

Spectacular & Secure NO YOU CAN'T

Lê Nguyễn Hoàng,
Calicarpa & Tournesol
@lenhoang@mastodon.social



CIA Seminar, November 2022



YES










WE

CAN





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](#)

 686  20,948










    eReader  PDF

     54   

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 $1.73 \times$ communication expansion for 2^{10} users and 2^{20} -dimensional vectors, and $1.98 \times$ expansion for 2^{14} users and 2^{24} -dimensional vectors over sending data in the clear.






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](#)

 686  20,948

    eReader  PDF

     54  


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.

Practical secure aggregation for privacy-preserving machine learning










[K Bonawitz](#), [V Ivanov](#), [B Kreuter](#), [A Marcedone](#)... - proceedings of the ..., 2017 - dl.acm.org

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 (ie 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 ...

☆ [Enregistrer](#)  [Citer](#) [Cité 1620 fois](#) [Autres articles](#) [Les 7 versions](#)

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](#)

 686  20,948





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.

Practical secure aggregation for privacy-preserving machine learning

[K Bonawitz](#), [V Ivanov](#), [B Kreuter](#), [A Marcedone](#)... - proceedings of the ..., 2017 - dl.acm.org

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 (ie 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 ...

 [Enregistrer](#)  [Citer](#) [Cité 1620 fois](#) [Autres articles](#) [Les 7 versions](#)












"federated learning is a privacy preserving"

Articles

Environ 43 résultats (0,11 s)






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](#)

 686  20,948

    eReader  PDF

     54  


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.

Practical secure aggregation for privacy-preserving machine learning

[K Bonawitz](#), [V Ivanov](#), [B Kreuter](#), [A Marcedone](#)... - proceedings of the ..., 2017 - dl.acm.org

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 (ie 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 ...

☆ [Enregistrer](#)  [Citer](#) [Cité 1620 fois](#) [Autres articles](#) [Les 7 versions](#)

 "federated learning is a privacy preserving"

Articles Environ 43 résultats (0,11 s)

Federated Learning allows for smarter models, lower latency, and less power consumption, **all while ensuring privacy**. And this approach has another immediate benefit: in addition to

ACM DIGITAL LIBRARY

CCS ▼

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](#)

686 20,948

[eReader](#) [PDF](#)

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 $1.73 \times$ communication expansion for 2^{10} users and 2^{20} -dimensional vectors, and $1.98 \times$ expansion for 2^{14} users and 2^{24} -dimensional vectors over sending data in the clear.

Practical secure aggregation for privacy-preserving machine learning

K Bonawitz, V Ivanov, B Kreuter, A Marcedone... - proceedings of the ... , 2017 - dl.acm.org

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 (ie 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 ...

☆ Enregistrer Citer Cité 1620 fois Autres articles Les 7 versions

Google Scholar "federated learning is a privacy preserving"

Articles Environ 43 résultats (0,11 s)

Federated Learning allows for smarter models, lower latency, and less power consumption, all while ensuring privacy. And this approach has another immediate benefit: in addition to

APPLIED FEDERATED LEARNING: IMPROVING GOOGLE KEYBOARD QUERY SUGGESTIONS

Timothy Yang*, Galen Andrew*, Hubert Eichner*
Haicheng Sun, Wei Li, Nicholas Kang, Daniel Ramage, Françoise Beaufays

Google LLC,
Mountain View, CA, U.S.A.

FEDERATED LEARNING FOR MOBILE KEYBOARD PREDICTION

Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, Daniel Ramage

Google LLC,
Mountain View, CA, U.S.A.
{harda, kanishkarao, mathews, swaroopram, fsb saugenst, huberte, loeki, dramage}@google.com

ABSTRACT

Federating where both client and server are decentralized. In this paper, we describe 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 $1.73 \times$ communication expansion for 2^{10} users and 2^{20} -dimensional vectors, and $1.98 \times$ expansion for 2^{14} users and 2^{24} -dimensional vectors over sending data in the clear.

We train a recurrent neural network language model using a distributed, on-device learning framework called federated learning for the purpose of next-word prediction in a virtual keyboard for smartphones. Server-based training using stochastic gradient descent is compared with training on client devices using the Federated Averaging algorithm. The federated algorithm, which enables training on a higher-quality dataset for this use case, is shown to achieve better prediction recall. This work demonstrates the feasibility and benefit of training language models on client devices without exporting sensitive user data to servers. The federated learning environment gives users greater control over the use of their data and simplifies the task of incorporating privacy by default with distributed training and aggregation across a population of client devices.

Index Terms— Federated learning, keyboard, language modeling, NLP, CIFG.

1. INTRODUCTION

Gboard — the Google keyboard¹ — is a virtual keyboard for touchscreen mobile devices with support for more than 600 language varieties and over 1 billion installs as of 2019. In addition to decoding noisy signals from input modalities including tap and word-gesture typing, Gboard provides auto-correction, word completion, and next-word prediction features.

As users increasingly shift to mobile devices [1], reliable and fast mobile input methods become more important. Next-word predictions provide a tool for facilitating text entry. Based on a small amount of user-generated preceding text, language models (LMs) can predict the most probable next word or phrase. Figure 1 provides an example: given the text, “I love you”, Gboard predicts the user is likely to type “and”, “too”, or “so much”. The center position in the suggestion strip is reserved for the highest-probability

¹gboard.app.google.com



Fig. 1. Next word predictions in Gboard. Based on the context “I love you”, the keyboard predicts “and”, “too”, and “so much”.

candidate, while the second and third most likely candidates occupy the left and right positions, respectively.

Prior to this work, predictions were generated with a word n -gram finite state transducer (FST) [2]. The mechanics of the FST decoder in Gboard — including the role of the FST in literal decoding, corrections, and completions — are described in Ref. [3]. Next word predictions are built by searching for the highest-order n -gram state that matches the preceding text. The n -best output labels from this state are returned. Paths containing back-off transitions to lower orders are also considered. The primary (static) language model for the English language in Gboard is a Katz smoothed Bayesian interpolated [4] 5-gram LM containing 1.25 million n -grams, including 164,000 unigrams. Personalized user history, contacts, and email n -gram models augment the primary LM.

Mobile keyboard models are constrained in multiple ways. In order to run on both low and high-end devices, models should be small and inference-time latency should be low. Users typically expect a visible keyboard response

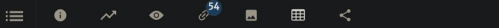
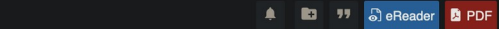
Practical Secure Aggregation for Privacy-Preserving Machine Learning

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](#)

686 20,948



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 $1.73 \times$ communication expansion for 2^{10} users and 2^{20} -dimensional vectors, and $1.98 \times$ expansion for 2^{14} users and 2^{24} -dimensional vectors over sending data in the clear.

Practical secure aggregation for privacy-preserving machine learning

K Bonawitz, V Ivanov, B Kreuter, A Marcedone... - proceedings of the ..., 2017 - dl.acm.org

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 (ie 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 ...

☆ Enregistrer Citer Cité 1620 fois Autres articles Les 7 versions



"federated learning is a privacy preserving"

Articles

Environ 43 résultats (0,11 s)

Federated Learning allows for smarter models, lower latency, and less power consumption, all while ensuring privacy. And this approach has another immediate benefit: in addition to

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.

APPLIED FEDERATED LEARNING: IMPROVING GOOGLE KEYBOARD QUERY SUGGESTIONS

Timothy Yang*, Galen Andrew*, Hubert Eichner*
Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, Françoise Beaufays

Google LLC,
Mountain View, CA, U.S.A.

FEDERATED LEARNING FOR MOBILE KEYBOARD PREDICTION

Andrew Hard, Kanishka Rao, Rajiv Mathews, Svaroop Ramaswamy, Françoise Beaufays
Sean Augenstein, Hubert Eichner, Chloé Kiddon, Daniel Ramage

Google LLC,
Mountain View, CA, U.S.A.
{harda, kanishkara, mathews, svaroopram, fsb saugenst, huberte, loeki, dramage}@google.com

Federated learning where both client and server are distributed. In this paper, we describe a federated learning framework for improving keyboard query suggestions on mobile devices. We describe the system architecture and the federated learning algorithm. We describe the data collection process and the federated learning algorithm. We describe the evaluation process and the results of the federated learning algorithm.

The introduction of federated learning where the training is done on the client devices and the aggregation is done on the server. This approach has several advantages: it reduces the amount of data that needs to be sent to the server, it reduces the risk of data leakage, and it allows for training on devices that are not always connected to the internet.

ABSTRACT

We train a recurrent neural network language model using a distributed, on-device learning framework called federated learning for the purpose of next-word prediction in a virtual keyboard for smartphones. Server-based training using stochastic gradient descent is compared with training on client devices using the Federated Averaging algorithm. The federated algorithm, which enables training on a higher-quality dataset for this use case, is shown to achieve better prediction recall. This work demonstrates the feasibility and benefit of training language models on client devices without exporting sensitive user data to servers. The federated learning environment gives users greater control over the use of their data and simplifies the task of incorporating privacy by default with distributed training and aggregation across a population of client devices.

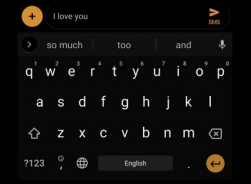


Fig. 1. Next word predictions in Gboard. Based on the context "I love you", the keyboard predicts "and", "text", and "so".

ACM DIGITAL LIBRARY

CCS ▾

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](#)

686 20,948

Here, we argue that our protocol is a secure multiparty computation in the honest but curious setting, regardless of how and when parties abort. In particular, we prove that when executing the protocol with threshold t , the joint view of the server and any set of less than t (honest) users does not leak any information about the other users' inputs, besides what can be inferred from the output of the computation. Before formally stating our result, we introduce some notation.

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 $1.73 \times$ communication expansion for 2^{10} users and 2^{20} -dimensional vectors, and $1.98 \times$ expansion for 2^{14} users and 2^{24} -dimensional vectors over sending data in the clear.

ACM DIGITAL LIBRARY

CCS

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](#)

686 20,948

eReader PDF

54

Here, we argue that our protocol is a secure multiparty computation in the honest but curious setting, regardless of how and when parties abort. In particular, we prove that when executing the protocol with threshold t , the joint view of the server and any set of less than t (honest) users does not leak any information about the other users' inputs, besides what can be inferred from the output of the computation. Before formally stating our result, we introduce some notation.

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.

The output is a model that *literally* learned from the data!

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](#)

686 20,948

🔔 📄 🗨️ [eReader](#) [PDF](#)

📄 📊 👁️ 🔗 📄 📅 🔗

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.

Extracting Training Data from Large Language Models

Authors:

Nicholas Carlini, *Google*; Florian Tramèr, *Stanford University*; Eric Wallace, *UC Berkeley*; Matthew Jagielski, *Northeastern University*; Ariel Herbert-Voss, *OpenAI and Harvard University*; Katherine Lee and Adam Roberts, *Google*; Tom Brown, *OpenAI*; Dawn Song, *UC Berkeley*; Úlfar Erlingsson, *Apple*; Alina Oprea, *Northeastern University*; Colin Raffel, *Google*

Abstract:

It has become common to publish large (billion parameter) language models that have been trained on private datasets. This paper demonstrates that in such settings, an adversary can perform a training data extraction attack to recover individual training examples by querying the language model.

We demonstrate our attack on GPT-2, a language model trained on scrapes of the public Internet, and are able to extract hundreds of verbatim text sequences from the model's training data. These extracted examples include (public) personally identifiable information (names, phone numbers, and email addresses), IRC conversations, code, and 128-bit UUIDs. Our attack is possible even though each of the above sequences are included in just one document in the training data.





We comprehensively evaluate our extraction attack to understand the factors that contribute to its success. Worryingly, we find that larger models are more vulnerable than smaller models. We conclude by drawing lessons and discussing possible safeguards for training large language models.

ACM DIGITAL LIBRARY

EC



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](#)

2   114

   [Get Access](#)

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 between 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.

A full version of this paper is available at <https://papers.ssrn.com>

[/sol3/papers.cfm?abstract_id=3685465](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3685465)

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.

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/10.1145/3485456.3467580>

Online: 18 July 2021 Publication History

114

Get Access

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 between 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.

A full version of this paper is available at <https://papers.ssrn.com>

/sol3/papers.cfm?abstract_id=3685465

Overall, our results indicate that online platforms can effectively combat fraudulent users without incurring large costs by designing new learning

Summary of Contributions. In this work, we follow a regret analysis framework and assess the performance of learning algorithms by proving worst-case guarantees parameterized by the number of fake users F , which we refer to as the *fakeness budget*. Given the above model, we show the following results.

1. We show that commonly used learning algorithms for product ranking are vulnerable to fake users in that their regret can be $\Omega(T)$, even when the number of fake users is small.
2. For the setting where the *fakeness budget* F is known to the platform, we design a deterministic online algorithm called *Fake-Aware Ranking (FAR)* whose worst-case regret is $O(\log(T) + F)$.
3. For a more challenging setting where the fakeness budget is unknown to the platform, we design a randomized online algorithm called *Fake-Oblivious Ranking with Cross-Learning (FORC)* whose worst-case regret is $O(F \log(T))$.
4. Finally, we carry out a numerical study using synthetic data that illustrates the superior performance of FORC even though the algorithm is unaware of the fakeness budget.

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](#)

114

Get Access

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 between 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.

A full version of this paper is available at <https://papers.ssrn.com>

[/sol3/papers.cfm?abstract_id=3685465](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3685465)

Overall, our results indicate that online platforms can effectively combat fraudulent users without incurring large costs by designing new learning algorithms even when the number of fake users is large.

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.

ACM DIGITAL LIBRARY

EC

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](#)

2 114

[Get Access](#)

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 between 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.

A full version of this paper is available at <https://papers.ssrn.com>

[/sol3/papers.cfm?abstract_id=3685465](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3685465)

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.

Making Byzantine Decentralized Learning Efficient

Sadegh Farhadkhani¹, Rachid Guerraoui¹, Nirupam Gupta¹,
Lê Nguyễn Hoàng², Rafael Pinot¹, and John Stephan^{*1}

¹IC, EPFL

`[firstname.lastname]@epfl.ch`

²Tournesol Association

`len@tournesol.app`

Abstract

Decentralized-SGD (D-SGD) distributes heavy learning tasks across multiple machines (a.k.a., *nodes*), effectively dividing the workload per node by the size of the system. However, a handful of *Byzantine* (i.e., misbehaving) nodes can jeopardize the entire learning procedure. This vulnerability is further amplified when the system is *asynchronous*. Although approaches that confer Byzantine resilience to D-SGD have been proposed, these significantly impact the efficiency of the process to the point of even negating the benefit of decentralization. This naturally raises the question: *can decentralized learning simultaneously enjoy Byzantine resilience and reduced workload per node?*

A few prior works addressed the challenge of Byzantine resilience in asynchronous decentralized learning [15, 16, 28]. However, in [28] the loss function is assumed to be strongly convex, which is seldom true in modern-day ML. Furthermore, in [16] convergence is only guaranteed asymptotically. Lastly, the algorithm proposed in [15] inflicts a prohibitively large computational overhead on honest (i.e., non-Byzantine) nodes: each honest node computes $\mathcal{O}(1/\epsilon^5)$ gradients to attain ϵ -stationarity. This nullifies the computational benefit of distributing the learning process. Indeed, an honest node can solve the learning problem more efficiently using only $\mathcal{O}(1/\epsilon^2)$ gradients by simply running SGD locally (a.k.a., *local SGD*), without coordination. This naturally raises an important question:

Can asynchronous decentralized learning simultaneously enjoy Byzantine resilience and reduced computational workload per node?

Contributions. We answer this question positively by introducing MONNA, a new asynchronous decentralized learning algorithm. In a system of n nodes including f Byzantines, an honest node executing MONNA computes $\mathcal{O}((1+f)/n\epsilon^2)$ gradients to reach ϵ -stationarity, which is comparable to D-SGD when $f \ll n$. Essentially, we grant Byzantine robustness to D-SGD by incorporating three key components.

A few prior works addressed the challenge of Byzantine resilience in asynchronous decentralized learning [15, 16, 28]. However, in [28] the loss function is assumed to be strongly convex, which is seldom true in modern-day ML. Furthermore, in [16] convergence is only guaranteed asymptotically. Lastly, the algorithm proposed in [15] inflicts a prohibitively large computational overhead on honest (i.e., non-Byzantine) nodes: each honest node computes $\mathcal{O}(1/\epsilon^5)$ gradients to attain ϵ -stationarity. This nullifies the computational benefit of distributing the learning process. Indeed, an honest node can solve the learning problem more efficiently using only $\mathcal{O}(1/\epsilon^2)$ gradients by simply running SGD locally (a.k.a., *local SGD*), without coordination. This naturally raises an important question:

Can asynchronous decentralized learning simultaneously enjoy Byzantine resilience and reduced computational workload per node?

Contributions. We answer this question positively by introducing MONNA, a new asynchronous decentralized learning algorithm. In a system of n nodes including f Byzantines, an honest node executing MONNA computes $\mathcal{O}((1+f)/n\epsilon^2)$ gradients to reach ϵ -stationarity, which is comparable to D-SGD when $f \ll n$. Essentially, we grant Byzantine robustness to D-SGD by incorporating three key components.

Extension to heterogeneity. When the honest nodes do not have identical data distributions, achieving Byzantine resilience becomes much more challenging [15]. Nevertheless, we also analyze the convergence of our algorithm under data heterogeneity, and obtain an error matching the existing lower bound [31]. The result is deferred to Appendix B for pedagogical reasons.

No you can't
(mathematically)



[Submitted on 30 Sep 2022]

SoK: On the Impossible Security of Very Large Foundation Models

El-Mahdi El-Mhamdi, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Lê-Nguyen Hoang, Rafael Pinot, John Stephan

Large machine learning models, or so-called foundation models, aim 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.

Heterogeneity is a (privacy) killer

Theorem 1 (Theorem 4 in [96]¹¹). *For any (ε, δ) -differentially private mechanism $\widehat{\text{MEAN}}$ for the mean estimation problem, there exists an input \vec{x} with large mean squared error, as*

$$\mathbb{E} \left[\left\| \widehat{\text{MEAN}}(\vec{x}) - \vec{x} \right\|_2^2 \right] \geq \Omega \left(\frac{\sigma(\varepsilon, \delta) d \Delta^2}{N^2 (\log 2d)^4} \right), \quad (4)$$

where σ is a positive and non-increasing function.

Corollary

Assume $\Delta = \Theta(\sqrt{d})$.

Then privacy requires error $\Omega(d/N)$.

And yet differential privacy is (deeply) flawed...



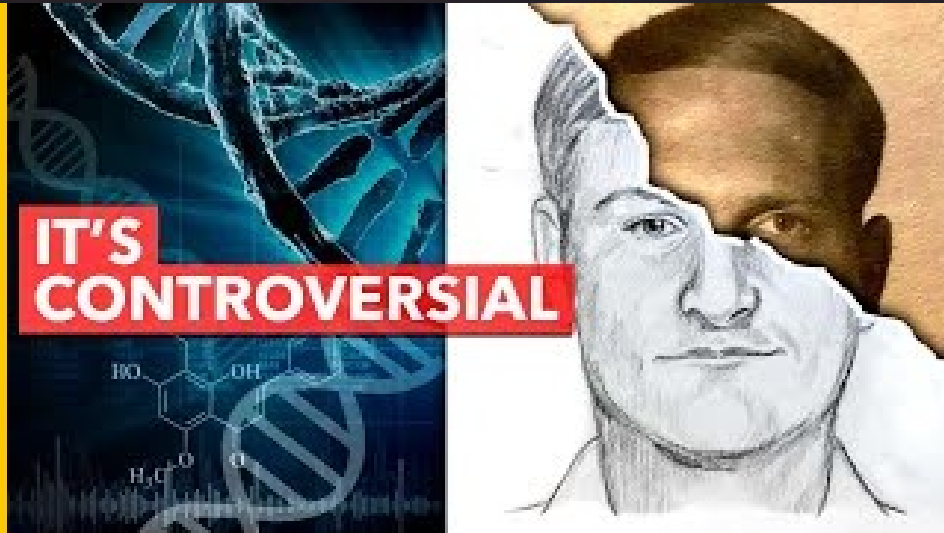
17

20

comparisons

3

contributors



Veritasium

How They Caught The Golden State Killer

Heterogeneity is a (security) killer

Theorem 2. *No algorithm $\widehat{\text{MEAN}}$ can guarantee*¹⁵

$$\forall \vec{x} \in \mathcal{B}_d(0, \Delta)^N, \forall H \subset [N] \text{ s.t. } |H| = N - f,$$

$$\left\| \widehat{\text{MEAN}}(\vec{x}) - \bar{x}_H \right\|_2^2 \leq \frac{f^2}{2(N - f)^2} \Delta^2,$$

where \bar{x}_H is the mean of honest vectors \vec{x}_H .

Corollary

Error grows as $\Omega(\Delta f / N)$.

And yet the (classical) Byzantine model is (deeply) flawed...



52

14

comparisons

4

contributors



Temps Présent

Fake news, une pandémie de mensonge...

No we can't
(morally speaking)



ACM DIGITAL LIBRARY

CCS

RESEARCH ARTICLE OPEN ACCESS

Practical Secure Aggregation for Privacy-Preserving Machine Learning

Authors: Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Macedone, 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](#)

686 20,948

[eReader](#) [PDF](#)

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 \times\$ communication expansion for \$2^{10}\$ users and \$2^{20}\$-dimensional vectors, and a \$1.98 \times\$ expansion for \$2^{14}\$ users and \$2^{24}\$-dimensional vectors over sending data in the clear.

OpenReview.net

Large Language Models Can Be Strong Differentially Private Learners

[PDF](#)

Xuechen Li, Florian Tramer, Percy Liang, Tatsunori Hashimoto

29 Sept 2021 (modified: 16 Mar 2022) ICLR 2022 Oral Readers: 48

Everyone [Show Bibtext](#) [Show Revisions](#)

Keywords: language model, differential privacy, language generation, fine-tuning, NLP

Abstract: Differentially Private (DP) learning has seen limited success for building large deep learning models of text, and straightforward attempts at applying Differentially Private Stochastic Gradient Descent (DP-SGD) to NLP tasks have resulted in large performance drops and high computational overhead.

We show that this performance drop can be mitigated with (1) the use of large pretrained language models; (2) non-standard hyperparameters that suit DP optimization; and (3) fine-tuning objectives which are aligned with the pretraining procedure.

With the above, we obtain NLP models that outperform state-of-the-art DP-trained models under the same privacy budget and strong non-private baselines—by directly fine-tuning pretrained models with DP optimization on moderately-sized corpora.

To address the computational challenge of running DP-SGD with large Transformers, we propose a memory saving technique that allows clipping in DP-SGD to run without instantiating per-example gradients for any linear layer in the model.

The technique enables privately training Transformers with almost the same memory cost as non-private training at a modest run-time overhead. Contrary to conventional wisdom that DP optimization fails at learning high-dimensional models (due to noise that scales with dimension) empirical results reveal that private learning with pretrained language models tends to not suffer from dimension-dependent performance degradation.

Code to reproduce results can be found at <https://github.com/lxuechen/private-transformers>.

One-sentence Summary: We show how to build highly performant differentially private NLP models by fine-tuning large pretrained models.

Add [Public Comment](#)

Privacy solved!

ACM DIGITAL LIBRARY

EC

EXTENDED-ABSTRACT

Learning Product Rankings Robust to Fake Users

Authors: Negin Golrezaei, Yahideh 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](#)

2 114

[Get Access](#)

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 between 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.

A full version of this paper is available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3685465

Security solved!

References • 2022 IEEE Symposium on Security and Privacy

Back to the Drawing Board: A Critical Evaluation of Poisoning Attacks on Production Federated Learning

Publisher: IEEE [Cite This](#)

Virat Shejwalkar ; Amir Houmansadr ; Peter Kairouz ; Daniel Ramage

[Sign In or Purchase](#)

1 Paper Citation 349 Full Text Views

Abstract	Authors	Figures	References	Citations
Abstract: While recent works have indicated that federated learning (FL) may be vulnerable to poisoning attacks by compromised clients, their real impact on production FL systems is not fully understood. In this work, we aim to develop a comprehensive systematization for poisoning attacks on FL by enumerating all possible threat models, variations of poisoning, and adversary capabilities. We specifically put our focus on untargeted poisoning attacks, as we argue that they are significantly relevant to production FL deployments. We present a critical analysis of untargeted poisoning attacks under practical, production FL environments by carefully characterizing the set of realistic threat models and adversarial capabilities. Our findings are rather surprising: contrary to the established belief, we show that FL is highly robust in practice even when using simple, low-cost defenses. We go even further and propose novel, state-of-the-art data and model poisoning attacks, and show via an extensive set of experiments across three benchmark datasets how (in)effective poisoning attacks are in the presence of simple defense mechanisms. We aim to correct previous misconceptions and offer concrete guidelines to conduct more accurate (and more realistic) research on this topic. View less				

Privacy solved! Security solved!

ACM DIGITAL LIBRARY

RESEARCH ARTICLE OPEN ACCESS

Practical Secure Aggregation for Privacy-Preserving Machine Learning

Authors: Keith Bonawitz, Vladimir Ivanov, Ber...

Antonio Marcedone, H. Brendan McMahan, Sa...

Daniel Ramage, Aaron Segal, Karn Seth, Auth...

Online: 30 October 2017 Publication History

686 20,948

ABSTRACT

We design a novel, communication-efficient, failure-robust aggregation of high-dimensional data. Our protocol allows the sum of large, user-held data vectors from mobile devices (i.e. without learning each user's individual contribution), for example, in a federated learning setting, to aggregate user updates for a deep neural network. We prove the security in honest-but-curious and active adversary settings, and show it is maintained even if an arbitrarily chosen subset of users collude. We evaluate the efficiency of our protocol and show, by comparing with a concrete implementation, that its runtime and communication overhead are low even on large data sets and client pools. For 16-bit integers, our protocol offers 1.73 x communication expansion for 2¹⁰ users and 2²⁰-dimensional vectors, and 1.98 x expansion for 2¹⁴ users and 2²⁴-dimensional vectors over sending data in the clear.

OpenReview.net

Large Language Models Can Be Strong Differentially Private

EXTENDED-ABSTRACT

Learning Product Rankings Robust to Fake Users

Authors: Negim Golezraei, Yahideh Marshadi, Jon Schneider, Shroyas Sekar

Authors Info & Claims

Conferences > 2022 IEEE Symposium on Security and Privacy

Back to the Drawing Board: A Critical Evaluation of Poisoning Attacks on Production Federated Learning

Publisher: IEEE

Cite This

Authors: Amir Houmansadr; Peter Kairouz; Daniel Ramage

Purchase

349 Full Text Views

Authors	Figures	References	Citations
---------	---------	------------	-----------



One-sentence summary: we show how to build highly performant differentially private NLP models by fine-tuning large pretrained models.

Add [Public Comment](#)

to the number and identity of the fake users.

A full version of this paper is available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3685465

While recent works have indicated that federated learning (FL) may be vulnerable to poisoning attacks by malicious clients, their real impact on production FL systems is not fully understood. In this work, we aim to develop a comprehensive systemization for poisoning attacks on FL by exploring all possible threat models, variations of poisoning attacks, and their capabilities. We specifically put our focus on untargeted poisoning attacks, as we argue that they are the most relevant to production FL deployments. We provide a critical analysis of untargeted poisoning attacks on production FL environments by carefully exploring the set of realistic threat models and adversarial capabilities. Our findings are rather surprising: contrary to the common belief, we show that FL is highly robust in practice to poisoning attacks using simple, low-cost defenses. We go even further and propose novel, state-of-the-art data and model poisoning attacks, and show via an extensive set of experiments across three benchmark datasets how these attacks perform. Our poisoning attacks are in the presence of simple defense mechanisms. We aim to correct previous misconceptions and offer concrete guidelines to conduct more accurate (and more realistic) research on this topic. [View less](#)

Privacy solved! Security solved!

RESEARCH ARTICLE OPEN ACCESS
Practical Secure Aggregation for Privacy-Preserving Machine Learning
Authors: Keith Bonawitz, Vladimir Antonov, Antonio Marcedone, H. Brendan Koenig, Daniel Ramage, Aaron Segal
CCS '17: Proceedings of the 2017 ACM Symposium on Communications Security • October 2017 /10.1145/3133956.3133982
Online: 30 October 2017 Publication History
686 views 20,948 downloads

 72
246 comparisons
114 contributors



Science4All
La technologie que tous les écologistes oublient...

Peter Kairouz ; Daniel Ramage
Figures References

...s have indicated that federated learning is not able to protect against poisoning attacks by real impact on production FL deployments. In this work, we aim to develop a defense for poisoning attacks on FL by using heat models, variations of poisoning, We specifically put our focus on untargeted poisoning attacks on FL environments by carefully designing realistic threat models and adversarial models rather surprising: contrary to the intuition that FL is highly robust in practice, we show that FL is highly vulnerable to poisoning attacks. We go even further by showing that poisoning attacks are effective on state-of-the-art data and model poisoning defenses via an extensive set of experiments on benchmark datasets how poisoning attacks are in the presence of simple defenses. We aim to correct previous research by providing concrete guidelines to conduct more accurate (and more realistic) research on this topic. [View less](#)

dimensional vectors, and 1.98 x expansion for 2¹⁰ users and 2¹⁰-dimensional vectors over sending data in the clear.

Add Public Comment

/api/papers.cfm?abstract_id=3685465

Privacy solved!

Security solved!

RESEARCH ARTICLE | OPEN ACCESS
Practical Secure Aggregation for Privacy-Preserving Machine Learning
Authors: Keith Bonawitz, Vladislav Izraeli, Antonio M. S. Correia, H. Brendan Kwon, Daniel Ramage, Aaron Segal, et al.
CCS '17: Proceedings of the 2017 ACM Symposium on Communications Security • October 2017 /10.1145/3133956.3133982
Online: 30 October 2017
686 views, 20,948 downloads

 72
246 comparisons
114 contributors


Éthique 18
Science4All
La technologie qu

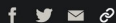
 **David Chavalarias** @chavalarias
Rappelons que la #désinformations en ligne est l'arme géopolitique qui a le meilleur rapport qualité/prix contre les démocraties. Un jour de bombardement sur l'Ukraine a coûté autour de 700M\$ au Kremlin. Imaginez ce que vous pouvez faire avec cette somme sur les réseaux sociaux!
Translate Tweet
12:27 PM · Oct 15, 2022 · Twitter Web App
7 Retweets 18 Likes

Privacy solved!

Security solved!

Facebook turned over chat messages between mother and daughter now charged over abortion

The company was served with a warrant for the messages, which experts worry could become common.



Aug. 9, 2022, 10:39 PM CEST / Updated Aug. 10, 2022, 4:51 AM CEST

By Kevin Collier and Minnyvonne Burke

Facebook turned over the chats of a mother and her daughter to Nebraska police after they were served with a warrant as part of an investigation into an illegal abortion, [court documents](#) show.

The investigation, which was launched in April before the Supreme Court overturned Roe v. Wade, is one of the few known instances of Facebook's turning over information to help law enforcement officials pursue an abortion case – but it is also an example of a scenario that abortion rights experts have warned will be more common as all abortions becomes [illegal in many states](#).

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

ACT NOW

News

© Amnesty International (Photo: Ahmer Khan)

MYANMAR PRESS RELEASE SOUTH EAST ASIA AND THE PACIFIC TECHNOLOGY AND HUMAN RIGHTS

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](#).





DEMOCRACY WORLDWIDE IN 2021

- *The level of democracy enjoyed by the average global citizen in 2021 is down to 1989 levels. The last 30 years of democratic advances are now eradicated.*
- *Dictatorships are on the rise and harbor 70% of the world population – 5.4 billion people.*
- *There are signals that the nature of autocratization is changing.*

Back to 1989 Levels

- Liberal democracies peaked in 2012 with 42 countries and are now down to the lowest levels in over 25 years – 34 nations home to only 13% of the world population.
- The democratic decline is especially evident in Asia-Pacific, Eastern Europe and Central Asia, as well as in parts of Latin America and the Caribbean.

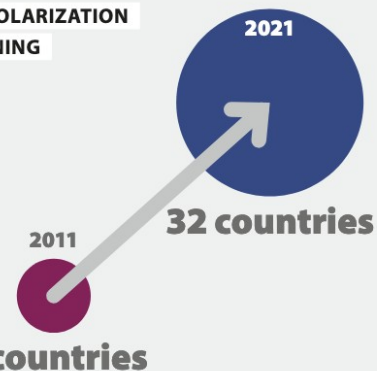
Dictatorships on the Rise

- The increasing number of closed autocracies – up from 25 to 30 countries with 26% of the world population – contributes to the changing nature of autocratization.
- Electoral autocracy remains the most common regime type and harbors 44% of the world's population, or 3.4 billion people.

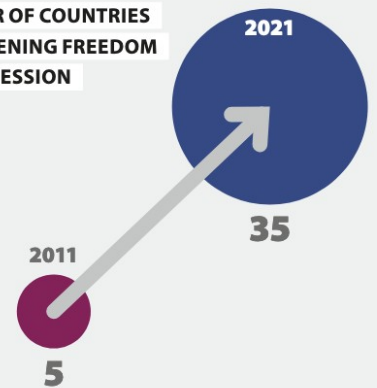
Ten Years Ago – A Different World

- A record of 35 countries suffered significant deteriorations in freedom of expression at the hands of governments – an increase from only 5 countries 10 years ago.
- A signal of toxic polarization, respect for counter-arguments and associated aspects of the deliberative component of democracy got worse in more than 32 countries – another increase from only 5 nations in 2011.

TOXIC POLARIZATION WORSENING



NUMBER OF COUNTRIES THREATENING FREEDOM OF EXPRESSION





2023 ICLR Organizing Committee

General Chair

Yan Liu (University of Southern California)

Senior Program Chair

Been Kim (Google Brain)

Program Chairs

Maximilian Nickel (FAIR)

Mengdi Wang (Princeton University)

Nancy F Chen (A*STAR)

Vukosi Marivate (University of Pretoria - Deep Learning Indaba)

Workshop Chairs

Aisha Walcott-Bryant (Google Research)

Celia Cintas (IBM Research Africa)

Hang Zhao (Tsinghua University)

Rose Yu (UC San Diego)

Diversity Equity & Inclusion Chairs

Krystal Maughan (University of Vermont)

Rosanne Liu (ML Collective, Google Brain)

2023 ICLR Organizing Committee

General Chair

Yan Liu (University of Southern California)

Senior Program Chair

Been Kim (Google Brain)

Program Chairs

Maximilian Nickel (FAIR)

Mengdi Wang (Princeton University)

Nancy F Chen (A*STAR)

Vukosi Marivate (University of Pretoria - Deep Learning Indaba)

Workshop Chairs

Aisha Walcott-Bryant (Google Research)

Celia Cintas (IBM Research Africa)

Hang Zhao (Tsinghua University)

Rose Yu (UC San Diego)

Diversity Equity & Inclusion Chairs

Krystal Maughan (University of Vermont)

Rosanne Liu (ML Collective, Google Brain)

1. Please enter a detailed review describing the strengths and weaknesses of the submission.

This paper proposes a method for equalizing the per-group means of a given set of scores while minimizing the maximum difference between the modified score and original score for any individual. The authors first present algorithms for doing this optimally in the two-group case, before using an approximation algorithm to achieve the result when more groups are considered (since they speculate that an optimal solution is no longer feasible).

The authors are careful to make it clear that they are agnostic to the desirability of the intervention they're proposing. This is the key issue with the paper in my view: it proposes a (reasonable, mathematically-sound) method for addressing a problem, but it provides no compelling reason (in the paper or the surrounding literature) why the problem needs to be addressed. As a result, the mathematical rigor feels misplaced.

2. Please provide an overall score for the submission.

Reject: Clearly below the acceptance threshold

3. Please enter a 2-3 sentence summary of your review explaining your overall score.

Unfortunately, I don't think the problem the authors introduce is one that has value to the academic community or to ML practitioners. Given this, I can't recommend the paper for publication.

There are barely any proofs, and the few sketches are not very well written. In general, while the paper's aim of establishing strategyproofness characteristics are laudable, the structure is a bit messy, and the results themselves are not very surprising (more dimensions make SP harder; when agents care about particular indices, SP cannot even be close). While the geometrical intuition provided is quite nice and rather approachable (if in a "hand wavy" way), this cannot replace a better technical insight (that is mainly in the additional material).


Ultimately, the topic -- despite the claims of deep ML connections -- does not seem significant enough, and the results aren't really surprising. I suggest expanding the paper by comparing to other mechanisms, and showing their pro and cons.

There is a disconnect between the theoretical results presented and the risks that are discussed in the paper. For example, for autocompletion, a single model does not have to be accurate for all users, in fact technologies cited in the paper most employ personalized models. Although it has been shown that extracting training data is possible [18], the claims that this can be done for all training data is not accurate. Furthermore, the data for autocompletion is not publicly recorded, eg. messages but public data that large language models are trained on, such as Wikipedia so these scen. The paper also points to search and recommendation algorithms

How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument

Published online by Cambridge University Press: 27 July 2017

GARY KING, JENNIFER PAN and MARGARET E. ROBERTS

Show author details 

Article

Figures

Supplementary materials

Metrics



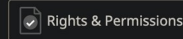
Save PDF



Share



Cite



Rights & Permissions

Abstract

The Chinese government has long been suspected of hiring as many as 2 million people to surreptitiously insert huge numbers of pseudonymous and other deceptive writings into the stream of real social media posts, as if they were the genuine opinions of ordinary people. Many academics, and most journalists and activists, claim that these so-called 50c party posts vociferously argue for the government's side in political and policy debates. As we show, this is also true of most posts openly accused on social media of being 50c. Yet almost no systematic empirical evidence exists for this claim or, more importantly, for the Chinese regime's strategic objective in pursuing this activity. In the first large-scale empirical analysis of this operation, we show how to identify the secretive authors of these posts, the posts written by them, and their content. We estimate that the government fabricates and posts about 448 million social media comments a year. In contrast to prior claims, we show that the Chinese regime's strategy is to avoid arguing with skeptics of the party and the government, and to not even discuss controversial issues. We show that the goal of this massive secretive operation is instead to distract the public and change the subject, as most of these posts involve cheerleading for China, the revolutionary history of the Communist Party, or other symbols of the regime. We discuss how these results fit with what is known about the Chinese censorship program and suggest how they may change our broader theoretical understanding of "common knowledge" and information control in authoritarian regimes.

Cause & Effect

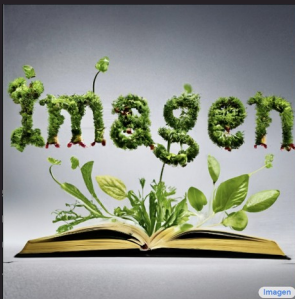
Prompt

Which of the following sentences makes more sense?

1. I studied hard because I got an A on the test.
2. I got an A on the test because I studied hard.

Model Response

I got an A on the test because I studied hard.



Sprouts in the shape of text 'Imagen' coming out of a fairytale book.



A small cactus wearing a straw hat and neon sunglasses in the Sahara desert.




An art gallery displaying Monet paintings. The art gallery is flooded. Robots are going around the art gallery using paddle boards.

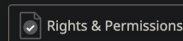
How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument

Published online by Cambridge University Press: 27 July 2017

GARY KING, JENNIFER PAN and MARGARET E. ROBERTS

Show author details 

Article Figures Supplementary materials Metrics



Abstract

The Chinese government has long been suspected of hiring as many as 2 million people to surreptitiously insert huge numbers of pseudonymous and other deceptive writings into the stream of real social media posts, as if they were the genuine opinions of ordinary people. Many academics, and most journalists and activists, claim that these so-called 50c party posts vociferously argue for the government's side in political and policy debates. As we show, this is also true of most posts openly accused on social media of being 50c. Yet almost no systematic empirical evidence exists for this claim or, more importantly, for the Chinese regime's strategic objective in pursuing this activity. In the first large-scale empirical analysis of this operation, we show how to identify the secretive authors of these posts, the posts written by them, and their content. We estimate that the government fabricates and posts about 448 million social media comments a year. In contrast to prior claims, we show that the Chinese regime's strategy is to avoid arguing with skeptics of the party and the government, and to not even discuss controversial issues. We show that the goal of this massive secretive operation is instead to distract the public and change the subject, as most of these posts involve cheerleading for China, the revolutionary history of the Communist Party, or other symbols of the regime. We discuss how these results fit with what is known about the Chinese censorship program and suggest how they may change our broader theoretical understanding of "common knowledge" and information control in authoritarian regimes.

PSG accused of having bought,
via an agency, a "digital
army" on the networks to
attack players and media

10/12/2022, 5:41:48 PM



Mediapart reveals this Wednesday that the Parisian club would have paid an agency to create several accounts on social networks in order to



PSG accused of having bought, via an agency, a "digital army" on the networks to attack players and media

10/12/2022, 5:41:48 PM



Mediapart reveals this Wednesday that the Parisian club would have paid an agency to create several accounts on social networks in order to



Members of Google's ethical AI team reportedly complained about harassment and bias years before being fired



Martin Coulter

April 22, 2021 · 4 min read



Timnit Gebru was ousted from Google in December 2020. Kimberly White/Getty Images

- Google's former lead AI ethics experts reportedly raised complaints of harassment years before being fired.
- Timnit Gebru and Margaret Mitchell flagged bullying and misconduct in 2018, Bloomberg reported.
- Google said some of the accounts were inaccurate and that it investigated harassment allegations thoroughly.
- [See more stories on Insider's business page.](#)

Conclusion



jarran29 6 hours ago

Même si le propos semble sensé et raisonnable, c'est compliqué de remettre en cause le système scientifique ainsi...

Car au final, cette vidéo déclenche des émotions, et on ne juge pas sur cette base. Et comme le sujet est ultra-technique, on ne peut pas juger du fond et on doit faire faire confiance...

Je suis donc tiraillé entre ma misanthropie, ma colère et ma prudence... :/

Show less



jarran29 6 hours ago

Même si le propos semble sensé et raisonnable, c'est compliqué de remettre en cause le système scientifique ainsi...

Car au final, cette vidéo déclenche des émotions, et on ne juge pas sur cette base. Et comme le sujet est ultra-technique, on ne peut pas juger du fond et on doit faire faire confiance...

Je suis donc tiraillé entre ma misanthropie, ma colère et ma prudence... :/

Show less

- The **science community** must **urgently**:
1. Correct widespread **misinformation**.
 2. Adopt **realistic** working assumptions
(or clearly acknowledge that our work does not apply to the most impactful AIs).

SoK: On the Impossible Security of Very Large Foundation Models

El-Mahdi El-Mhamdi
École 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 Nguyễn Hoang
Association Tournesol,
Switzerland
lh@tournesol.app

Rafaël 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*, aim 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 [151], 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 [23]). Such models also excel in few-shot learning [23], which has motivated their wide use as pre-trained “foundation” (or “base”) models, to be *fine-tuned* to

any task of interest [35], [34], [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 [41]. 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 discourses¹. Military agencies, private companies and even universities, are now all racing for ever more impressive performance [29], [63].

However, numerous voices have raised serious concerns about the rushed deployment of such technologies [87]. These concerns are well illustrated by the anti-Muslim bias of OpenAI’s (deployed and commercialized) GPT-3 *foundation model* [23]. 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 own *foundation model*, generated insults against Facebook’s CEO Mark Zuckerberg [205].

¹Politico 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” [18]. This implicitly calls for racing to build ever larger *foundation models*.

40
10
6
comparisons
contributors

1,600+ scientific citations!

Science4All (english)
Google poisoned the science community...

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

Lê-Nguyễn Hoang^{1,2}, Louis Faucon², Aidan Jungo², Sergei Volodin², Dalia Papuc^{1,2}, Orfeas Liossatos^{1,2}, Ben Crulis³, Mariame Tighanimine^{2,4}, Isabela Constantin², Anastasiia Kucherenko^{1,2}, Alexandre Maurer^{2,5}, Felix Grimberg^{1,2}, Vlad Nitu^{2,6}, Chris Vossen², Sébastien Rouault^{1,2}, and El-Mahdi El-Mhamdi^{2,7}

¹IC, EPFL, Switzerland

²Tournesol Association, Switzerland

³University of Tours, France

⁴LISE, CNAM-CNRS, France

⁵UM6P, Benguerir, Morocco

⁶CNRS, INSA Lyon, France

⁷École 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.

Robust Sparse Voting

Youssef Allouah¹, Rachid Guerraoui¹, Lê-Nguyễn Hoang¹, and Oscar Villemaud¹

¹IC, 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.

Tournesol: Permissionless Collaborative Algorithmic Governance with Security Guarantees

Anonymous Author(s)

Submission Id: ???

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.

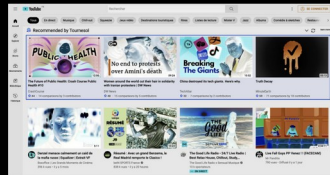


Figure 1: Our browser extension provides Tournesol's recommendations directly on the users' YouTube home page.

Perhaps most importantly, every day, social media's 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,