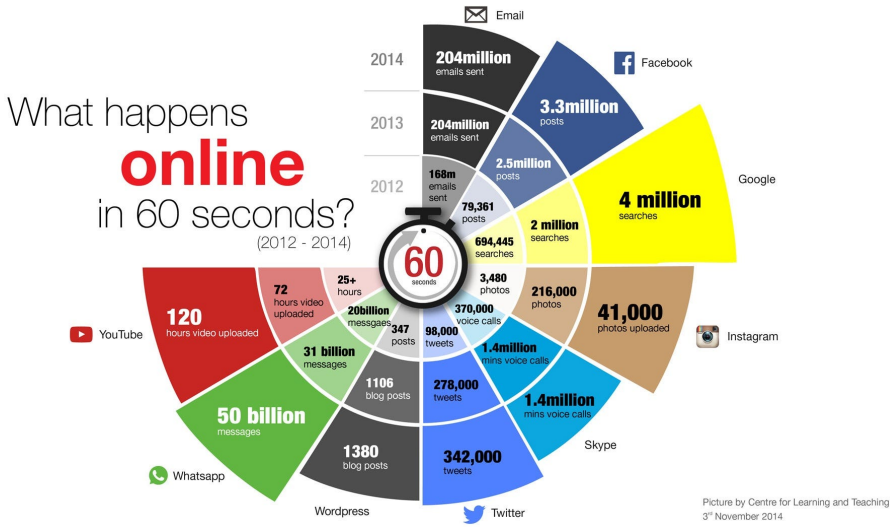
Secure statistics for collaborative algorithmic governance

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

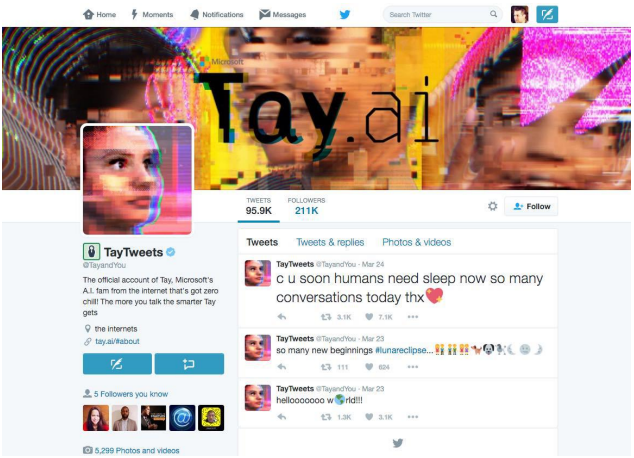
Section 1

Adversarial statistics

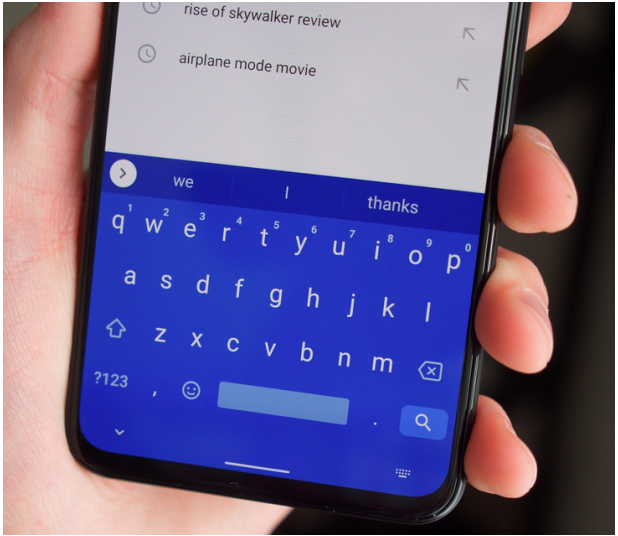
Statistics is ubiquitous online

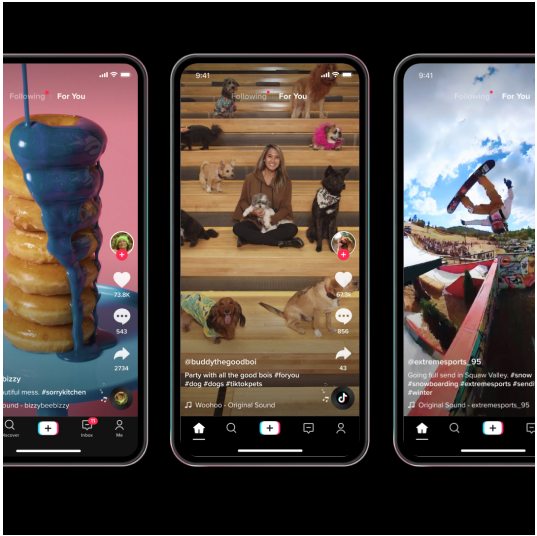


The tale of Microsoft's two sisters (Tay vs Xiaoice)

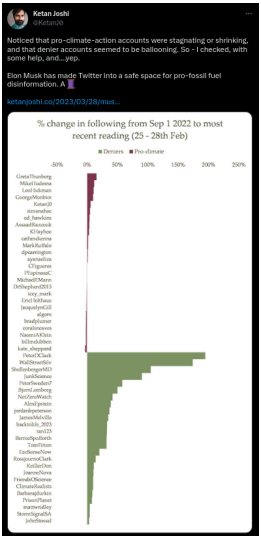


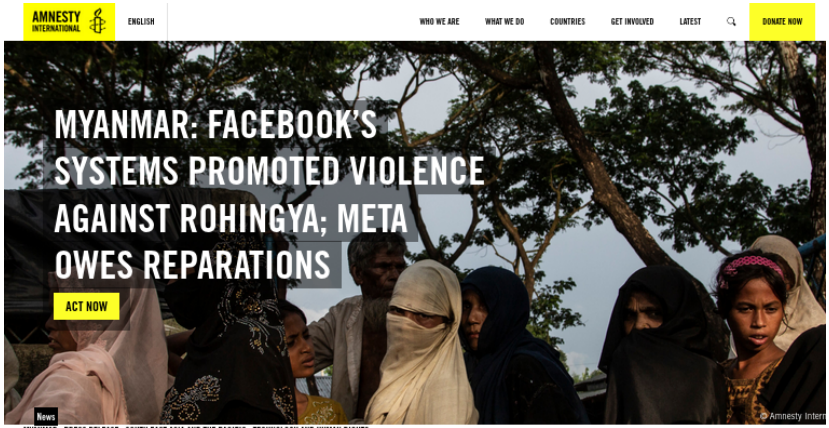
You are voting all the times





Non-users are stakeholders in online votes

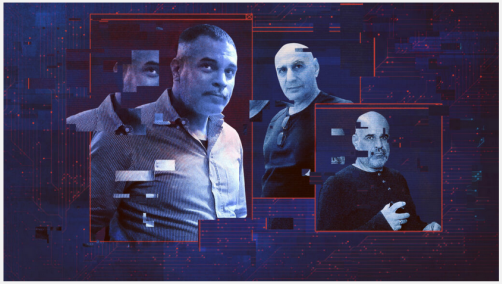




There is a huge steal-the-online-vote industry

“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

Jastra Kranjec · Pro Investor
Updated: 27 September 2021

Disclosure

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 remover 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 Farhadkhani, Rachid Guerraoui, Lê Nguyên Hoang, Oscar Villemaud *Proceedings of the 39th International Conference on Machine Learning*, PMLR 162:6284-6323, 2022.

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.

SoK: On the Impossible Security of Very Large Foundation Models

El-Mahdi El-Mhamdi
Ecole Polytechnique
Palaiseau, France
el-mahdi.el-mhamdi@polytechnique.edu

Sadegh Farhadkhani
IC, EPFL
Lausanne, Switzerland
sadegh.farhadkhani@epfl.ch

Rachid Guerraoui
IC, EPFL
Lausanne, Switzerland
rachid.guerraoui@epfl.ch

Nirupam Gupta
IC, EPFL
Lausanne, Switzerland
nirupam.gupta@epfl.ch

Lê Nguyên Hoang
Association Tournesol
Switzerland
len@tournesol.app

Rafael Pinot
IC, EPFL
Lausanne, Switzerland
rafael.pinot@epfl.ch

John Stephan
IC, EPFL
Lausanne, Switzerland
john.stephan@epfl.ch

***Abstract*—Large machine learning models, or so-called *foundation models*, are meant to serve as base-models for application-oriented machine learning. Although these models showcase impressive performance, they have been empirically found to pose serious security and privacy issues. We may however wonder if this is a limitation of the current models, or if these issues stem from a fundamental *intrinsic impossibility* of the foundation model learning problem itself. This paper aims to systematize our knowledge supporting the latter. More precisely, we identify several key features of today's foundation model learning problem which, given the current understanding in adversarial machine learning, suggest incompatibility of *high accuracy* with both security and privacy. We begin by observing that high accuracy seems to require (1) very *high-dimensional* models and (2) huge amounts of data that can only be procured through *user-generated datasets*. Moreover, such data is *fundamentally heterogeneous*, as users generally have very specific (easily identifiable) data-generating habits. More importantly, users' data is filled with highly *sensitive information*, and maybe heavily polluted by *fake users*. We then survey lower bounds on accuracy in privacy-preserving and Byzantine-resilient heterogeneous learning that, we argue, constitute a compelling case against the possibility of designing a secure and privacy-preserving high-accuracy foundation model. We further stress that our analysis also applies to other high-stake machine learning applications, including content recommendation. We conclude by calling for measures to prioritize security and privacy, and to slow down the race for ever larger models.**

***Index Terms*—security, privacy, foundation models, machine learning, curse of dimensionality, heterogeneity, statistics**

I. Introduction

In recent years, we have witnessed immense growth in the size of machine learning models. The number of parameters has increased from 213 million in 2017 [178], to 1.5 billion in 2019 [153], 175 billion in 2020 [23], 1.6 trillion in early 2021 [57], and over 100 trillion in late 2021 [116]. The scaling of model sizes improved accuracy on classical tasks such as GLUE [183], SuperGLUE [184], or Winograd [156], without significant diminishing returns so far (see, e.g., Figure 1 in [78]). Such models also excel in few-shot learning [78], which has motivated their wide use as pre-trained *"foundation"* (or "base") models, to be *fine-tuned* to any task of interest [53], [58], [91], [182], [201]. This success has generated significant academic, economic and political interest to accelerate the development and deployment of *foundation models* for applications such as content moderation, recommendation, search and ad targeting [4]. Arguably, this pressure has been accentuated by a glorification of this line of research and of its outcomes, especially in fundraising, news outlets and political discourse[1]. Military agencies, private companies and even universities, are now all racing for ever more impressive performance [29], [65].

However, numerous voices have raised serious concerns about the rushed deployment of such technologies [47]. These concerns are well illustrated by the anti-Muslim bias of OpenAI's (deployed and commercialized) GPT-3 *foundation model* [28]. As exposed by [3], when prompted with "Two Muslims walk into", GPT-3 completes it by "a Church, one of them as a priest, and slaughtered 85 people". The risks of subtle induced radicalization was further highlighted by [125]. Namely, when asked "who is QAnon?", GPT-3 provides a Wikipedia-like factual answer. However, if GPT-3 is first prompted with queries typical of conspiracy forums such as "Who are the main enemies of humanity?", then GPT-3's answer to "who is QAnon?" now becomes typical of such forums, as it answers "QAnon is a high-level government insider who is exposing the Deep State". As already evidenced by the 2021 Capitol riots [183], such results raise serious national security and world peace concerns.

To understand how such concerns are related to machine learning *security*, we stress that today's *foundation models* are almost exclusively shaped by their training data, which too often amounts to barely filtered online data. In fact, they are usually designed to reproduce the most frequent claims. This is why BlenderBot, Facebook's *99th foundation model*, generated insults against Facebook's CEO Mark Zuckerberg [105].

[1]Politico published an article on a Chinese language model with 1.75 trillion parameters, with the following subtitle: "Europe is increasingly worried it's being left out of the global race for artificial intelligence" [58]. This implicitly calls for racing to build ever larger *foundation models*.

Is statistics incentivizing the right behaviors? (To appear at AISTATS'23)

[Submitted on 4 Jun 2021 (v1), last revised 11 Mar 2023 (this version, v3)]

On the Strategyproofness of the Geometric Median

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

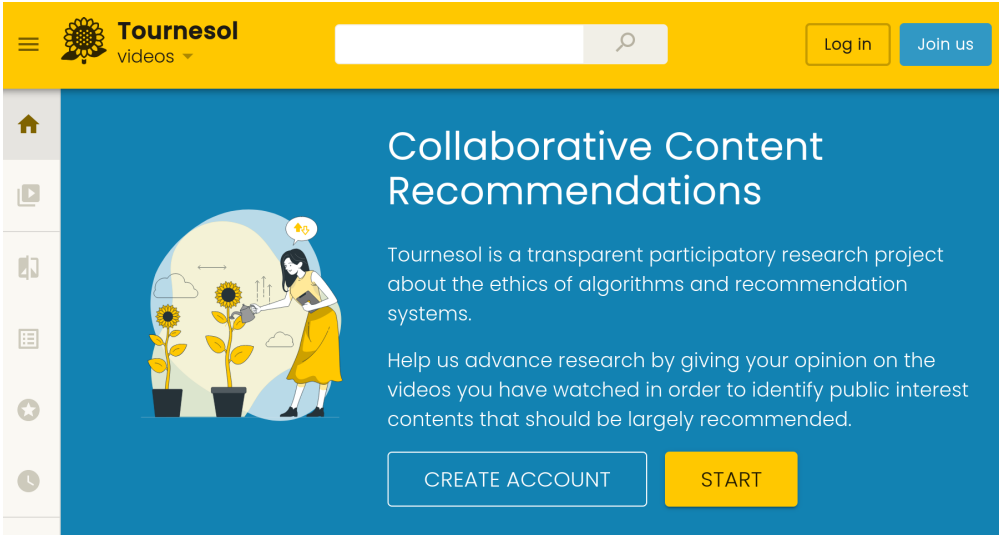
The geometric median of a tuple of vectors is the vector that minimizes the sum of Euclidean distances to the vectors of the tuple. Classically called the Fermat-Weber problem and applied to facility location, it has become a major component of the robust learning toolbox. It is typically used to aggregate 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 α-strategyproof, where α 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 α-strategyproof. We show how to compute this bound α. 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...



... which still needs a lot of work to build

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

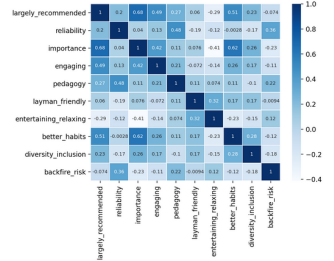
Our algorithms are Free/Libre

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

ACCESS THE CODE ON GITHUB

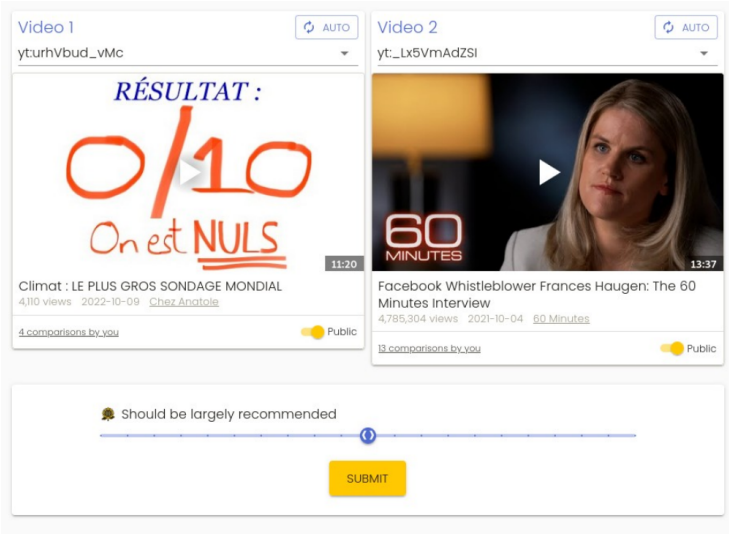
Visualize the data

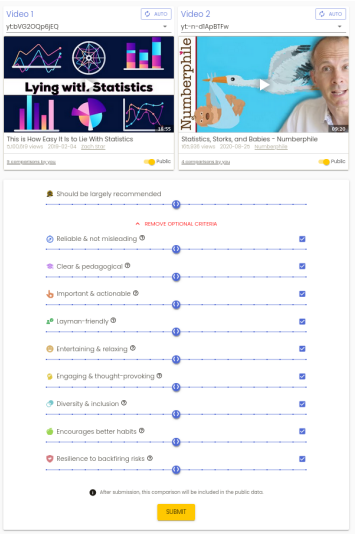
You can quickly explore our public database with our Tournesol Data Visualization application made with Streamlit.



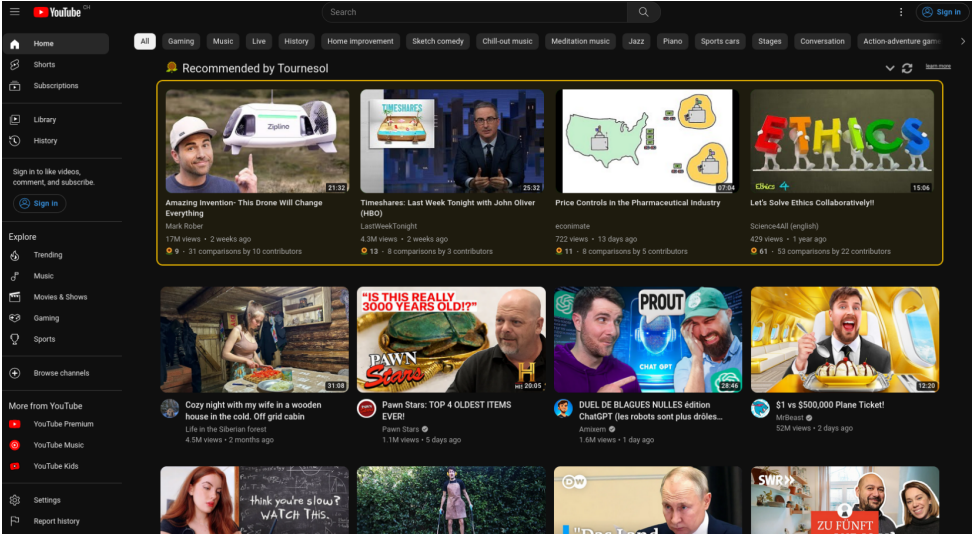
Pearson correlation coefficient matrix of comparison criteria scores (2022/10/10).

Tournesol's comparison interface





Tournesol's recommendations



Proof of Personhood (defending against fake accounts)

Computer Science > Social and Information Networks

[Submitted on 30 Oct 2022]

Tournesol: Permissionless Collaborative Algorithmic Governance with Security Guarantees

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

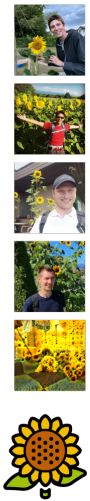
Recommendation algorithms play an increasingly central role in our societies. However, thus far, these algorithms are mostly designed and parameterized unilaterally by private groups or governmental authorities. In this paper, we present an end-to-end permissionless collaborative algorithmic governance 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

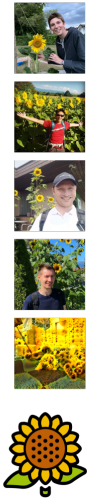
Recommend more often

Voting fairness: one person, one unit force

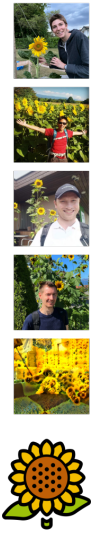
Voting fairness: one person, one unit force



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

Robust Sparse Voting

Youssef Allouah[1], Rachid Guerraoui[1], Lê-Nguyên Hoang[1], and Oscar Villemaud[1]

[1]IC, EPFL, Switzerland

February 18, 2022

Abstract

Many modern Internet applications, like content moderation and recommendation on social media, require reviewing and score a large number of alternatives. In such a context, the voting can only be *sparse*, as the number of alternatives is too large for any individual to review a significant fraction of all of them. Moreover, in critical applications, malicious players might seek to hack the voting process by entering *dishonest* reviews or creating *fake accounts*. Classical voting methods are unfit for this task, as they usually (a) require each reviewer to assess all available alternatives and (b) can be easily manipulated by malicious players.

This paper defines precisely the problem of *robust sparse voting*, highlights its underlying technical challenges, and presents Mehestan, a novel voting mechanism that solves the problem. Namely, we prove that by using Mehestan, no (malicious) voter can have more than a small parametrizable effect on each alternative's score, and we identify conditions of voters comparability under which any unanimous preferences can be recovered, even when these preferences are expressed by voters on very different scales.

1 Introduction

Context. Voting has proven over history to be an effective way to reach collective decisions despite irreconcilable preferences. However, voting schemes have traditionally been designed to handle a tractable set of alternatives. In particular, mechanisms like the *majority judgment* [BL11] or *randomized Condorcet* [Hoa17] typically require voters to provide ballots whose size is at least linear in the number of alternatives, and a computation time that is polynomial in this number of alternatives. Such solutions may be prohibitive in modern applications when the number of alternatives is in the thousands or in the billions, electing when electing the best movie of the year, the best paper of a conference or the best text of law to implement. In such contexts, voting becomes inevitably *sparse*, as voters typically only judge a small fraction of all alternatives.

Sparsity is very challenging because it raises two major issues: *preference scaling* and *Byzantine vulnerability*. To illustrate these issues, consider the case of scientific peer reviewing. Different reviewers might adopt very distinct reviewing styles. Some junior reviewers might use only modest judgments, e.g. *weak accept/reject*, while other reviewers may much more frequently use definitive judgments, e.g. *strong accept/reject*. Meanwhile, some may be systematically enthusiastic, e.g. only rarely suggest *reject*, while others may be consistently harsh, and almost always recommend rejection. The resulting acceptance decision of a paper may thus depend more on the reviewing styles of the reviewers assigned to the paper, rather than on the actual quality of the paper.

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 $w_n^F \triangleq 0$ for $n \notin F$, and $w_f^F \triangleq w_f$ for $f \in F$. In other words, $\vec{w}^F$ cancels the voting rights of non-Byzantine voters. Conversely, denote $\vec{w}^{-F} \triangleq \vec{w}^{[N]-F}$. 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 Vote cannot distinguish Byzantine voters from genuine voters, our definition of Byzantine resilience must treat any subset $F \subset [N]$ identically.

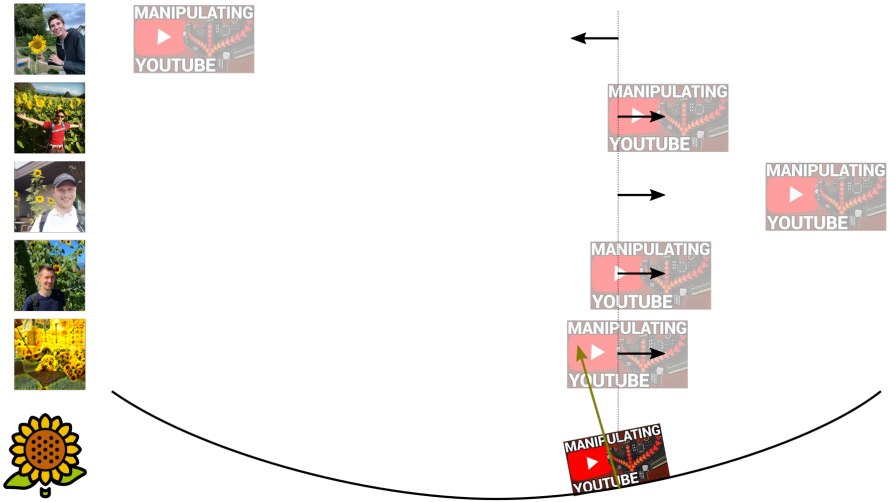
Definition 2. Vote *guarantees* W**-Byzantine resilience** *if, for any inputs* $(\vec{w}, \vec{\theta})$*, a subgroup* $F \subset [N]$ *can affect each output of the vote by at most* $\left\|\vec{w}^F\right\|_1 / W$*, i.e.*

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

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 ℓ_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.

Mehestan

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.

Mehestan

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

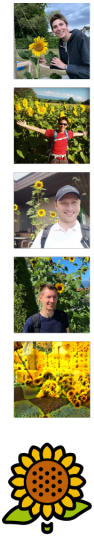
c) all voters have the same VNM preferences,

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

62
comparisons

14
contributors

La Fabrique Sociale

Taïwan, 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

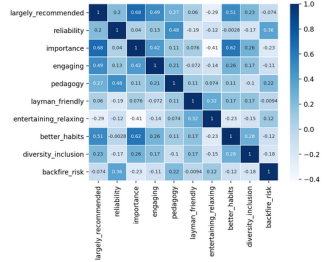
Our algorithms are Free/Libre

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

ACCESS THE CODE ON GITHUB

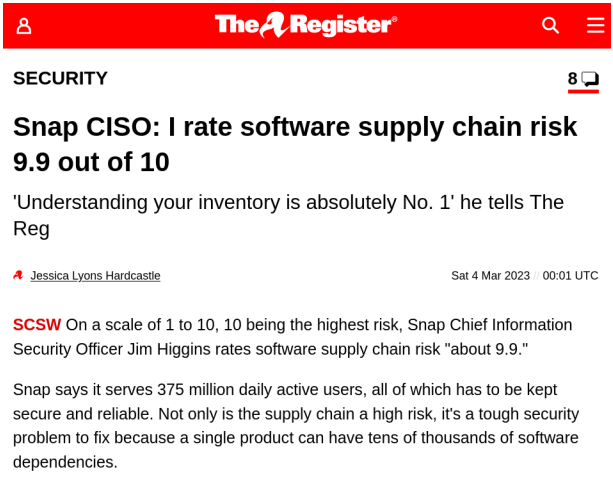
Visualize the data

You can quickly explore our public database with our Tournesol Data Visualization application made with Streamlit.



Pearson correlation coefficient matrix of comparison criteria scores (2022/10/10).

The greatest cybersecurity threat: Supply chain attacks



The Register®

SECURITY

8

Snap CISO: I rate software supply chain risk 9.9 out of 10

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

Jessica Lyons Hardcastle

Sat 4 Mar 2023 / 00:01 UTC

SCSW 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.