

2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML 2024)

**Toronto, Ontario, Canada
9 – 11 April 2024**



**IEEE Catalog Number: CFP24BT6-POD
ISBN: 979-8-3503-4951-1**

**Copyright © 2024 by the Institute of Electrical and Electronics Engineers, Inc.
All Rights Reserved**

Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

For other copying, reprint or republication permission, write to IEEE Copyrights Manager, IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854. All rights reserved.

****** This is a print representation of what appears in the IEEE Digital Library. Some format issues inherent in the e-media version may also appear in this print version.***

IEEE Catalog Number:	CFP24BT6-POD
ISBN (Print-On-Demand):	979-8-3503-4951-1
ISBN (Online):	979-8-3503-4950-4

Additional Copies of This Publication Are Available From:

Curran Associates, Inc
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: (845) 758-0400
Fax: (845) 758-2633
E-mail: curran@proceedings.com
Web: www.proceedings.com

CURRAN ASSOCIATES INC.
proceedings
.com

2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML) **SaTML 2024**

Table of Contents

Message from the Program Chairs	x
Organizing Committee	xi
Program Committee	xii
Steering Committee	xv

Session A: Privacy

Probabilistic Dataset Reconstruction from Interpretable Models	1
<i>Julien Ferry (LAAS-CNRS, Université de Toulouse, CNRS, France), Ulrich Aïvodji (École de Technologie Supérieure, Canada), Sébastien Gambs (Université du Québec à Montréal, Canada), Marie-José Huguet (LAAS-CNRS, Université de Toulouse, CNRS, INSA, France), and Mohamed Siala (LAAS-CNRS, Université de Toulouse, CNRS, INSA, France)</i>	
Shake to Leak: Fine-tuning Diffusion Models Can Amplify the Generative Privacy Risk	18
<i>Zhangheng Li (University of Texas at Austin), Junyuan Hong (University of Texas at Austin), Bo Li (University of Illinois Urbana-Champaign; University of Chicago), and Zhangyang Wang (University of Texas at Austin)</i>	
Improved Differentially Private Regression via Gradient Boosting	33
<i>Shuai Tang (Amazon Web Services), Sergul Aydore (Amazon Web Services), Michael Kearns (University of Pennsylvania; Amazon Web Services), Saeyoung Rho (Columbia University), Aaron Roth (University of Pennsylvania; Amazon Web Services), Yichen Wang (Amazon Web Services), Yu-Xiang Wang (University of California, Santa Barbara; Amazon Web Services), and Zhiwei Steven Wu (Carnegie Mellon University; Amazon Web Services)</i>	
SoK: A Review of Differentially Private Linear Models For High Dimensional Data	57
<i>Amol Khanna (Booz Allen Hamilton, USA), Edward Raff (Booz Allen Hamilton; University of Maryland, Baltimore County, USA), and Nathan Inkawhich (Air Force Research Laboratory, USA)</i>	
Concentrated Differential Privacy for Bandits	78
<i>Achraf Azize (Univ. Lille, Inria, CNRS, Centrale Lille, CRIStAL, France) and Debabrota Basu (Univ. Lille, Inria, CNRS, Centrale Lille, CRIStAL, France)</i>	

PILLAR: How to Make Semi-Private Learning More Effective	110
<i>Francesco Pinto (University of Oxford, England), Yaxi Hu (Max Planck Institute for Intelligent Systems, Germany), Fanny Yang (ETH Zürich, Switzerland), and Amartya Sanyal (ETH Zürich, Switzerland; Max Planck Institute for Intelligent Systems, Germany)</i>	

Session B: Fairness

Fair Federated Learning via Bounded Group Loss	140
<i>Shengyuan Hu (Carnegie Mellon University), Zhiwei Steven Wu (Carnegie Mellon University), and Virginia Smith (Carnegie Mellon University)</i>	
Estimating and Implementing Conventional Fairness Metrics With Probabilistic Protected Features	161
<i>Hadi Elzayn (Stanford University), Emily Black (Barnard College), Patrick Vossler (Stanford University), Nathanael Jo (Massachusetts Institute of Technology), Jacob Goldin (University of Chicago), and Daniel E. Ho (Stanford University)</i>	

Session C: Verifying/Certifying Properties of ML Systems

Evaluating Superhuman Models with Consistency Checks	194
<i>Lukas Fluri (ETH Zurich, Switzerland), Daniel Paleka (ETH Zurich, Switzerland), and Florian Tramèr (ETH Zurich, Switzerland)</i>	
Certiably Robust Reinforcement Learning through Model-Based Abstract Interpretation	233
<i>Chenxi Yang (The University of Texas at Austin, United States), Greg Anderson (Reed College, United States), and Swarat Chaudhuri (The University of Texas at Austin, United States)</i>	
Fast Certification of Vision-Language Models Using Incremental Randomized Smoothing	252
<i>Ashutosh Nirala (Iowa State University), Ameya Joshi (New York University), Soumik Sarkar (Iowa State University), and Chinmay Hegde (New York University)</i>	

Session D: Training-Time Integrity (backdoor Attacks, etc.)

Backdoor Attack on Unpaired Medical Image-Text Foundation Models: A Pilot Study on MedCLIP.....	272
<i>Ruinan Jin (The University of British Columbia), Chun-Yin Huang (The University of British Columbia), Chenyu You (Yale University), and Xiaoxiao Li (The University of British Columbia)</i>	
REStore: Exploring a Black-Box Defense Against DNN Backdoors using Rare Event Simulation	286
<i>Quentin Le Roux (Thales DIS & INRIA, France), Kassem Kallas (INRIA, France), and Teddy Furon (INRIA, France)</i>	
EdgePruner: Poisoned Edge Pruning in Graph Contrastive Learning	309
<i>Hiroya Kato (KDDI Research, Inc., Japan), Kento Hasegawa (KDDI Research, Inc., Japan), Seira Hidano (KDDI Research, Inc., Japan), and Kazuhide Fukushima (KDDI Research, Inc., Japan)</i>	

Indiscriminate Data Poisoning Attacks on Pre-Trained Feature Extractors	327
<i>Yiwei Lu (University of Waterloo), Matthew Y.R. Yang (University of Waterloo), Gautam Kamath (University of Waterloo), and Yaoliang Yu (University of Waterloo)</i>	
ImpNet: Imperceptible and Blackbox-Undetectable Backdoors in Compiled Neural Networks	344
<i>Eleanor Clifford (University of Cambridge), Ilia Shumailov (University of Oxford), Yiren Zhao (Imperial College London), Ross Anderson (University of Cambridge), and Robert Mullins (University of Cambridge)</i>	
The Devil's Advocate: Shattering the Illusion of Unexploitable Data using Diffusion Models.....	358
<i>Hadi M. Dolatabadi (University of Melbourne, Australia), Sarah Erfani (University of Melbourne, Australia), and Christopher Leckie (University of Melbourne, Australia)</i>	

Session E: Inference-Time Integrity (adversarial Examples, etc.)

SoK: Pitfalls in Evaluating Black-Box Attacks	387
<i>Fnu Suya (University of Maryland College Park, USA), Anshuman Suri (University of Virginia, USA), Tingwei Zhang (Cornell University, USA), Jingtao Hong (Columbia University, USA), Yuan Tian (University of California Los Angeles, USA), and David Evans (University of Virginia, USA)</i>	
Evading Black-box Classifiers Without Breaking Eggs	408
<i>Edoardo Debenedetti (ETH Zurich, Switzerland), Nicholas Carlini (Google DeepMind, United States), and Florian Tramèr (ETH Zurich, Switzerland)</i>	
Segment (Almost) Nothing: Prompt-Agnostic Adversarial Attacks on Segmentation Models	425
<i>Francesco Croce (University of Tübingen, Tübingen AI Center, Germany) and Matthias Hein (University of Tübingen, Tübingen AI Center, Germany)</i>	

Session F: Collaborative Learning

Improving Privacy-Preserving Vertical Federated Learning by Efficient Communication with ADMM	443
<i>Chulin Xie (University of Illinois Urbana-Champaign), Pin-Yu Chen (International Business Machines), Qinbin Li (University of California, Berkeley), Arash Nourian (Amazon Web Services), Ce Zhang (University of Chicago), and Bo Li (University of Chicago)</i>	
Differentially Private Multi-Site Treatment Effect Estimation	472
<i>Tatsuki Koga (University of California, San Diego, USA), Kamalika Chaudhuri (University of California, San Diego, USA), and David Page (Duke University, USA)</i>	

ScionFL: Efficient and Robust Secure Quantized Aggregation	490
<i>Yaniv Ben-Itzhak (VMware Research Group), Helen Möllering (Technical University of Darmstadt), Benny Pinkas (Aptos Labs and Bar-Ilan University), Thomas Schneider (Technical University of Darmstadt), Ajith Suresh (Technology Innovation Institute), Oleksandr Tkachenko (DFINITY Foundation), Shay Vargaftik (VMware Research Group), Christian Weinert (Royal Holloway, University of London), Hossein Yalame (Technical University of Darmstadt), and Avishay Yanai (VMware Research Group)</i>	
Differentially Private Heavy Hitter Detection using Federated Analytics	512
<i>Karan Chadha (Stanford University, USA), Junye Chen (Apple, USA), John Duchi (Stanford University, Apple, USA), Vitaly Feldman (Apple, USA), Hanieh Hashemi (Apple, USA), Omid Javidbakht (Apple, USA), Audra McMillan (Apple, USA), and Kunal Talwar (Apple, USA)</i>	
Olympia: A Simulation Framework for Evaluating the Concrete Scalability of Secure Aggregation Protocols	534
<i>Ivoline C. Ngong (University of Vermont, USA), Nicholas Gibson (University of Vermont, USA), and Joseph P. Near (University of Vermont, USA)</i>	

Session G: Patching

Model Reprogramming Outperforms Fine-Tuning on Out-of-Distribution Data in Text-Image Encoders	552
<i>Andrew Geng (University of Wisconsin-Madison, United States) and Pin-Yu Chen (IBM Research, United States)</i>	
Data Redaction from Conditional Generative Models	569
<i>Zhifeng Kong (University of California San Diego, USA) and Kamalika Chaudhuri (University of California San Diego, USA)</i>	
Towards Scalable and Robust Model Versioning	592
<i>Wenxin Ding (University of Chicago, United States), Arjun Nitin Bhagoji (University of Chicago, United States), Ben Y. Zhao (University of Chicago, United States), and Haitao Zheng (University of Chicago, United States)</i>	

Session H: Auditing

AI Auditing: The Broken Bus on the Road to AI Accountability	612
<i>Abeba Birhane (Mozilla Foundation and Trinity College Dublin, Ireland), Ryan Steed (Carnegie Mellon University Pittsburgh, USA), Victor Ojewale (Brown University Providence, USA), Briana Vecchione (Data & Society New York City, USA), and Inioluwa Deborah Raji (Mozilla Foundation and University of California, Berkeley, USA)</i>	
Under Manipulations, are Some AI Models Harder to Audit?	644
<i>Augustin Godinot (Univ Rennes, Inria, CNRS, IRISA, PEReN, France), Erwan Le Merrer (Univ Rennes, Inria, CNRS, IRISA, France), Gilles Trédan (LAAS/CNRS, France), Camilla Penzo (PEReN, France), and François Taïani (Univ Rennes, Inria, CNRS, IRISA, France)</i>	

Session I: Forensic Analysis of ML Systems

Unifying Corroborative and Contributive Attributions in Large Language Models	665
<i>Theodora Worledge (Stanford University), Judy Hanwen Shen (Stanford University), Nicole Meister (Stanford University), Caleb Winston (Stanford University), and Carlos Guestrin (Stanford University)</i>	
CodeLMSec Benchmark: Systematically Evaluating and Finding Security Vulnerabilities in Black-Box Code Language Models	684
<i>Hossein Hajipour (CISPA Helmholtz Center for Information Security, Germany), Keno Hassler (CISPA Helmholtz Center for Information Security, Germany), Thorsten Holz (CISPA Helmholtz Center for Information Security, Germany), Lea Schönherr (CISPA Helmholtz Center for Information Security, Germany), and Mario Fritz (CISPA Helmholtz Center for Information Security, Germany)</i>	
Navigating the Structured What-If Spaces: Counterfactual Generation via Structured Diffusion	710
<i>Nishtha Madaan (Indian Institute of Delhi, India) and Srikanta Bedathur (Indian Institute of Delhi, India)</i>	
Understanding, Uncovering, and Mitigating the Causes of Inference Slowdown for Language Models	723
<i>Kamala Varma (University of Maryland, College Park), Arda Numanoğlu (Middle East Technical University, Ankara), Yigitcan Kaya (University of California, Santa Barbara), and Tudor Dumitraş (University of Maryland, College Park)</i>	
Author Index	741