Gradient-Leaks: Understanding Deanonymization in Federated Learning

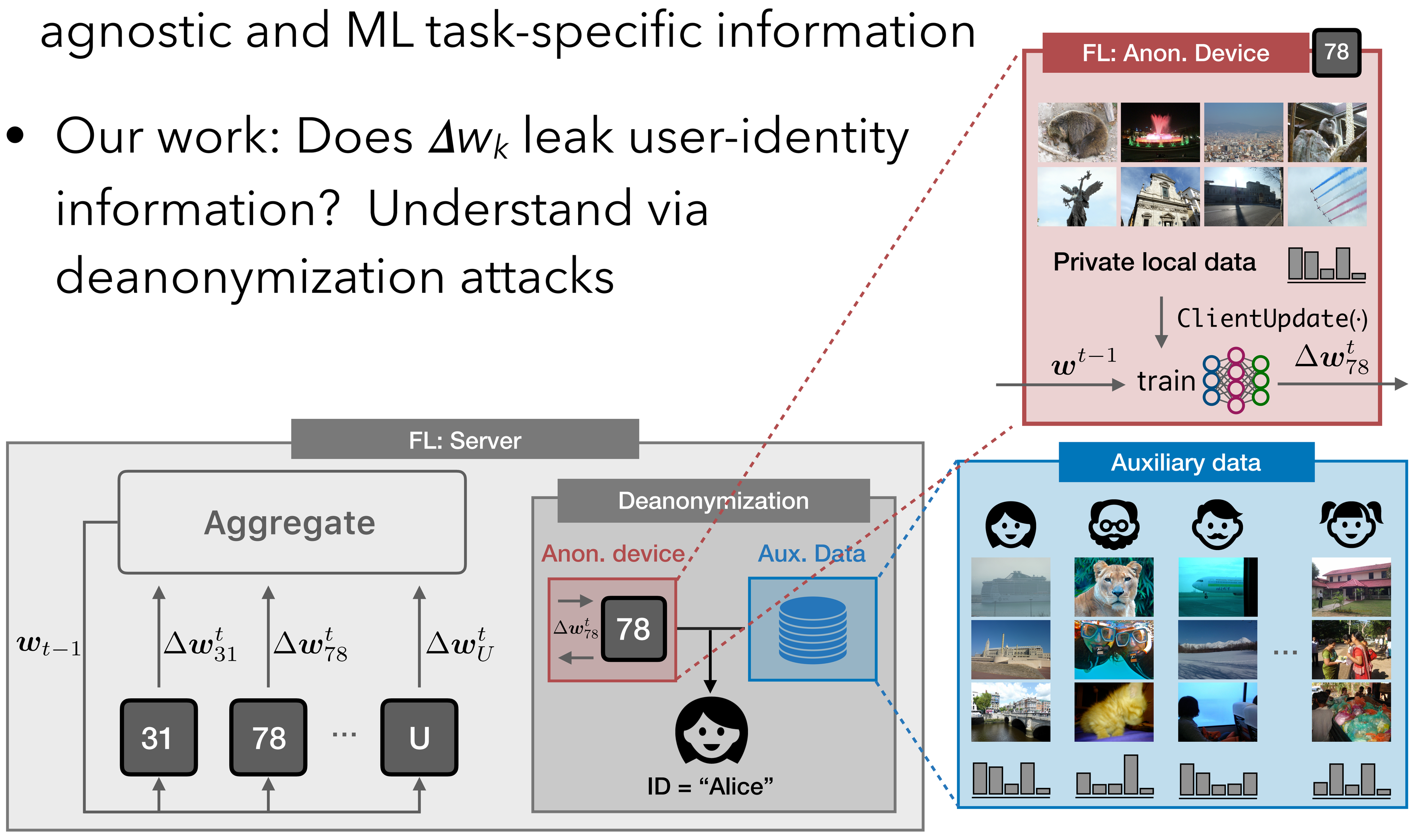
Tribhuvanesh Orekondy[1], Seong Joon Oh[2], Yang Zhang[3], Bernt Schiele[1], Mario Fritz[3]

[1] Max Planck Institute for Informatics [2] LINE Plus (Naver) Clova ML

[3] CISPA Helmholtz Centre for Information Security

Motivation

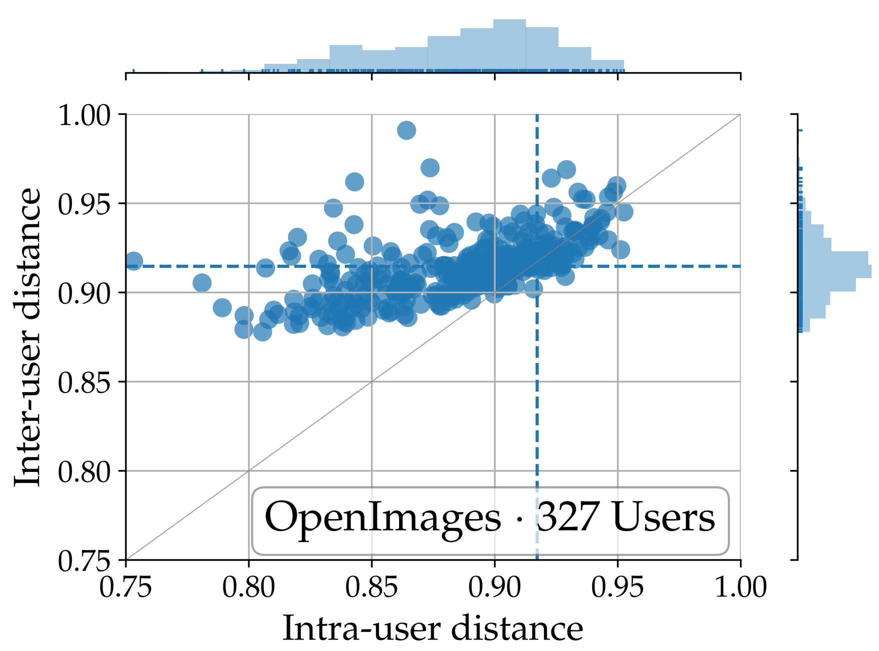
- Federated Learning: promising approach to ensure user/data privacy for training ML models on devices e.g., smartphones
- Only model parameter deltas Δw_k are anonymously shared: Δw_k = `ClientUpdate`(Private user data)
- To protect users' privacy, want Δw_k to encode only user-agnostic and ML task-specific information
- Our work: Does Δw_k leak user-identity information? Understand via deanonymization attacks

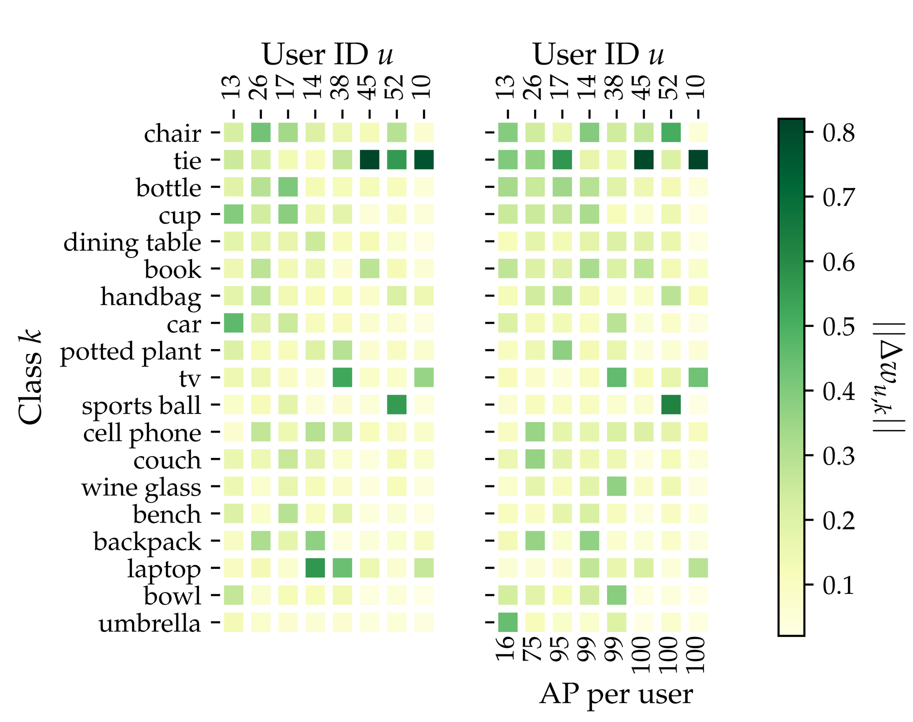


Deanonymization Attacks in FL

Insight

Biased User Data → Biased Estimators





Threat Model

$$\left(\ \boxed{\Delta \boldsymbol{w}_{\text{anon}}}\ ,\ \ \Delta \boldsymbol{w}_u^{\text{aux}}\ \right) \mapsto \mathbb{P}(\ \text{anon} = u\ )$$

=`ClientUpdate`(Aux. data of user u) Observed by Attacker

Attack Models

- Re-identification:

$$f^{\text{re-id}} : \Delta \boldsymbol{w}_{\text{anon}} \mapsto u$$

⇒ Learnt using MLP

- Matching:

$$f^{\text{mat}} : (\Delta \boldsymbol{w}_{u_i},\ \Delta \boldsymbol{w}_{u_j}) \mapsto \mathbb{P}(u_i = u_j)$$

⇒ Siamese Network

Acknowledgment This research was partially supported by the German Research Foundation (DFG CRC 1223)

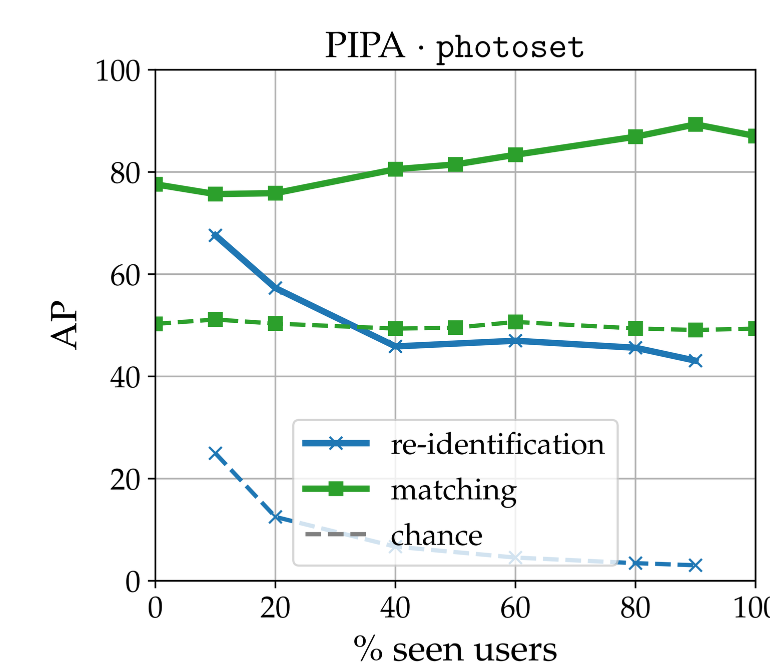
Take-aways

- Task: Deanonymization attack of devices in FL
- Insight: Use user selection bias as a quasi-identification statistical signal to perform deanonymization
- Attacks are effective: 16–91% AP, 19–175x chance-level re-identification performance
- Poses a threat to privacy and anonymity of users participating in FL

Evaluation

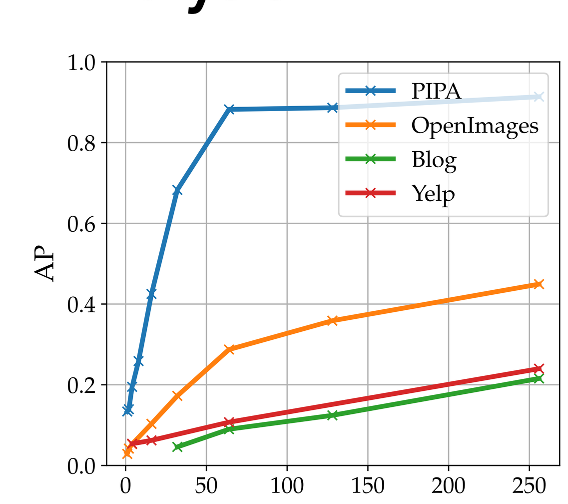
How Effective are Deanonymization Attacks?

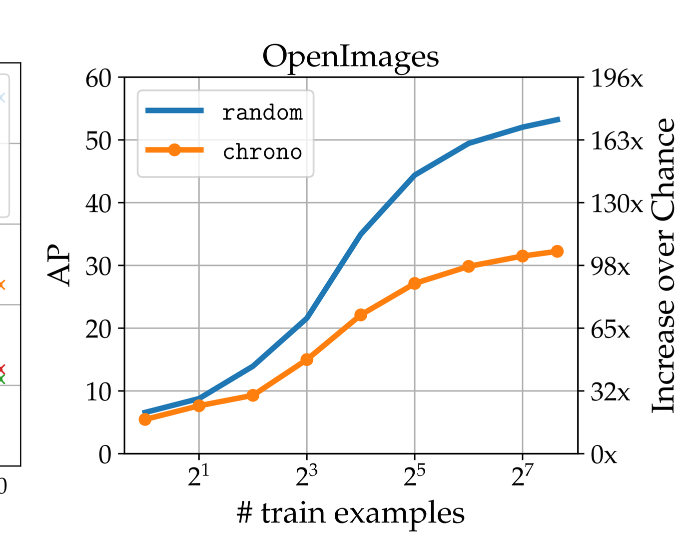
		Re-ID	Matching
PIPA	random	91 (48×)	99.5 (2×)
	photoset	42.2 (22×)	91.2 (1.85×)
Open Images	random	53.7 (175×)	98.2 (1.93×)
	chrono	32.5 (106×)	94.8 (1.93×)
Blog	random	52.9 (29×)	95.3 (1.9×)
	chrono	44.8 (25×)	91.9 (1.89×)
Yelp	random	23.5 (28×)	83.4 (1.7×)
	chrono	16 (19×)	79.3 (1.56×)



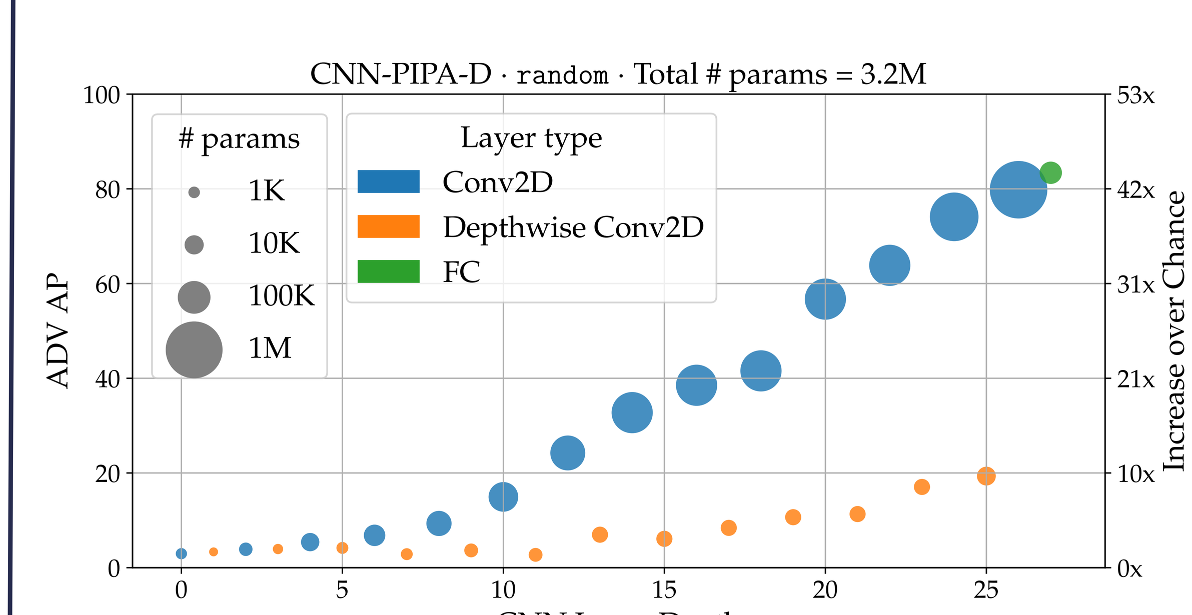
- Closed-world: 19-175x chance-level performance Re-ID performance
- Open-world: Robust to encountering new unseen users at test-time

Analysis

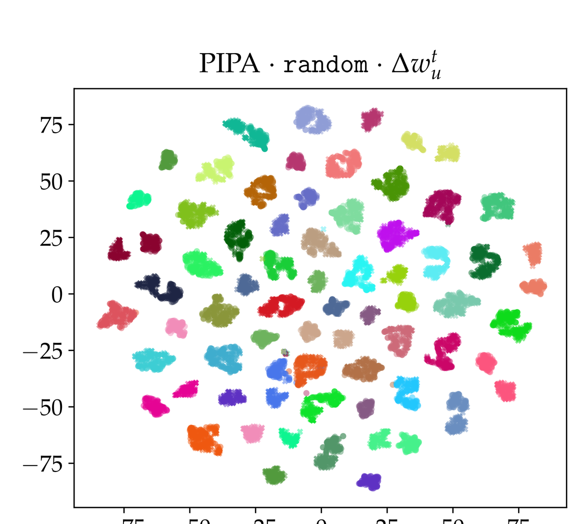


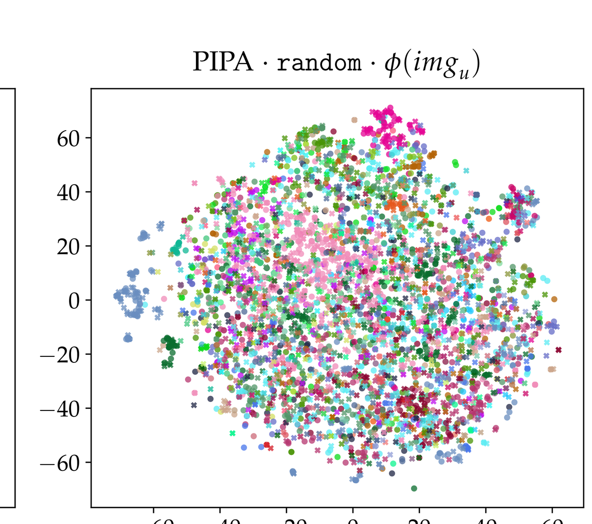


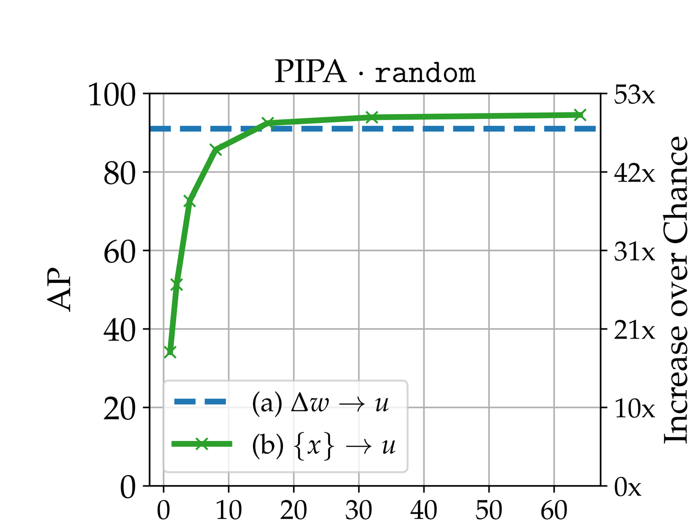
- vs. # attacker's training examples: Attacks possible in few-shot training settings



- vs. parameter type: All parameters provide above-chance level information
- FC most informative







- Δw_k encodes aggregated data information → sometimes easier to deanonymize via Δw_k than raw images themselves