Secure statistics for collaborative algorithmic governance

Lê Nguyên Hoang,
Calicarpa, Tournesol & Science4All,
STATLEARN, April 2023
Section 1

Adversarial statistics
Statistics is ubiquitous online

What happens online in 60 seconds? (2012 - 2014)

- 120 hours video uploaded
- 72 hours video uploaded
- 25+ hours
- 20 billion messages
- 347 posts
- 31 billion messages
- 1106 blog posts
- 50 billion messages
- 1380 blog posts
- 342,000 tweets
- 278,000 tweets
- 1.4 million minutes video calls
- 370,000 voice calls
- 1.4 million minutes voice calls
- 216,000 photos
- 41,000 photos uploaded
- 694,445 searches
- 4 million searches
- 3.3 million posts
- 2.5 million posts
- 204 million emails sent
- 168 million emails sent
- 79,361 posts

Picture by Centre for Learning and Teaching
3rd November 2014
The tale of Microsoft’s two sisters (Tay vs Xiaoice)
You are voting all the times
Non-users are stakeholders in online votes.

MYANMAR: FACEBOOK'S SYSTEMS PROMOTED VIOLENCE AGAINST ROHINGYA; META OWES REPARATIONS

ACT NOW

September 29, 2022

Facebook owner Meta’s dangerous algorithms and reckless pursuit of profit substantially contributed to the atrocities perpetrated by the Myanmar military against the Rohingya people in 2017. Amnesty International said in a new report published today.
There is a huge steal-the-online-vote industry

“TEAM JORGE”: IN THE HEART OF A GLOBAL DISINFORMATION MACHINE

In Part 2 of the “Story Killers” project, which continues the work of assassinated Indian journalist Gauri Lankesh on disinformation, the Forbidden Stories consortium investigated an ultra-secret Israeli company involved in manipulating elections and hacking African politicians. We took an unprecedented dive into a world where troll armies, cyber espionage and influencers are intertwined.

Facebook Removed More than 15 Billion Fake Accounts in Two Years, Five Times more than its Active User Base

As the world’s largest social networking platform, Facebook has witnessed a surge in the number of users in the past few years. Hundreds of millions of people have joined its social media space to communicate, keep in touch with the latest trends or promote business, especially after the pandemic hit. Although the COVID-19 restrictions have loosened in most countries, Facebook’s active user base continues growing, but so does the number of fake accounts.

According to data presented by Stock Apps, the social media giant removed over 15 billion fake accounts in the last two years, five times more than its active user base.

3 Billion Fake Accounts Removed in the First Half of 2021, 20x More than the Number of New Active Users

Scammers use fake Facebook accounts to connect with users, get their personal information and steal identities. Most of them will reach out to anyone who’s accepted their friend request to try and scam them out of money.

Many fake accounts are also driven by spammers who are constantly trying to invade Facebook’s systems. Although the social media giant invested in enhanced technology to detect automated and coordinated spam, the problem is still getting worse.

According to the company’s official data, in 2019, Facebook removed 6.5 billion fake accounts, the highest number to date.
Defend data poisoning? Heterogeneity and large models are security killers

An Equivalence Between Data Poisoning and Byzantine Gradient Attacks

Sadegh Farhadi Khani, Rachid Guerraoui, Lê Nguyên Hoang, Oscar Villemant

Abstract

To study the resilience of distributed learning, the “Byzantine” literature considers a strong threat model where workers can report arbitrary gradients to the parameter server. Whereas this model helped obtain several fundamental results, it has sometimes been considered unrealistic, when the workers are mostly trustworthy machines. In this paper, we show a surprising equivalence between this model and data poisoning, a threat considered much more realistic. More specifically, we prove that every gradient attack can be reduced to data poisoning, in any personalized federated learning system with PAC guarantees (which we show are both desirable and realistic). This equivalence makes it possible to obtain new impossibility results on the resilience of any “robust” learning algorithm to data poisoning in highly heterogeneous applications, as corollaries of existing impossibility theorems on Byzantine machine learning. Moreover, using our equivalence, we derive a practical attack that we show (theoretically and empirically) can be very effective against classical personalized federated learning models.
On the Strategyproofness of the Geometric Median

El-Mahdi El-Mhamdi, Sadegh Farhadkhani, Rachid Guerraoui, Lê-Nguyên Hoàng

The geometric median of a tuple of vectors is the vector that minimizes the sum of Euclidean distances to the vectors of the tuple. Classically called the Fermat-Weber problem and applied to facility location, it has become a major component of the robust learning toolbox. It is typically used to aggregate the (processed) inputs of different data providers, whose motivations may diverge, especially in applications like content moderation. Interestingly, as a voting system, the geometric median has well-known desirable properties: it is a provably good average approximation, it is robust to a minority of malicious voters, and it satisfies the "one voter, one unit force" fairness principle. However, what was not known is the extent to which the geometric median is strategyproof. Namely, can a strategic voter significantly gain by misreporting their preferred vector? We prove in this paper that, perhaps surprisingly, the geometric median is not even $\alpha$-strategyproof, where $\alpha$ bounds what a voter can gain by deviating from truthfulness. But we also prove that, in the limit of a large number of voters with i.i.d. preferred vectors, the geometric median is asymptotically $\alpha$-strategyproof. We show how to compute this bound $\alpha$. We then generalize our results to voters who care more about some dimensions. Roughly, we show that, if some dimensions are more polarized and regarded as more important, then the geometric median becomes less strategyproof. Interestingly, we also show how the skewed geometric medians can improve strategyproofness. Nevertheless, if voters care differently about different dimensions, we prove that no skewed geometric median can achieve strategyproofness for all. Overall, our results constitute a coherent set of insights into the extent to which the geometric median is suitable to aggregate high-dimensional disagreements.
Section 2

Tournesol
A deployed functional vote...

Collaborative Content Recommendations

Tournesol is a transparent participatory research project about the ethics of algorithms and recommendation systems.

Help us advance research by giving your opinion on the videos you have watched in order to identify public interest contents that should be largely recommended.

CREATE ACCOUNT  START
Activated accounts
17,540  
+ 347
Comparisons
99,015  
+ 8,984
Rated videos
20,351  
+ 1,260

Research
“We seek to support research on the ethics of algorithms by providing a large and reliable database of human judgments.”

Our data are open
We hope that other projects can benefit from the efforts of the Tournebelle community. To this end we are making available a database made up of all public contributions that anyone can use.

These data are published under the terms of the Open Data Commons Attribution License (ODC 1.0).

Download the database

Our algorithms are Free/Libre
In a perspective of transparency and knowledge sharing, the algorithms and all source code we created are Free Software.

Access the code on GitHub

Visualize the data
You can quickly explore our public database with our Tournebelle Data Visualization application made with Streamlit.

pearson correlation coefficient matrix of comparison criteria scores (2022.07.19)
Tournesol’s comparison interface

Facebook Whistleblower Frances Haugen: The 60 Minutes Interview

Should be largely recommended

Secure stats
Tournesol's recommendations

- Amazing Invention: This Drone Will Change Everything
- Timeshares: Last Week Tonight with John Oliver (HBO)
- Price Controls in the Pharmaceutical Industry
- Let's Solve Ethics Collaboratively!
- Cozy night with my wife in a wooden house in the cold. Off grid cabin
- Pawn Stars: TOP 4 OLDEST ITEMS EVER
- DUEL DE BLAGUES NULLES édition ChatGPT (les robots sont plus drôles...)
- $1 vs $500,000 Plane Ticket!
Tournesol: Permissionless Collaborative Algorithmic Governance with Security Guarantees

Romain Beylerian, Bérangère Colbois, Louis Faucon, Lê Nguyên Hoang, Aidan Jungo, Alain Le Noach, Adrien Matissart

Recommendation algorithms play an increasingly central role in our societies. However, thus far, these algorithms are mostly designed and parameterized unilaterally 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 https://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 the 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 key challenges to make our solution applicable to larger-scale settings.
Voting fairness: one person, one unit force

Recommend more often
Voting fairness: one person, one unit force
Voting fairness: one person, one unit force
Voting fairness: one person, one unit force
Voting fairness: one person, one unit force
Under extreme sparsity, the median is not robust enough!

Most web items are never reviewed!
The median of a single (malicious) voter’s score is the voter’s score.
Limiting each voter’s influence: \( W \)-Byzantine resilience

2.2 Byzantine resilience

Our second desirable property under study is what we call **Byzantine resilience**. To formalize it, for any subset \( F \subset [N] \) of (Byzantine) voters, denote \( \vec{w}^F \) the tuple of voting rights defined by \( w^F_n = 0 \) for \( n \notin F \), and \( w^F_f = w_f \) for \( f \in F \). In other words, \( \vec{w}^F \) cancels the voting rights of non-Byzantine voters. Conversely, denote \( \vec{w}^{-F} = \vec{w}^{[N] - F} \). Clearly, we have \( \vec{w} = \vec{w}^F + \vec{w}^{-F} \). Byzantine resilience then demands that canceling (or activating) the Byzantine voters’ voting rights will only have a limited effect on the vote outcome, whose scale is bounded by the Byzantine’s total voting rights. Evidently, since we assume that \textsc{Vote} cannot distinguish Byzantine voters from genuine voters, our definition of Byzantine resilience must treat any subset \( F \subset [N] \) identically.

**Definition 2.** **Vote guarantees** \( W \)-Byzantine resilience if, for any inputs \((\vec{w}, \vec{\theta})\), a subgroup \( F \subset [N] \) can affect each output of the vote by at most \( \frac{\|\vec{w}^F\|_1}{W} \), i.e.

\[
\forall \vec{w}, \vec{\theta}, \forall F \subset [N], \forall a \in [A], \quad \left| \text{VOTE}_a(\vec{w}^{-F}, \vec{\theta}) - \text{VOTE}_a(\vec{w}, \vec{\theta}) \right| \leq \frac{\|\vec{w}^F\|_1}{W}. \tag{1}
\]

We say that **Vote** is **Byzantine resilient**, if there exists \( W > 0 \) such that **Vote** is \( W \)-Byzantine resilient.

The variable \( W \) can be interpreted as a resilience measure. Intuitively, it protects the vote against Byzantine voters whose cumulative voting right is bounded by \( W \). More precisely, the Byzantine voters must have at least \( W \) voting rights to move an alternative’s score by one unit. Put differently, this amounts to \( 1/W \)-Lipschitz continuity in voters’ voting rights (with respect to \( \ell_1 \) norm).
The quadratically regularized median (QrMed)
Section 3

Noise and biases
The French reviewers problems

Note: Some of my best friends are Parisian and Marseillais.
The French reviewers problems

Note: Some of my best friends are Parisian and Marseillais.

The Parisien reviewer problem
Some content may be mostly scored by complain-addict reviewers.
The French reviewers problems

Note: Some of my best friends are Parisian and Marseillais.

The Parisien reviewer problem
Some content may be mostly scored by complain-addict reviewers.

The Marseillais reviewer problem
Some content may be mostly scored by exaggeration-addict reviewers.
Definition (Sparse unanimity, simplified)

A voting algorithm is sparsely unanimous if, assuming

a) each pair of voters scores two common alternatives,

b) each alternative is scored by sufficiently many voters and

c) all voters have the same VNM preferences,

the vote returns the unanimous VNM preferences.
Definition (Sparse unanimity, simplified)
A voting algorithm is sparsely unanimous if, assuming
  a) each pair of voters scores two common alternatives,
  b) each alternative is scored by sufficiently many voters and
  c) all voters have the same VNM preferences,
the vote returns the unanimous VNM preferences.

Theorem (AGHV'21)
For any $W$, there exists a voting algorithm, called Mehestan and deployed on Tournesol, which guarantees both $W$-Byzantine resilience and sparse unanimity.
Accounting for varying data uncertainties
Unsolved challenges

- Active learning
- Provably approximately correct heuristics
- Volition learning (include priors on psychological behaviors)
- Epistocratical (robust) voting
- Bayesian (robust) voting
Taiwan’s pol.is experience

73 contributors

62 comparisons

La Fabrique Sociale
Taiwan, la démocratie du futur?
Section 4

Conclusion
Statistical hypotheses must urgently be revised for online applications

The most widespread dangerously unrealistic assumption for web-applied statistics

“Assume iid data...”
Statistical hypotheses must urgently be revised for online applications

The most widespread dangerously unrealistic assumption for web-applied statistics
“Assume iid data…”

The most widespread politically biased assumption for web-applied statistics
“We fit the data…”
Tournesol’s data are publicly available!

Activated accounts
17,540
+ 347
Comparisons
99,015
+ 8,984
Rated videos
20,351
+ 1,260

Research

“We seek to support research on the ethics of algorithms by providing a large and reliable database of human judgments.”

Our data are open

We hope that other projects can benefit from the efforts of the Tournesol community. To this end we are making available a database made up of all public contributions that anyone can use.

These data are published under the terms of the Open Data Commons Attribution License (ODC: BY 1.0).

Download the database

Our algorithms are Free/Libre

In a perspective of transparency and knowledge sharing, the algorithms and all source code we created are Free Software.

Access the code on GitHub

Visualize the data

You can quickly explore our public database with our Tournesol Data Visualization application made with Streamlit.
Snap CISO: I rate software supply chain risk 9.9 out of 10

"Understanding your inventory is absolutely No. 1" he tells The Reg

On a scale of 1 to 10, 10 being the highest risk, Snap Chief Information Security Officer Jim Higgins rates software supply chain risk "about 9.9."

Snap says it serves 375 million daily active users, all of which has to be kept secure and reliable. Not only is the supply chain a high risk, it's a tough security problem to fix because a single product can have tens of thousands of software dependencies.