OpenDP Privacy Attacks & Auditing Working Group

Privacy Attacks Respository

URL	BibTex (Please add a bibtex entry for this paper to facilitate writing our summary document)	Authors	Title	Short Description	Type of Data	Type of Release	Threat Model	Research Type	Links to Artifacts	Comments	Submitter (your name, affiliation)
https://dl.acm.org/doi/10.1145/773153.773173	@inproceedings{10.1145/773153.773173, author = {Dinur, Irit and Nissim, Kobbi}, title = {Revealing information while preserving privacy}, year = {2003}, }	Irit Dinur and Kobbi Nissim	Revealing Information While Preserving Privacy	Seminal paper on the theory of data reconstruction attacks.	Tabular	Linear-Queries	Reconstruction	Theoretical			Jon Ullman, Northeastern University
https://arxiv.org/abs/1810.05692	@article{cohen2018linear, title={Linear program reconstruction in practice}, author={Cohen, Aloni and Nissim, Kobbi}, year={2018} }	Aloni Cohen and Kobbi Nissim	Linear program reconstruction in practice.	Implemented linear reconstruction attacks against a production private query system called Diffix	Tabular	Linear-Queries	Reconstruction	Applications			Jon Ullman, Northeastern University
https://arxiv.org/abs/2405.10994	@article{annamalai2024you, title={" What do you want from theory alone?" Experimenting with Tight Auditing of Differentially Private Synthetic Data Generation}, author={Annamalai, Meenatchi Sundaram Muthu Selva and Ganev, Georgi and De Cristofaro, Emiliano}, journal={USENIX Security}, year={2024} }	Meenatchi Sundaram Muthu Selva Annamalai, Georgi Ganev, Emiliano De Cristofaro	"What do you want from theory alone?" Experimenting with Tight Auditing of Differentially Private Synthetic Data Generation	Audits six implementations of DP synthetic data generative models using different datasets and threat models and finds that commonly used black-box MIAs are severely limited in power, yielding remarkably loose empirical privacy estimates. Considers MIAs in stronger threat models, i.e., passive and active white-box, using both existing and newly proposed attacks.	Tabular	Generative-Model	Membership-Inference	Empirical	https://github.com/spalabucr/synth-audit		Georgi Ganev, UCL
https://proceedings.neurips.cc/paper/2020/file/fc4ddc15f9f4b4b06ef7844d6bb53abf-Paper.pdf	@article{jagielski2020auditing, title={Auditing differentially private machine learning: How private is private {SGD}?}, author={Jagielski, Matthew and Ullman, Jonathan and Oprea, Alina}, journal={Advances in Neural Information Processing Systems}, }	Matthew Jagielski, Jonathan Ullman, Alina Oprea	Auditing Differentially Private Machine Learning: How Private is Private SGD?	Connects the success rate of a membership inferency adversary to a lower bound on the privacy loss of the underlying DP mechanism.	Image	Predictive-Model	Membership-Inference	Empirical			Tudor Cebere, Inria
https://www.pnas.org/doi/10.1073/pnas.2218605120	@article{dick2023confidence, title={Confidence-ranked reconstruction of census microdata from published statistics}, author={Dick, Travis and Dwork, Cynthia and Kearns, Michael and Liu, Terrance and Roth, Aaron and Vietri, Giuseppe and Wu, Zhiwei Steven}, journal={Proceedings of the National Academy of Sciences}, volume={120}, number={8}, pages={e2218605120}, year={2023}, publisher={National Acad Sciences} }	Travis Dick, Cynthia Dwork, Michael Kearns, Terrance Liu, Aaron Roth, Giuseppe Vietri, and Zhiwei Steven Wu	Confidence-ranked reconstruction of census microdata frompublished statistics	Ranking rows in reconstructed microdata by how confident the adversary is that are in the true dataset.	Tabular	Linear-Queries	Membership-Inference	Empirical		I didn't want to give it it's own row but linking a rebuttal paper which is interesting in that it is indicative of arguments people make against reconstruction attacks: https://arxiv.org/abs/2311.03171	Audra McMillan, Apple
https://dl.acm.org/doi/abs/10.1145/3548606.3560581	@inproceedings{cretu2022querysnout, title={QuerySnout: Automating the discovery of attribute inference attacks against query-based systems}, author={Cretu, Ana-Maria and Houssiau, Florimond and Cully, Antoine and de Montjoye, Yves-Alexandre}, booktitle={Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security}, pages={623--637}, year={2022} }	Ana-Maria Cretu, Florimond Houssiau, Antoine Cully, Yves-Alexandre de Montjoye	QuerySnout: Automating the Discovery of Attribute Inference Attacks against Query-Based Systems	A method to automatically discover attacks against interactive query systems	Tabular	Linear-Queries	Attribute-Inference	Empirical			Ana-Maria Cretu, EPFL
https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_05B-5_Pyrgelis_paper.pdf	@article{pyrgelis2017knock, title={Knock knock, who's there? Membership inference on aggregate location data}, author={Pyrgelis, Apostolos and Troncoso, Carmela and De Cristofaro, Emiliano}, journal={arXiv preprint arXiv:1708.06145}, year={2017} }	Apostolos Pyrgelis, Carmela Troncoso, Emiliano de Cristofaro	Knock Knock, Who's There? Membership Inference on Aggregate Location Data	Membership inference attacks against location aggregates	Tabular	Linear-Queries	Membership-Inference	Empirical			Ana-Maria Cretu, EPFL Yves-Alexandre de Montjoye, Imperial College
https://arxiv.org/abs/2301.10053	@inproceedings{annamalai2023linear, title={A linear reconstruction approach for attribute inference attacks against synthetic data}, author={Annamalai, Meenatchi Sundaram Muthu Selva and Gadotti, Andrea and Rocher, Luc}, booktitle={Usenix Security}, year={2024} }	Meenatchi Sundaram Muthu Selva Annamalai, Andrea Gadotti, Luc Rocher	A linear reconstruction approach for attribute inference attacks against synthetic data	Introduces a new attribute inference attack against synthetic data based on linear reconstruction methods for aggregate statistics, which target all records in the dataset, not only outliers.	Tabular	Generative-Model	Attribute inference	Empirical	https://github.com/synthetic-society/recon-synth		Georgi Ganev, UCL
https://arxiv.org/pdf/2206.10469.pdf	@article{carlini2022privacy, title={The privacy onion effect: Memorization is relative}, author={Carlini, Nicholas and Jagielski, Matthew and Zhang, Chiyuan and Papernot, Nicolas and Terzis, Andreas and Tramer, Florian}, journal={Advances in Neural Information Processing Systems}, volume={35}, pages={13263--13276}, year={2022} }	Carlini et al.	The Privacy Onion Effect: Memorization is Relative	Removing vulnerable records makes other records vulnerable	Tabular	Predictive-Model	Membership-Inference	Empirical			Yves-Alexandre de Montjoye, Imperial College
https://arxiv.org/abs/2112.03570	@inproceedings{carlini2022membership, title={Membership inference attacks from first principles}, author={Carlini, Nicholas and Chien, Steve and Nasr, Milad and Song, Shuang and Terzis, Andreas and Tramer, Florian}, booktitle={2022 IEEE Symposium on Security and Privacy (SP)}, pages={1897--1914}, year={2022}, organization={IEEE} }	Carlini et al.	Membership Inference Attacks From First Principles	Takes a step back + new attack (LIRA)	Tabular	Predictive-Model	Membership-Inference	Theoretical			Yves-Alexandre de Montjoye, Imperial College
https://arxiv.org/abs/2003.14053	@article{geiping2020inverting, title={Inverting gradients-how easy is it to break privacy in federated learning?}, author={Geiping, Jonas and Bauermeister, Hartmut and Dr{\"o}ge, Hannah and Moeller, Michael}, journal={Advances in neural information processing systems}, volume={33}, pages={16937--16947}, year={2020} }	Geiping et al.	Inverting Gradients -- How easy is it to break privacy in federated learning?	Gradient inversion attack (reconstruct data point from gradient) - application to federated learning	Image	Predictive-Model	Reconstruction	Empirical	https://github.com/JonasGeiping/breaching		Aurélien Bellet, Inria
https://archive.dimacs.rutgers.edu/~graham/pubs/papers/empiricalpriv.pdf	@inproceedings{cormode2013empirical, title={Empirical privacy and empirical utility of anonymized data}, author={Cormode, Graham and Procopiuc, Cecilia M and Shen, Entong and Srivastava, Divesh and Yu, Ting}, booktitle={2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)}, pages={77--82}, year={2013}, organization={IEEE} }	Cormode, G., Procopiuc, C.M., Shen, E., Srivastava, D. and Yu, T	Empirical privacy and empirical utility of anonymized data.	Naive Bayes attacks against simple workloads of statistics	Tabular	Linear-Queries	Reconstruction	Empirical			James Honaker, Anonym
https://projects.iq.harvard.edu/files/privacytools/files/pdf_02.pdf	@article{dwork2017exposed, title={Exposed! a survey of attacks on private data}, author={Dwork, Cynthia and Smith, Adam and Steinke journal={Annual Review of Statistics and Its Applicati volume={4}, number={1}, pages={61--84}, year={2017}, publisher={Annual Reviews} }	Dwork, C., Smith, A., Steinke, T. and Ullman,	Exposed! a survey of attacks on private data.	Survey of privacy attacks	Tabular			Theoretical			James Honaker, Anonym
https://proceedings.neurips.cc/paper_files/paper/2023/file/9a6f6e0d6781d1cb8689192408946d73-Paper-Conference.pdf	@article{steinke2024privacy, title={Privacy auditing with one (1) training run}, author={Steinke, Thomas and Nasr, Milad and Jagielsk journal={Advances in Neural Information Processing S volume={36}, year={2024} }	Steinke, T., Nasr, M. and Jagielski, M.	Privacy Auditing with One (1) Training Run.				Membership-Inference	Empirical			James Honaker, Anonym
https://arxiv.org/pdf/1610.05820.pdf	@inproceedings{shokri2017membership, title={Membership inference attacks against machine l author={Shokri, Reza and Stronati, Marco and Song, C booktitle={2017 IEEE symposium on security and priva pages={3--18}, year={2017}, organization={IEEE} }	Shokri, R., Stronati, M., Song, C. and Shmatik	Membership inference attacks against machine learning models.	The original membership inference based on shadow models			Membership-Inference	Empirical			James Honaker, Anonym
https://arxiv.org/abs/1512.00327		Wagner, I. and Eckhoff, D.	Technical privacy metrics: a systematic survey.	Survey of privacy-loss metrics	Tabular	Linear-Queries	Information Leakage	Theoretical			James Honaker, Anonym
https://petsymposium.org/popets/2023/popets-2023-0055.php	@misc {giomi2022unifiedframeworkquantifyingprivacy, title={A Unified Framework for Quantifying Privacy Risk in Synthetic Data}, author={Matteo Giomi and Franziska Boenisch and Christoph Wehmeyer and Borbála Tasnádi}, year={2022}, eprint={2211.10459}, archivePrefix={arXiv}, primaryClass={cs.CR}, url={https://arxiv.org/abs/2211.10459}, }	Giomi, M., Boenisch, F., Wehmeyer, C., and Tasnádi, B.	A Unified Framework for Quantifying Privacy Risk in Synthetic Data	Adversarial evaluation of singling out, linkability, and inference risk in tabular synthetic data	Tabular			Empirical	https://github.com/statice/anonymeter		Matteo Giomi, Anonos
https://arxiv.org/abs/2211.10459	@inproceedings{giomi2022unified, title={A unified framework for quantifying privacy risk in synthetic data}, author={Giomi, Matteo and Boenisch, Franziska and Wehmeyer, Christoph and Tasn{\'a}di, Borb{\'a}la}, booktitle={PETs}, year={2023}, }	Matteo Giomi, Franziska Boenisch, Christoph Wehmeyer, Borbála Tasnádi	A Unified Framework for Quantifying Privacy Risk in Synthetic Data	Present Anonymeter, a statistical framework to jointly quantify different types of privacy risks in synthetic tabular datasets. Equips this framework with attack-based evaluations for the singling out, linkability, and inference risks, the three key indicators of factual anonymization according to the GDPR.	Tabular	Generative-Model	Membership-Inference	Empirical			Georgi Ganev, UCL

OpenDP Privacy Attacks & Auditing Working Group

Privacy Attacks Respository

URL	BibTex (Please add a bibtex entry for this paper to facilitate writing our summary document)	Authors	Title	Short Description	Type of Data	Type of Release	Threat Model	Research Type	Links to Artifacts	Comments	Submitter (your name, affiliation)
https://arxiv.org/abs/2310.16789	@article{shi2023detecting, title={Detecting pretraining data from large language models}, author={Shi, Weijia and Ajith, Anirudh and Xia, Mengzhou and Huang, Yangsibo and Liu, Daogao and Blevins, Terra and Chen, Danqi and Zettlemoyer, Luke}, journal={arXiv preprint arXiv:2310.16789}, year={2023} }	Weijia Shi et al.	Detecting Pretraining Data from Large Language Models	introduces Min-K prob attack: Membership Inference atttack agaisnt LLMs using average of lowest k probable tokens in the target sequence.	Text	Generative-Model	Membership-Inference	Applications	https://swj0419.github.io/detect-pretrain.github.io/		Hamid Mozaffari, Oracle Labs
https://arxiv.org/abs/2402.07841	@misc{duan2024membership, title={Do Membership Inference Attacks Work on Large Language Models?}, author={Michael Duan and Anshuman Suri and Niloofar Mireshghallah and Sewon Min and Weijia Shi and Luke Zettlemoyer and Yulia Tsvetkov and Yejin Choi and David Evans and Hannaneh Hajishirzi}, year={2024}, eprint={2402.07841}, archivePrefix={arXiv}, primaryClass={id='cs.CL' full_name='Computation and Language' is_active=True alt_name='cmp-lg' in_archive='cs' is_general=False description='Covers natural language processing. Roughly includes material in ACM Subject Class I.2.7. Note that work on artificial languages (programming languages, logics, formal systems) that does not explicitly address natural-language issues broadly construed (natural-language processing, computational linguistics, speech, text retrieval, etc.) is not appropriate for this area.'} }	Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, Hannaneh Hajishirzi	Do Membership Inference Attacks Work on Large Language Models?	Examining MIA attacks on LLM, proposing a github repo with set of standard attacks	Text	Generative-Model	Membership-Inference	Empirical	https://github.com/iamgroot42/mimir		Daniil Filienko, University of Washington
https://arxiv.org/abs/2301.13188	@article{carlini2023extracting, title={Extracting training data from diffusion models}, author={Carlini, Nicholas and Hayes, Jamie and Nasr, Milad and Jagielski, Matthew and Sehwag, Vikash and Tram{\`e}r, Florian and Balle, Borja and Ippolito, Daphne and Wallace, Eric}, journal={USENIX Security}, year={2023} }	Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, Eric Wallace	Extracting training data from diffusion models	Show that diffusion models memorize individual images from their training data and emit them at generation time. With a generate-and-filter pipeline, extracts over a thousand training examples from state-of-the-art models, ranging from photographs of individual people to trademarked company logos.	Image	Generative-Model	Data-Extraction	Empirical			Georgi Ganev, UCL
https://arxiv.org/abs/2012.07805	@inproceedings{carlini2021extracting, title={Extracting training data from large language models}, author={Carlini, Nicholas and Tramer, Florian and Wallace, Eric and Jagielski, Matthew and Herbert-Voss, Ariel and Lee, Katherine and Roberts, Adam and Brown, Tom and Song, Dawn and Erlingsson, Ulfar and others}, year={2021} }	Nicolas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel	Extracting training data from large language models.	Showed how to prompt models like GPT2 to reveal specific training examples	Text	Generative-Model	Data-Extraction	Applications			Jon Ullman, Northeastern University
https://arxiv.org/pdf/2211.06550.pdf	@article{houssiau2022tapas, title={Tapas: a toolbox for adversarial privacy auditing of synthetic data}, author={Houssiau, Florimond and Jordon, James and Cohen, Samuel N and Daniel, Owen and Elliott, Andrew and Geddes, James and Mole, Callum and Rangel-Smith, Camila and Szpruch, Lukasz}, journal={arXiv preprint arXiv:2211.06550}, year={2022} }	Florimond Houssiau, James Jordon, Samuel N. Cohen, Owen Daniel, Andrew Elliott, James Geddes, Callum Mole, Camila Rangel-Smith, Lukasz Szpruch	Tapas: Toolbox for Adversarial Privacy Auditing of Synthetic Data	A system of classification of MIA attacks presented as part of a toolbox of attacks to evaluate synthetic data privacy under a wide range of scenarios	Tabular						Daniil Filienko, University of Washington
https://arxiv.org/abs/1909.03935	@inproceedings{chen2020gan, title={Gan-leaks: a taxonomy of membership inference attacks against generative models}, author={Chen, Dingfan and Yu, Ning and Zhang, Yang and Fritz, Mario}, booktitle={ACM CCS}, year={2020}, }	Dingfan Chen, Ning Yu, Yang Zhang, Mario Fritz	Gan-leaks: a taxonomy of membership inference attacks against generative models	Presents a taxonomy of membership inference attacks, encompassing not only existing attacks but also our novel ones. Moreover, provides a theoretically grounded attack calibration technique, which consistently boosts the attack performance in all cases, across different attack settings, data modalities, and training configurations.	Image	Generative-Model	Membership-Inference	Empirical			Georgi Ganev, UCL
https://arxiv.org/abs/1705.07663	@inproceedings{hayes2019logan, title={{LOGAN}: membership inference attacks against generative models}, author={Hayes, Jamie and Melis, Luca and Danezis, George and De Cristofaro, Emiliano}, booktitle={PoPETs}, year={2019}, }	Jamie Hayes, Luca Melis, George Danezis, Emiliano De Cristofaro	LOGAN: Membership Inference Attacks Against Generative Models	Presents the first membership inference attacks against generative models (GANs): given a data point, the adversary determines whether or not it was used to train the model.	Image	Generative-Model	Membership-Inference	Empirical			Georgi Ganev, UCL
https://arxiv.org/pdf/2112.05307.pdf	@inproceedings{jin2022we, title={Are we there yet? timing and floating-point attacks on differential privacy systems}, author={Jin, Jiankai and McMurtry, Eleanor and Rubinstein, Benjamin IP and Ohrimenko, Olga}, booktitle={2022 IEEE Symposium on Security and Privacy (SP)}, pages={473--488}, year={2022}, organization={IEEE} }	Jin, McMurtry, Rubinstein, Ohrimenko	Are We There Yet? Timing and Floating-Point Attacks on Differential Privacy Systems	Attacks on DP implementations that use floating-point arithmetic; timing side channels on discrete samplers			Reconstruction	Applications			Zachary Ratliff, Harvard + OpenDP
https://arxiv.org/abs/2305.18462	@article{mattern2023membership, title={Membership inference attacks against language models via neighbourhood comparison}, author={Mattern, Justus and Mireshghallah, Fatemehsadat and Jin, Zhijing and Sch{\"o}lkopf, Bernhard and Sachan, Mrinmaya and Berg-Kirkpatrick, Taylor}, journal={arXiv preprint arXiv:2305.18462}, year={2023} }	Justus Mattern et al.	Membership Inference Attacks against Language Models via Neighbourhood Comparison	Memership inference attack agaisnt LLMs by generating neighbours which are perturbed text using a mask model like BERT	Text	Generative-Model	Membership-Inference	Applications	https://github.com/mireshghallah/neighborhood-curvature-mia		Hamid Mozaffari, Oracle Labs
https://arxiv.org/abs/2312.03262	@inproceedings{zarifzadeh2024low, title={Low-Cost High-Power Membership Inference Attacks}, author={Zarifzadeh, Sajjad and Liu, Philippe and Shokri, Reza}, booktitle={Forty-first International Conference on Machine Learning}, year={2024} }	Sajjad Zarifzadeh, Philippe Liu, Reza Shokri	Low-Cost High-Power Membership Inference Attacks	Efficient MIA with shadow models and auxiliary reference data. Outperfroms LiRA in cases where one has plenty of reference data and almost no reference models	Tabular	Predictive-Model	Membership-Inference	Empirical			Luca Melis, Meta
https://arxiv.org/abs/2302.12580	@article{van2023membership, title={Membership inference attacks against synthetic data through overfitting detection}, author={van Breugel, Boris and Sun, Hao and Qian, Zhaozhi and van der Schaar, Mihaela}, journal={AISTATS}, year={2023} }	Boris van Breugel, Hao Sun, Zhaozhi Qian, Mihaela van der Schaar	Membership inference attacks against synthetic data through overfitting detection	Proposes DOMIAS, a density-based MIA model that aims to infer membership by targeting local overfitting of the generative model assuming the attacker has some knowledge of the underlying data distribution.	Tabular	Generative-Model	Membership-Inference	Empirical			Georgi Ganev, UCL
https://arxiv.org/abs/2111.09679	@inproceedings{ye2022enhanced, title={Enhanced membership inference attacks against machine learning models}, author={Ye, Jiayuan and Maddi, Aadyaa and Murakonda, Sasi Kumar and Bindschaedler, Vincent and Shokri, Reza}, booktitle={Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security}, pages={3093--3106}, year={2022} }	Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, Reza Shokri	Enhanced Membership Inference Attacks against Machine Learning Models	Framework for MIA based on approximate LRT. Theory behind general class of attacks + a few instantiations.	Tabular	Predictive-Model	Membership-Inference	Empirical			
https://arxiv.org/abs/2208.14933	@inproceedings{liu2022membership, title={Membership inference attacks by exploiting loss author={Liu, Yiyong and Zhao, Zhengyu and Backes, M booktitle={Proceedings of the 2022 ACM SIGSAC Cor pages={2085--2098}, year={2022} }	Yiyong Liu, Zhengyu Zhao, Michael Backes, Y	Membership inference attacks by exloiting loss trajectories	Membership inference attack based on knowledge distillation to form signals to attack via loss trajectories over multiple epochs.	Image	Predictive-Model	Membership-Inference	Empirical	https://github.com/DennisLiu2022/Membership-Inference-Attacks-by-Exploiting-Loss-Trajectory		Johan Östman, AI Sweden
https://arxiv.org/abs/2307.03694	@article{bertran2024scalable, title={Scalable membership inference attacks via quantile regression}, author={Bertran, Martin and Tang, Shuai and Roth, Aaron and Kearns, Michael and Morgenstern, Jamie H and Wu, Steven Z}, journal={Advances in Neural Information Processing Systems}, volume={36}, year={2024} }	Martin Betram, Shuai Tang, Michael Kearns, Jamie Morgenstern, Aaron Roth, Zhiwei Steven Wu	Scalable Membership Inference Attacks via Quantile Regression	Membership inference attack that does not require shadow models but only to train a single regression model to predict quantiles of the logits.	Image	Predictive-Model	Membership-Inference	Empirical			Johan Östman, AI Sweden
https://arxiv.org/abs/2007.14321	@inproceedings{choquette2021label, title={Label-only membership inference attacks}, author={Choquette-Choo, Christopher A and Tramer, Florian and Carlini, Nicholas and Papernot, Nicolas}, booktitle={International conference on machine learning}, pages={1964--1974}, year={2021}, organization={PMLR} }	Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini. Nicolas Papernot	Label-only membership inference attacks	Black-box membership inference attacl with label-only access. Signals are created by probing the mode with permutations around a given datapoint.	Image	Predictive-Model	Membership-Inference	Empirical			Johan Östman, AI Sweden
https://openreview.net/pdf?id=7WsivwyHrS	@inproceedings{wuyou, title={You Only Query Once: An Efficient Label-Only Membership Inference Attack}, author={Wu, Yutong and Qiu, Han and Guo, Shangwei and Li, Jiwei and Zhang, Tianwei}, booktitle={The Twelfth International Conference on Learning Representations} }	Yutong Wu, Han Qiu, Shangwei Guo, Jiwei Li, Tianwei Zhang	You only query once: an efficient label-only membership inference attack	Strategies to craft query examples to reduce the required number of queries	Image	Predictive-Model	Membership-Inference	Empirical	https://github.com/WU-YU-TONG/YOQO		Johan Östman, AI Sweden

OpenDP Privacy Attacks & Auditing Working Group

Privacy Attacks Respository

URL	BibTex (Please add a bibtex entry for this paper to facilitate writing our summary document)	Authors	Title	Short Description	Type of Data	Type of Release	Threat Model	Research Type	Links to Artifacts	Comments	Submitter (your name, affiliation)
https://openreview.net/forum?id=fwzUgo0FM9v	@article{fowl2021robbing, title={Robbing the fed: Directly obtaining private data in federated learning with modified models}, author={Fowl, Liam and Geiping, Jonas and Czaja, Wojtek and Goldblum, Micah and Goldstein, Tom}, journal={arXiv preprint arXiv:2110.13057}, year={2021} }	L Fowl, J Geiping, W Czaja, M Goldblum, T Goldstein	Robbing the Fed: Directly Obtaining Private Data in Federated Learning with Modified Models	Malicious adversary to perform exact reconstruction of the training data in FL	Image	Predictive-Model	Reconstruction	Empirical	https://github.com/JonasGeiping/breaching		Dmitrii Usynin, TUM/Imperial College London
https://arxiv.org/abs/2201.12675	@article{fowl2022decepticons, title={Decepticons: Corrupted transformers breach privacy in federated learning for language models}, author={Fowl, Liam and Geiping, Jonas and Reich, Steven and Wen, Yuxin and Czaja, Wojtek and Goldblum, Micah and Goldstein, Tom}, journal={arXiv preprint arXiv:2201.12675}, year={2022} }	Liam Fowl, Jonas Geiping, Steven Reich, Yuxin Wen, Wojtek Czaja, Micah Goldblum, Tom Goldstein	Decepticons: Corrupted Transformers Breach Privacy in Federated Learning for Language Models	Model inversion attacks using a malicious attacker in transformer-based FL settings	Text	Predictive-Model	Reconstruction	Empirical	https://github.com/JonasGeiping/breaching		Dmitrii Usynin, TUM/Imperial College London
https://arxiv.org/abs/2202.00580	@article{wen2022fishing, title={Fishing for user data in large-batch federated learning via gradient magnification}, author={Wen, Yuxin and Geiping, Jonas and Fowl, Liam and Goldblum, Micah and Goldstein, Tom}, journal={arXiv preprint arXiv:2202.00580}, year={2022} }	Yuxin Wen, Jonas Geiping, Liam Fowl, Micah Goldblum, Tom Goldstein	Fishing for User Data in Large-Batch Federated Learning via Gradient Magnification	Malicious adversary to perform exact reconstruction of the training data in FL, this time with (almost) arbitrarily large batch sizes	Image	Predictive-Model	Reconstruction	Empirical	https://github.com/JonasGeiping/breaching		Dmitrii Usynin, TUM/Imperial College London
https://arxiv.org/abs/2112.02918	@inproceedings{boenisch2023curious, title={When the curious abandon honesty: Federated learning is not private}, author={Boenisch, Franziska and Dziedzic, Adam and Schuster, Roei and Shamsabadi, Ali Shahin and Shumailov, Ilia and Papernot, Nicolas}, booktitle={2023 IEEE 8th European Symposium on Security and Privacy (EuroS\&P)}, pages={175--199}, year={2023}, organization={IEEE} }	Franziska Boenisch, Adam Dziedzic, Roei Schuster, Ali Shahin Shamsabadi, Ilia Shumailov, Nicolas Papernot	When the Curious Abandon Honesty: Federated Learning Is Not Private	Malicious FL input reconstruction using trap weights (model modification)	Image	Predictive-Model	Reconstruction	Empirical	https://github.com/JonasGeiping/breaching		Dmitrii Usynin, TUM/Imperial College London
https://openaccess.thecvf.com/content/CVPR2021/html/Yin_See_Through_Gradients_Image_Batch_Recovery_via_GradInversion_CVPR_2021_paper.html	@inproceedings{yin2021see, title={See through gradients: Image batch recovery via gradinversion}, author={Yin, Hongxu and Mallya, Arun and Vahdat, Arash and Alvarez, Jose M and Kautz, Jan and Molchanov, Pavlo}, booktitle={Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition}, pages={16337--16346}, year={2021} }	Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M. Alvarez, Jan Kautz, Pavlo Molchanov	See Through Gradients: Image Batch Recovery via GradInversion	First demonstration of large batch-size model inversion in FL	Image	Predictive-Model	Reconstruction	Empirical	https://github.com/JonasGeiping/breaching		Dmitrii Usynin, TUM/Imperial College London
https://proceedings.mlr.press/v202/kariyappa23a/kariyappa23a.pdf	@inproceedings{kariyappa2023cocktail, title={Cocktail party attack: Breaking aggregation-based privacy in federated learning using independent component analysis}, author={Kariyappa, Sanjay and Guo, Chuan and Maeng, Kiwan and Xiong, Wenjie and Suh, G Edward and Qureshi, Moinuddin K and Lee, Hsien-Hsin S}, booktitle={International Conference on Machine Learning}, pages={15884--15899}, year={2023}, organization={PMLR} }	Sanjay Kariyappa, Chuan Guo, Kiwan Maeng, Wenjie Xiong, G. Edward Suh, Moinuddin K Qureshi, Hsien-Hsin S. Lee	Cocktail Party Attack: Breaking Aggregation-Based Privacy in Federated Learning Using Independent Component Analysis	Theoretical attempts to present (and attack) secure aggregation in FL	Image	Predictive-Model	Reconstruction	Theoretical			Dmitrii Usynin, TUM/Imperial College London
https://dl.acm.org/doi/abs/10.1145/3592800	@article{usynin2023beyond, title={Beyond gradients: Exploiting adversarial priors in model inversion attacks}, author={Usynin, Dmitrii and Rueckert, Daniel and Kaissis, Georgios}, journal={ACM Transactions on Privacy and Security}, volume={26}, number={3}, pages={1--30}, year={2023}, publisher={ACM New York, NY} }	Dmitrii Usynin, Daniel Rueckert, Georgios Kaissis	Beyond Gradients: Exploiting Adversarial Priors in Model Inversion Attacks	More successful model inversion using HbC adversary with the knowledge of context and style of the training data	Image	Predictive-Model	Reconstruction	Empirical			Dmitrii Usynin, TUM/Imperial College London
https://arxiv.org/abs/2404.02936	@article{zhang2024min, title={Min-K\%++: Improved Baseline for Detecting Pre-Training Data from Large Language Models}, author={Zhang, Jingyang and Sun, Jingwei and Yeats, Eric and Ouyang, Yang and Kuo, Martin and Zhang, Jianyi and Yang, Hao and Li, Hai}, journal={arXiv preprint arXiv:2404.02936}, year={2024} }	Jingyang Zhang et al.	Min-K%++: Improved Baseline for Detecting Pre-Training Data from Large Language Models	Impvroved the Min-K attack by normalizing the log prob of tokens	Text	Generative-Model	Membership-Inference	Applications	https://zjysteven.github.io/mink-plus-plus/		Hamid Mozaffari, Oracle Labs
https://petsymposium.org/2019/files/papers/issue4/popets-2019-0067.pdf	@inproceedings{hilprecht2019monte, title={{Monte carlo and reconstruction membership inference attacks against generative models}}, author={Hilprecht, Benjamin and H{\"a}rterich, Martin and Bernau, Daniel}, booktitle={PoPETs}, year={2019}, }	Benjamin Hilprecht, Martin Härterich, and Daniel Bernau	Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models	Present two information leakage attacks that outperform previous work on membership inference against generative models.	Image	Generative-Model	Membership-Inference	Empirical			Georgi Ganev, UCL
https://arxiv.org/abs/2312.05114	@article{ganev2023inadequacy, title={On the Inadequacy of Similarity-based Privacy Metrics: Reconstruction Attacks against "Truly Anonymous Synthetic Data"}, author={Ganev, Georgi and De Cristofaro, Emiliano}, journal={arXiv:2312.05114}, year={2023} }	Georgi Ganev, Emiliano De Cristofaro	On the Inadequacy of Similarity-based Privacy Metrics: Reconstruction Attacks against "Truly Anonymous Synthetic Data"	Reviews the privacy metrics offered by leading companies in this space and sheds light on a few critical flaws in reasoning about privacy entirely via empirical evaluations. Presents a reconstruction attack, ReconSyn, which successfully recovers at least 78% of the low-density train records (or outliers) with only black-box access to a single fitted generative model and the privacy metrics.	Tabular	Generative-Model	Reconstruction	Empirical			Georgi Ganev, UCL
https://arxiv.org/abs/2102.03314	@inproceedings{oprisanu2022on, title={On utility and privacy in synthetic genomic data}, author={Oprisanu, Bristena and Ganev, Georgi and De Cristofaro, Emiliano}, booktitle={NDSS}, year={2022} }	Bristena Oprisanu, Georgi Ganev, Emiliano De Cristofaro	On utility and privacy in synthetic genomic data	Provides the first evaluation of both utility and privacy protection of six state-of-the-art models for generating synthetic genomic data. The experiments show that no single approach to generate synthetic genomic data yields both high utility and strong privacy across the board.	Tabular	Generative-Model	Membership-Inference	Empirical			Georgi Ganev, UCL
https://arxiv.org/pdf/2302.03098.pdf	@article{andrew2023one, title={One-shot empirical privacy estimation for federated learning}, author={Andrew, Galen and Kairouz, Peter and Oh, Sewoong and Oprea, Alina and McMahan, H Brendan and Suriyakumar, Vinith}, journal={arXiv preprint arXiv:2302.03098}, year={2023} }	Galen Andrew, Peter Kairouz, Sewoong Oh, Alina Oprea, H. Brendan McMahan, Vinith M. Suriyakumar	One-shot Empirical Privacy Estimation for Federated Learning	Presents an apporach for performing a strong white-box attack for measuring the DP epsilon of ML training algos in 1 training run.		Generative-Model	Membership-Inference				Peter Kairouz, Google
https://ieeexplore.ieee.org/documer	@INPROCEEDINGS{8429311, author={Yeom, Samuel and Giacomelli, Irene and Fredrikson, Matt and Jha, Somesh}, booktitle={2018 IEEE 31st Computer Security Foundations Symposium (CSF)}, title={Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting}, year={2018}, volume={}, number={}, pages={268-282}, keywords={Privacy;Machine learning algorithms;Data models;Training data;Machine learning;Training;privacy;machine-learning;inference-attacks}, doi={10.1109/CSF.2018.00027}}	Samuel Yeom; Irene Giacomelli; Matt Fredrikson; Somesh Jha	Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting	MIA attack based on final loss for LLMs	Text	Generative-Model	Membership-Inference	Empirical			Daniil Filienko, University of Washington
https://arxiv.org/abs/2203.03929	@article{mireshghallah2022quantifying, title={Quantifying privacy risks of masked language models using membership inference attacks}, author={Mireshghallah, Fatemehsadat and Goyal, Kartik and Uniyal, Archit and Berg-Kirkpatrick, Taylor and Shokri, Reza}, journal={arXiv preprint arXiv:2203.03929}, year={2022} }	Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, Reza Shokri	Quantifying Privacy Risks of Masked Language Models Using Membership Inference Attacks	Membership inference attack based on likelihood ratio hypothesis testing that involves an additional reference MLM to more accurately quantify the privacy risks of memorization in MLMs.	Text	Generative-Model	Membership-Inference	Applications			Daniil Filienko, University of Washington
https://arxiv.org/abs/2311.17035	@article{nasr2023scalable, title={Scalable extraction of training data from (production) language models}, author={Nasr, Milad and Carlini, Nicholas and Hayase, Jonathan and Jagielski, Matthew and Cooper, A Feder and Ippolito, Daphne and Choquette-Choo, Christopher A and Wallace, Eric and Tram{\`e}r, Florian and Lee, Katherine}, journal={arXiv:2311.17035}, year={2023} }	Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, Katherine Lee	Scalable extraction of training data from (production) language models	Studies extractable memorization: training data that an adversary can efficiently extract by querying a machine learning model without prior knowledge of the training dataset. Shows an adversary can extract gigabytes of training data from open-source language models like Pythia or GPT-Neo, semi-open models like LLaMA or Falcon, and closed models like ChatGPT.	Text	Generative-Model	Data-Extraction	Empirical			Georgi Ganev, UCL
https://arxiv.org/abs/2101.04535	@inproceedings{nasr2021adversary, title={Adversary instantiation: lower bounds for differentially private machine learning}, author={Nasr, Milad and Songi, Shuang and Thakurta, Abhradeep and Papernot, Nicolas and Carlin, Nicholas}, booktitle={IEEE S\&P}, year={2021} }	Milad Nasr, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, Nicholas Carlini	Adversary instantiation: lower bounds for differentially private machine learning	Instantiates a hypothetical adversary in order to establish lower bounds on the probability that the distinguishing game can be won. Uses this adversary to evaluate the importance of the adversary capabilities allowed in the privacy analysis of DP training algorithms.	Image	Predictive-Model	Membership-Inference	Empirical			Georgi Ganev, UCL

OpenDP Privacy Attacks & Auditing Working Group											
Privacy Attacks Respository											
URL	**BibTex (Please add a bibtex entry for this paper to facilitate writing our summary document)**	**Authors**	**Title**	**Short Description**	**Type of Data**	**Type of Release**	**Threat Model**	**Research Type**	**Links to Artifacts**	**Comments**	**Submitter (your name, affiliation)**
https://arxiv.org/abs/2403.06634	@article{carlini2024stealing, title={Stealing part of a production language model}, author={Carlini, Nicholas and Paleka, Daniel and Dvijotham, Krishnamurthy Dj and Steinke, Thomas and Hayase, Jonathan and Cooper, A Feder and Lee, Katherine and Jagielski, Matthew and Nasr, Milad and Conmy, Arthur and others}, journal={arXiv preprint arXiv:2403.06634}, year={2024} }	Nicholas Carlini, Daniel Paleka, Krishnamurthy (Dj) Dvijotham, Thomas Steinke, Jonathan Hayase, A. Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, Eric Wallace, David Rolnick, Florian Tramèr	Stealing Part of a Production Language Model	Recovers the embedding projection layer (up to symmetries) of a transformer model, given typical black-box (API) access	Text	Generative-Model	Data-Extraction	Applications			Daniil Filienko, University of Washington
https://arxiv.org/pdf/2011.07018.pdf	@inproceedings{stadler2022synthetic, title={Synthetic data--anonymisation groundhog day}, author={Stadler, Theresa and Oprisanu, Bristena and Troncoso, Carmela}, booktitle={31st USENIX Security Symposium (USENIX Security 22)}, pages={1451--1468}, year={2022} }	Stadler et al.	Synthetic Data - Anonymisation Groundhog Day	Black-box attack on synthetic data and quantitative evaluation of privacy of synthetic data	Tabular	Generative-Model	Membership-Inference	Empirical			Yves-Alexandre de Montjoye, Imperial College
https://arxiv.org/abs/2302.07956	@inproceedings{nasr2023tight, title={Tight Auditing of Differentially Private Machine Learning}, author={Milad Nasr and Jamie Hayes and Thomas Steinke and Borja Balle and Florian Tram{\`e}r and Matthew Jagielski and Nicholas Carlini and Andreas Terzis}, booktitle={USENIX Security}, year = {2023} }	Milad Nasr, Jamie Hayes, Thomas Steinke, Borja Balle, Florian Tramèr, Matthew Jagielski, Nicholas Carlini, Andreas Terzis	Tight Auditing of Differentially Private Machine Learning	Designs an improved auditing scheme that yields tight privacy estimates for natural (not adversarially crafted) datasets -- if the adversary can see all model updates during training. Moreover, the auditing scheme requires only two training runs (instead of thousands) to produce tight privacy estimates, by adapting recent advances in tight composition theorems for differential privacy.	Image	Predictive-Model	Membership-Inference	Empirical			Georgi Ganev, UCL
https://dl.acm.org/doi/10.1145/3291276.3295691	@article{garfinkel2019understanding, title={Understanding database reconstruction attacks on public data}, author={Garfinkel, Simson and Abowd, John M and Martindale, Christian}, journal={ACM Queue}, year={2019}, }	Simson Garfinkel, John M Abowd, Christian Martindale	Understanding database reconstruction attacks on public data	Database reconstruction attacks can be performed by using published statistical tables to create a set of mathematical constraints and then solving the resulting set of simultaneous equations. Shows how such an attack can be addressed by adding noise to the published tabulations, so that the reconstruction no longer results in the original data. This has implications for the 2020 Census.	Tabular	Linear-Queries	Reconstruction	Empirical			Georgi Ganev, UCL
https://arxiv.org/abs/2307.01701	@article{guepin2023synthetic, title={Synthetic is all you need: removing the auxiliary data assumption for membership inference attacks against synthetic data}, author={Gu{\'e}pin, Florent and Meeus, Matthieu and Cretu, Ana-Maria and de Montjoye, Yves-Alexandre}, journal={arXiv:2307.01701}, year={2023} }	Florent Guépin, Matthieu Meeus, Ana-Maria Cretu, Yves-Alexandre de Montjoye	Synthetic is all you need: removing the auxiliary data assumption for membership inference attacks against synthetic data	Develops new MIAs performed using only the synthetic data in three different scenarios: (S1) Black-box access to the generator, (S2) only access to the released synthetic dataset and (S3) a theoretical setup as upper bound for the attack performance.	Tabular	Generative-Model	Membership-Inference	Empirical			Georgi Ganev, UCL
https://arxiv.org/abs/2206.07758	@inproceedings{haim2022reconstructing, title={Reconstructing training data from trained neural networks}, author={Haim, Niv and Vardi, Gal and Yehudai, Gilad and Irani, Michal and Shamir, Ohad}, booktitle={NeurIPS}, year={2022}, }	Niv Haim, Gal Vardi, Gilad Yehudai, Ohad Shamir, Michal Irani	Reconstructing Training Data from Trained Neural Networks	Shows that in some cases a significant fraction of the training data can in fact be reconstructed from the parameters of a trained neural network classifier. Proposes a novel reconstruction scheme that stems from recent theoretical results about the implicit bias in training neural networks with gradient-based methods.	Image	Predictive-Model	Reconstruction	Empirical	https://giladude1.github.io/reconstruction/		Georgi Ganev, UCL
https://arxiv.org/abs/2201.04845	@inproceedings{balle2022reconstructing, title={Reconstructing training data with informed adversaries}, author={Balle, Borja and Cherubin, Giovanni and Hayes, Jamie}, booktitle={IEEE S\&P}, year={2022}, }	Borja Balle, Giovanni Cherubin, Jamie Hayes	Reconstructing training data with informed adversaries	Studies how given access to a machine learning model an adversary can reconstruct the model's training data from the lens of a powerful informed adversary who knows all the training data points except one.	Image	Predictive-Model	Reconstruction	Empirical			Georgi Ganev, UCL
https://www.usenix.org/system/files/sec20summer_salem_prepub.pdf	@inproceedings{salem2020updates, title={Updates-leak: Data set inference and reconstruction attacks in online learning}, author={Salem, Ahmed Mohamed Gamal and Bhattacharyya, Apratim and Backes, Michael and Fritz, Mario and Zhang, Yang}, booktitle={USENIX Security}, year={2020}, }	Ahmed Salem, Apratim Bhattacharya, Michael Backes, Mario Fritz, Yang Zhang	Updates-leak: Data set inference and reconstruction attacks in online learning	Investigate whether the change in the output of a black-box ML model before and after being updated can leak information of the dataset used to perform the update, namely the updating set. Proposes four attacks following an encoder-decoder formulation, which allows inferring diverse information of the updating set.	Image	Predictive-Model	Reconstruction	Empirical			Georgi Ganev, UCL
https://dl.acm.org/doi/pdf/10.1145/3576915.3616607	@inproceedings{lokna2023group, title={Group and Attack: Auditing Differential Privacy}, author={Lokna, Johan and Paradis, Anouk and Dimitrov, Dimitar I and Vechev, Martin}, booktitle={CCS}, year={2023} }	Johan Lokna, Anouk Paradis, Dimitar I. Dimitrov, Martin Vechev	Group and Attack: Auditing Differential Privacy	Present a novel method to efficiently discover (ϵ, δ) differential privacy violations based on the key insight that many (ϵ, δ) pairs can be grouped as they result in the same algorithm. Crucially, the method is orthogonal to existing approaches and, when combined, results in a faster and more precise violation search.			Information Leakage	Empirical	https://github.com/eth-sri/Delta-Siege		Georgi Ganev, UCL
https://arxiv.org/abs/1802.08232	@inproceedings{carlini2019secret, title={The secret sharer: Evaluating and testing unintended memorization in neural networks}, author={Carlini, Nicholas and Liu, Chang and Erlingsson, {\'U}lfar and Kos, Jernej and Song, Dawn}, booktitle={28th USENIX security symposium (USENIX security 19)}, pages={267--284}, year={2019} }	Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, Dawn Song	The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks	Describes a testing methodology for quantitatively assessing the risk that rare or unique training-data sequences are unintentionally memorized by generative sequence models.	Text	Generative-Model	Data-Extraction	Empirical			Georgi Ganev, UCL
https://arxiv.org/abs/2202.12219	@article{tramer2022debugging, title={Debugging Differential Privacy: A Case Study for Privacy Auditing}, author={Tramer, Florian and Terzis, Andreas and Steinke, Thomas and Song, Shuang and Jagielski, Matthew and Carlini, Nicholas}, journal={arXiv:2202.12219}, year={2022} }	Florian Tramer, Andreas Terzis, Thomas Steinke, Shuang Song, Matthew Jagielski, Nicholas Carlini	Debugging Differential Privacy: A Case Study for Privacy Auditing	Inspired by recent advances in auditing which have been used for estimating lower bounds on differentially private algorithms, shows that auditing can also be used to find flaws in (purportedly) differentially private schemes.	Image	Predictive-Model	Information Leakage	Empirical			Georgi Ganev, UCL
https://arxiv.org/abs/2206.05199	@inproceedings{zanella2023bayesian, title={Bayesian estimation of differential privacy}, author={Zanella-B{\'e}guelin, Santiago and Wutschitz, Lukas and Tople, Shruti and Salem, Ahmed and R{\"u}hle, Victor and Paverd, Andrew and Naseri, Mohammad and K{\"o}pf, Boris and Jones, Daniel}, booktitle={ICML}, year={2023} }	Santiago Zanella-Béguelin, Lukas Wutschitz, Shruti Tople, Ahmed Salem, Victor Rühle, Andrew Paverd, Mohammad Naseri, Boris Köpf, Daniel Jones	Bayesian Estimation of Differential Privacy	Proposes a novel Bayesian method that greatly reduces sample size, and adapt and validate a heuristic to draw more than one sample per trained model. The Bayesian method exploits the hypothesis testing interpretation of differential privacy to obtain a posterior for ε (not just a confidence interval) from the joint posterior of the false positive and false negative rates of membership inference attacks.	Image	Predictive-Model	Information Leakage	Empirical			Georgi Ganev, UCL
https://files.sri.inf.ethz.ch/website/papers/sp21-dpsniper.pdf	@inproceedings{bichsel2021dp, title={DP-Sniper: Black-Box Discovery of Differential Privacy Violations using Classifiers}, author={Bichsel, Benjamin and Steffen, Samuel and Bogunovic, Ilija and Vechev, Martin}, booktitle={IEEE S\&P}, year={2021} }	Benjamin Bichsel, Samuel Steffen, Ilija Bogunovic, Martin Vechev	DP-Sniper: Black-Box Discovery of Differential Privacy Violations using Classifier	Present DP-Sniper, a practical black-box method that automatically finds violations of differential privacy.	Tabular		Information Leakage	Empirical			Georgi Ganev, UCL
https://arxiv.org/pdf/2310.09266.pdf	@article{kandpal2023user, title={User inference attacks on large language models}, author={Kandpal, Nikhil and Pillutla, Krishna and Oprea, Alina and Kairouz, Peter and Choquette-Choo, Christopher A and Xu, Zheng}, journal={arXiv preprint arXiv:2310.09266}, year={2023} }	Nikhil Kandpal, Krishna Pillutla, Alina Oprea, Peter Kairouz, Christopher A. Choquette-Choo, Zheng Xu	User Inference Attacks on Large Language Models	Presents an attack for inferencing the precense/absence of a user in the fine-tuning set of an LLM. The adversary is not assumed to know all the fine-tuning examples of a user -- only a subset (including some examples that weren't used even if the user participated in the fine-tuning stage).		Generative-Model	Membership-Inference	Empirical			Peter Kairouz, Google
https://arxiv.org/abs/2406.11544	@inproceedings{suri2024do, title = {Do Parameters Reveal More than Loss for Membership Inference?}, author = {Anshuman Suri and Xiao Zhang and David Evans}, booktitle = {Workshop on High-dimensional Learning Dynamics (HiLD), ICML}, year = {2024}, url = {https://arxiv.org/abs/2406.11544} }	Anshuman Suri, Xiao Zhang, David Evans	Do Parameters Reveal More than Loss for Membership Inference?	The paper shows how prior claims about black-box access sufficing for optimal membership inference do not hold for most useful settings such as SGD	Tabular	Predictive-Model	Membership-Inference	Theoretical	https://github.com/iamgroot42/iha_hild		Anshuman Suri, UVA
https://arxiv.org/abs/2402.10001	@INPROCEEDINGS{ElMrini2024a, author = {{El Mrini}, Abdellah and {Cyffers}, Edwige and {Bellet}, Aur{\'e}lien}, title = {{P}rivacy {A}ttacks in {D}ecentralized {L}earning}, booktitle = {{ICML}}, year = {2024}}	Abdellah El Mrini, Edwige Cyffers, Aurélien E	Privacy Attacks in Decentralized Learning	The paper designs data reconstruction attacks against Decentralized SGD (where nodes in a communication graph alternate between local gradient steps and averaging steps with their neighbors). They show it is possible for a small subset of (honest but curious) attacker nodes to reconstruct the data from even distant nodes in the graph.	Image	Predictive-Model	Reconstruction	Empirical	https://github.com/AbdellahElmrini/decAttack		Aurélien Bellet, Inria

OpenDP Privacy Attacks & Auditing Working Group											
Privacy Attacks Respository											
URL	**BibTex (Please add a bibtex entry for this paper to facilitate writing our summary document)**	**Authors**	**Title**	**Short Description**	**Type of Data**	**Type of Release**	**Threat Model**	**Research Type**	**Links to Artifacts**	**Comments**	**Submitter (your name, affiliation)**
https://arxiv.org/abs/2106.03408	@inproceedings{malek2021labeldp, author = {Malek Esmaeili, Mani and Mironov, Ilya and Prasad, Karthik and Shilov, Igor and Tramer, Florian}, booktitle = {Advances in Neural Information Processing Systems}, editor = {M. Ranzato and A. Beygelzimer and Y. Dauphin and P.S. Liang and J. Wortman Vaughan}, pages = {6934--6945}, publisher = {Curran Associates, Inc.}, title = {Antipodes of Label Differential Privacy: {PATE} and {ALIBI}}, url = {https://proceedings.neurips.cc/paper_files/paper/2021/file/37ecd27608480aa3569a511a638ca74f-Paper.pdf}, volume = {34}, year = {2021} }	Mani Malek, Ilya Mironov, Karthik Prasad, Igor Shilov, Florian Tramèr	Antipodes of Label Differential Privacy: PATE and ALIBI	Label-inference attack against two LabelDP mechanisms for image classification	Image	Predictive-Model	Attribute inference	Empirical	https://github.com/facebookresearch/label_dp_antipodes/tree/main/memorization_attack		Ilya Mironov, Meta
https://arxiv.org/pdf/2212.10986	@INPROCEEDINGS {salem2023-sok, author = {A. Salem and G. Cherubin and D. Evans and B. Kopf and A. Paverd and A. Suri and S. Tople and S. Zanella-Beguelin}, booktitle = {2023 IEEE Symposium on Security and Privacy (S\& P)}, title = {{SoK}: {L}et the Privacy Games Begin! {A} Unified Treatment of Data Inference Privacy in Machine Learning}, year = {2023}, pages = {327-345}, doi = {10.1109/SP46215.2023.10179281}, url = {https://doi.ieeecomputersociety.org/10.1109/SP46215.2023.10179281}, publisher = {IEEE Computer Society}, address = {Los Alamitos, CA, USA}, month = {may} }	Ahmed Salem, Giovanni Cherubin, David Evans, Boris Köpf, Andrew Paverd, Anshuman Suri, Shruti Tople, Santiago Zanella-Béguelin	SoK: Let the Privacy Games Begin! A Unified Treatment of Data Inference Privacy in Machine Learning	Systematization of Knowledge of data inference attacks using a privacy-game framework.				Theoretical			Ilya Mironov, Meta
https://arxiv.org/abs/2111.08440	@inproceedings{watson2022-difficulty, author = {Lauren Watson and Chuan Guo and Graham Cormode and Alexandre Sablayrolles}, title = {On the Importance of Difficulty Calibration in Membership Inference Attacks}, booktitle = {The Tenth International Conference on Learning Representations, {ICLR} 2022, Virtual Event, April 25-29, 2022}, publisher = {OpenReview.net}, year = {2022}, url = {https://openreview.net/forum?id=3eIrli0TwQ},}	Lauren Watson, Chuan Guo, Graham Cormode, Alex Sablayrolles	On the Importance of Difficulty Calibration in Membership Inference Attacks	Improvement in membership inference attacks by taking into account the difficulty of correct classification	Image, tabular	Predictive-Model	Membership-Inference	Empirical	https://github.com/facebookresearch/calibration_membership		Ilya Mironov, Meta
https://oaklandsok.github.io/papers/papernot2018.pdf	@INPROCEEDINGS{papernot2018-sok, author={Papernot, Nicolas and McDaniel, Patrick and Sinha, Arunesh and Wellman, Michael P.}, booktitle={2018 IEEE European Symposium on Security and Privacy (EuroS\&P)}, title={SoK: Security and Privacy in Machine Learning}, year={2018}, volume={}, number={}, pages={399-414}, }	Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, Michael P. Wellman	SoK: Security and Privacy in Machine Learning	Systematization of Knowledge of attacks on privacy of ML				survey			Ilya Mironov, Meta
https://arxiv.org/abs/1807.09173	@ARTICLE{truex2021, author={Truex, Stacey and Liu, Ling and Gursoy, Mehmet Emre and Yu, Lei and Wei, Wenqi}, journal={IEEE Transactions on Services Computing}, title={Demystifying Membership Inference Attacks in Machine Learning as a Service}, year={2021}, volume={14}, number={6}, pages={2073-2089}, doi={10.1109/TSC.2019.2897554}}	Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei	Demystifying Membership Inference Attacks in Machine Learning as a Service	(1) MIA in the style of Shokri et al.'17 with shadow models with architecture different from that of the target model. (2) MIA with access to training gradients in the federated setting.	Image, tabular	Predictive-Model	Membership-Inference	Empirical			Ilya Mironov, Meta
https://arxiv.org/abs/2005.13702	@InProceedings{Rezaei_2021_CVPR, author = {Rezaei, Shahbaz and Liu, Xin}, title = {On the Difficulty of Membership Inference Attacks}, booktitle = {Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)}, month = {June}, year = {2021}, pages = {7892-7900} }	Shahbaz Rezaei, Xin Liu	On the Difficulty of Membership Inference Attacks	Critique of prior MIAs for their poor false positive rate.	Image	Predictive-Model	Membership-Inference	Empirical	https://github.com/shrezaei/MI-Attack		
https://arxiv.org/abs/2312.11283	@misc {abowd20232010censusconfidentialityprotections, title={The 2010 Census Confidentiality Protections Failed, Here's How and Why}, author={John M. Abowd and Tamara Adams and Robert Ashmead and David Darais and Sourya Dey and Simson L. Garfinkel and Nathan Goldschlag and Daniel Kifer and Philip Leclerc and Ethan Lew and Scott Moore and Rolando A. Rodríguez and Ramy N. Tadros and Lars Vilhuber}, year={2023}, eprint={2312.11283}, archivePrefix={arXiv}, }	John M. Abowd, Tamara Adams, Robert Ashmead, David Darais, Sourya Dey, Simson L. Garfinkel, Nathan Goldschlag, Daniel Kifer, Philip Leclerc, Ethan Lew, Scott Moore, Rolando A. Rodríguez, Ramy N. Tadros, Lars Vilhuber	The 2010 Census Confidentiality Protections Failed, Here's How and Why	The definitive analysis of the reconstruction-abetted reidentification attack on the 2010 Census data. Microdata were reconstructed from tabular summaries using a series of Integer Programs. An additional Integer Program assesses the solution variability. For 97 million of 308 million persons, the reconstruction is provably perfect. For perfectly reconstructed persons with atypical characteristics within their geography, inferences about these characteristics are correct with probability 0.95, far in excess of correct inferences from purely statistical models. Inferences are also correct with similar probability for typical persons within their geography, but such inferences could have been made with a purely statistical model.	Tabular	Linear-Queries	Reconstruction	Empirical	https://github.com/uscensusbureau/recon_replication		John Abowd, Cornell
					Tabular						