployed and available at tournesol.app, and the open-source code can be found on GitHub[1]. The recommendations are based on over 50,000 judgments made by over 10,000 contributors on over 10,000 videos. They are made to thousands of users on the website, as well as directly on YouTube.com through the Firefox and Chrome browser extensions, with more than 1,000 new contributions per week in September and October 2022. All the required processing is currently performed on a modest server with 4 CPU cores and no GPU.

[1]See https://github.com/tournesol-app/tournesol. Note that the hyperparameters we provide in the paper refer to commit 8f3c9881f5d99a97a879b55acc5140b7e2c6a4a5.

Section 4

Conclusion

Our scientific integrity is at stake. So is the world's security.

Figure: Google poisoned the science community, which now amplifies its disinformation