BETTER PRIVACY GUARANTEES FOR DECENTRALIZED FEDERATED LEARNING

Aurélien Bellet (Inria Lille)

Joint work with Edwige Cyffers (Inria Lille), Mathieu Even and Laurent Massoulié (Inria Paris)

Workshop FL-Day - Decentralized Federated Learning: Approaches and Challenges
January 10, 2023
In decentralized algorithms, such as decentralized SGD [Lian et al., 2017] [Koloskova et al., 2020], users communicate along the edges of a graph.

These algorithms are increasingly popular in machine learning due to their scalability.
In decentralized algorithms, such as decentralized SGD [Lian et al., 2017] [Koloskova et al., 2020], users communicate along the edges of a graph.

These algorithms are increasingly popular in machine learning due to their scalability.

Folklore belief: “Decentralized algorithms are good for privacy because users have a limited view of the system.”

Question: is this claim really true? can we formalize and quantify these gains? Yes!
1. Background: Differential Privacy & DP-SGD

2. A relaxation of local DP for decentralized algorithms

3. Private random walk-based decentralized SGD

4. Private gossip-based decentralized SGD

5. Conclusion & Perspectives
BACKGROUND: DIFFERENTIAL PRIVACY & DP-SGD
• ML models are susceptible to various attacks on data privacy

• Membership inference attack: infer whether a known individual data point was present in the training set

• For instance, one can exploit overconfidence in model predictions [Shokri et al., 2017] [Carlini et al., 2022]
• **Reconstruction attack**: extract training data points from the model

• For instance, one can extract sensitive text from large language models [Carlini et al., 2021] or run differencing attacks on ML models [Paige et al., 2020]
DIFFERENTIAL PRIVACY

• Neighboring datasets $D = \{x_1, x_2, \ldots, x_n\}$ and $D' = \{x_1, x'_2, x_3, \ldots, x_n\}$

• Requirement: $A(D)$ and $A(D')$ should have “similar” distributions

![Diagram showing randomized algorithms $A$ with input $x_1, x_2, \ldots, x_n$ and output $A(D)$ and $A(D')$ with probability distributions. The output probability ratio is bounded.]
### Definition (Rényi Differential Privacy [Mironov, 2017])

An algorithm $A$ satisfies $(\alpha, \epsilon)$-Rényi Differential Privacy (RDP) for $\alpha > 1$ and $\epsilon > 0$ if for all pairs of neighboring datasets $D \sim D'$:

$$D_{\alpha} (A(D) \| A(D')) \leq \epsilon,$$

where for two r.v. $X, Y$ with densities $\mu_X, \mu_Y$, $D_{\alpha}(X \| Y)$ is the Rényi divergence of order $\alpha$:

$$D_{\alpha}(X \| Y) = \frac{1}{\alpha - 1} \ln \int \left( \frac{\mu_X(z)}{\mu_Y(z)} \right)^{\alpha} \mu_Y(z)dz.$$

- Conversion to standard $(\epsilon, \delta)$-DP: $(\alpha, \epsilon)$-RDP implies $(\epsilon + \frac{\ln(1/\delta)}{\alpha - 1}, \delta)$-DP for any $\delta \in (0, 1)$
PROPERTIES OF RDP

• RDP is **robust to auxiliary knowledge**, as seen by its Bayesian interpretation:
  • Consider an adversary who seeks to infer whether the dataset is $\mathcal{D}$ or $\mathcal{D}'$
  • The adversary has prior knowledge $p$ and observes $X \sim \mathcal{A}(\mathcal{D})$
  • Let the r.v. $R_{prior} = \frac{p(D')}{p(D)}$ and $R_{post} = \frac{p(D' | X)}{p(D | X)} = \frac{p(X | D') p(D')}{p(X | D) p(D)}$ for $X \sim \mathcal{A}(\mathcal{D})$
  • RDP bounds the $\alpha$-th moment of $\frac{R_{post}}{R_{prior}}$ (for $\alpha \to \infty$, we recover “pure” $\epsilon$-DP)
  • “The adversary does not know much more after observing the output of the algorithm”

• **Immunity to post-processing**: for any $g$, if $\mathcal{A}(\cdot)$ is $(\alpha, \epsilon)$-RDP, then so is $g(\mathcal{A}(\cdot))$

• **Composition**: if $\mathcal{A}_1$ is $(\alpha, \epsilon_1)$-RDP and $\mathcal{A}_2$ is $(\alpha, \epsilon_2)$-RDP, then $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2)$ is $(\alpha, \epsilon_1 + \epsilon_2)$-RDP → simpler and tighter than composition for $(\epsilon, \delta)$-DP
ENFORCING RDP WITH THE GAUSSIAN MECHANISM

• Consider $f$ taking as input a dataset and returning a $p$-dimensional real vector

• Denote its sensitivity by $\Delta = \max_{D \sim D'} \| f(D) - f(D') \|_2$

**Theorem (Gaussian mechanism)**

Let $\sigma > 0$. The algorithm $A(\cdot) = f(\cdot) + \mathcal{N}(0, \sigma^2 \Delta^2)$ satisfies $(\alpha, \frac{\alpha}{2\sigma^2})$-RDP for any $\alpha > 1$.

**Theorem (Subsampled Gaussian mechanism, informal)**

If $A$ is executed on a random fraction $q$ of $D$, then it satisfies $(\alpha, \frac{q^2 \alpha}{2\sigma^2})$-RDP.

• DP induces a privacy-utility trade-off, here in terms of the variance of the estimate

• Random subsampling amplifies privacy guarantees
• A trusted curator wants to privately release a model trained on data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$

• We focus here on approximately solving an Empirical Risk Minimization (ERM) problem under a DP constraint:

$$\min_{\theta \in \mathbb{R}^p} \left\{ F(\theta; \mathcal{D}) := \frac{1}{n} \sum_{i=1}^n \ell(\theta; x_i, y_i) \right\}, \quad \text{where loss } \ell \text{ is differentiable in } \theta$$

• Note: in some cases, DP implies generalization [Bassily et al., 2016, Jung et al., 2021]
Algorithm  Differentially Private SGD (DP-SGD) [Bassily et al., 2014, Abadi et al., 2016]

Initialize $\theta^{(0)} \in \mathbb{R}^p$ (must be independent of $D$)

for $t = 0, \ldots, T - 1$ do

Pick $i_t \in \{1, \ldots, n\}$ uniformly at random

$\eta^{(t)} \leftarrow (\eta_1^{(t)}, \ldots, \eta_p^{(t)}) \in \mathbb{R}^p$ where each $\eta_i^{(t)} \sim \mathcal{N}(0, \sigma^2 \Delta^2)$

$\theta^{(t+1)} \leftarrow \theta^{(t)} - \gamma^{(t)} \left( \nabla \ell(\theta^{(t)}; x_{i_t}, y_{i_t}) + \eta^{(t)} \right)$

Return $\theta^{(T)}$

• The sensitivity $\Delta = \sup_\theta \sup_{x,y,x',y'} \| \nabla \ell(\theta; x, y) - \nabla \ell(\theta; x', y') \|$ can be controlled by assuming $\ell(\cdot; x, y)$ Lipschitz for all $x, y$, or using gradient clipping [Abadi et al., 2016]
PRIVACY-UTILITY TRADE-OFF OF DP-SGD

- **Utility analysis**: same as non-private SGD (with additional noise due to privacy)

- **Privacy analysis**: DP-SGD is $(\alpha, \frac{\alpha^T}{2n^2\sigma^2} )$ by subsampled Gaussian mechanism + composition over $T$ iterations

- Setting $\sigma^2$ to satisfy $(\epsilon, \delta)$-DP and choosing $T$ to balance optimization and privacy errors, we get the following suboptimality gap:

  \[
  \tilde{O}
  \begin{cases}
  \frac{\sqrt{p} \ln(1/\delta)}{n\epsilon} & \text{Convex, Lipschitz, smooth loss} \\
  \frac{p \ln(1/\delta)}{n^2\epsilon^2} & \text{Convex, Lipschitz, smooth loss, strongly convex } F
  \end{cases}
  \]

- This is optimal [Bassily et al., 2014]: cannot do better without additional assumptions
So far we considered the central DP model, which relies on a trusted curator to collect and process raw data → the output $A(D)$ is only the final result.

Central DP is good for utility but is an unrealistic trust model in applications where many parties contribute sensitive data, as in federated learning.

Instead we can consider for local DP, where each party must locally randomize its contributions → the output of $A(D)$ consists of all messages sent by all parties.

Unfortunately local DP induces a large cost in utility: for averaging $n$ private $p$-dimensional values in ball of radius $\Delta$ under $(\alpha, \epsilon)$-RDP, we have

$$\mathbb{E}[\|x^{\text{out}} - \bar{x}\|^2] = \Theta\left(\frac{\alpha p \Delta^2}{n \epsilon}\right) \text{ for local DP, and } \mathbb{E}[\|x^{\text{out}} - \bar{x}\|^2] = \Theta\left(\frac{\alpha p \Delta^2}{n^2 \epsilon}\right) \text{ for central DP}$$

→ study intermediate models allowing better utility without relying on trusted parties.
A RELAXATION OF LOCAL DP FOR DECENTRALIZED ALGORITHMS
A connected graph $G = (\mathcal{V}, \mathcal{E})$ on a set of $|\mathcal{V}| = n$ users (nodes)

Each user $v \in \mathcal{V}$ holds a local dataset $\mathcal{D}_v$

A decentralized algorithm relies only on communication along the edges $\mathcal{E}$ of $G$

Each user $v$ thus has a limited view: it only observes the messages that it receives

We want to use this to prove stronger privacy guarantees than under local DP
Let $\mathcal{O}_v$ be the set of messages sent and received by party $v$.

Denote by $\mathcal{D} \sim_u \mathcal{D}'$ two datasets $\mathcal{D} = (\mathcal{D}_1, \ldots, \mathcal{D}_u, \ldots, \mathcal{D}_n)$ and $\mathcal{D}' = (\mathcal{D}_1, \ldots, \mathcal{D}'_u, \ldots, \mathcal{D}_n)$ that differ only in the local dataset of user $u$.

**Definition (Network DP [Cyffers and Bellet, 2022])**

An algorithm $\mathcal{A}$ satisfies $(\alpha, \epsilon)$-Network DP (NDP) if for all pairs of distinct users $u, v \in \mathcal{V}$ and neighboring datasets $\mathcal{D} \sim_u \mathcal{D}'$:

$$D_\alpha(\mathcal{O}_v(\mathcal{A}(\mathcal{D}))) \parallel \mathcal{O}_v(\mathcal{A}(\mathcal{D}'))) \leq \epsilon.$$ 

This is a relaxation of local DP: if $\mathcal{O}_v$ contains the full transcript of messages, then network DP boils down to local DP.
• We will also consider privacy guarantees that are specific to each pair of nodes, rather than uniform over all pairs

**Definition (Pairwise Network DP [Cyffers et al., 2022])**

For \( f : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}^+ \), an algorithm \( \mathcal{A} \) satisfies \((\alpha, f)\)-Pairwise Network DP (PNDP) if for all pairs of distinct users \( u, v \in \mathcal{V} \) and neighboring datasets \( \mathcal{D} \sim_u \mathcal{D}' \):

\[
D_\alpha(\mathcal{O}_v(\mathcal{A}(\mathcal{D}))) \| \mathcal{O}_v(\mathcal{A}(\mathcal{D}'))) \leq f(u, v).
\]

• For comparison with central and local DP baselines, we will report the mean privacy loss \( \overline{\varepsilon}_v = \frac{1}{n} \sum_{u \in \mathcal{V} \setminus \{v\}} f(u, v) \) under the constraint \( \overline{\varepsilon} = \max_{v \in \mathcal{V}} \overline{\varepsilon}_v \leq \epsilon \)

• Note: \( \overline{\varepsilon}_v \) is not a proper privacy guarantee (we simply use it to summarize our gains)
PRIVATE RANDOM WALK-BASED DECENTRALIZED SGD
PRIVATE RANDOM WALK-BASED DECENTRALIZED SGD

- Consider the standard objective $F(\theta; D) = \frac{1}{n} \sum_{v=1}^{n} F_v(\theta; D_v)$
- We consider a decentralized SGD algorithm where the model is updated sequentially by following a random walk, aka incremental gradient [Johansson et al., 2009]
- We focus here on the complete graph

**Algorithm**  Private random walk-based SGD [Cyffers and Bellet, 2022]

Initialize $\theta \in \mathbb{R}^p$

for $t = 1$ to $T$ do

1. Draw random user $v \sim \mathcal{U}(1, \ldots, n)$
2. $\eta = [\eta_1, \ldots, \eta_p]$, where each $\eta_j \sim \mathcal{N}(0, \sigma^2 \Delta^2)$
3. $\theta \leftarrow \theta - \gamma [\nabla_{\theta} F_v(\theta; D_v) + \eta]$

return $\theta$
Consider the standard objective $F(\theta; D) = \frac{1}{n} \sum_{v=1}^{n} F_v(\theta; D_v)$.

We consider a decentralized SGD algorithm where the model is updated sequentially by following a random walk, aka incremental gradient [Johansson et al., 2009].

We focus here on the complete graph.

**Algorithm**  Private random walk-based SGD [Cyffers and Bellet, 2022]

Initialize $\theta \in \mathbb{R}^p$

for $t = 1$ to $T$
do

Draw random user $v \sim U(1, \ldots, n)$

$\eta = [\eta_1, \ldots, \eta_p]$, where each $\eta_j \sim \mathcal{N}(0, \sigma^2 \Delta^2)$

$\theta \leftarrow \theta - \gamma [\nabla_\theta F_v(\theta; D_v) + \eta]$

return $\theta$
• Consider the standard objective $F(\theta; D) = \frac{1}{n} \sum_{v=1}^{n} F_v(\theta; D_v)$

• We consider a decentralized SGD algorithm where the model is updated sequentially by following a random walk, aka incremental gradient [Johansson et al., 2009]

• We focus here on the complete graph

\begin{algorithm}
Initialize $\theta \in \mathbb{R}^p$

for $t = 1$ to $T$
do

Draw random user $v \sim U(1, \ldots, n)$

$\eta = [\eta_1, \ldots, \eta_p]$, where each $\eta_j \sim \mathcal{N}(0, \sigma^2 \Delta^2)$

$\theta \leftarrow \theta - \gamma [\nabla_\theta F_v(\theta; D_v) + \eta]$

return $\theta$
\end{algorithm}
### Theorem ([Cyffers and Bellet, 2022], informal)

Let \( F_1(\cdot; D_1), \ldots, F_n(\cdot; D_n) \) be convex and smooth. Given \( \alpha > 1, \epsilon > 0 \), let \( T = \tilde{\Omega}(n^2) \) and \( \sigma^2 \) be such that private random walk-based decentralized SGD on the complete graph satisfies \((\alpha, \epsilon)\)-local RDP. Then the algorithm also satisfies \((\alpha, \ln n - \epsilon)\)-network DP.

- In other words, accounting for the limited view in decentralized algorithms allows to recover the privacy-utility trade-off of DP-SGD under central DP! (up to a log factor)

- Note: for \( T = o(n^2) \), the amplification effect is still strong and can be computed numerically, see ([Cyffers and Bellet, 2022])

- Utility analysis: same as DP-SGD!

- Privacy analysis: leverages privacy amplification by iteration ([Feldman et al., 2018]) and exploits the randomness of the walk through “weak convexity” of Rényi divergence
• Numerical results are consistent with our theory: network DP-SGD significantly amplifies privacy guarantees compared to local DP-SGD
PRIVATE GOSSIP-BASED DECENTRALIZED SGD
GOSSIP-BASED DECENTRALIZED SGD

- Random walk-based SGD is sequential (no parallel computation)

- A popular parallel alternative is gossip-based decentralized SGD [Lian et al., 2017] [Koloskova et al., 2020], which builds upon gossip averaging [Boyd et al., 2006]

- A gossip matrix over the graph $G = (\mathcal{V}, \mathcal{E})$ is a matrix $W \in \mathbb{R}^{n \times n}$ which:
  - is symmetric with nonnegative entries
  - is stochastic, i.e., $W \mathbf{1} = \mathbf{1}$
  - for any $v, w \in \mathcal{V}$, $W_{v,w} > 0$ implies $\{v, w\} \in \mathcal{E}$ or $v = w$

**Algorithm**  GOSSIP_AVERAGING($\{x_v\}_{v \in \mathcal{V}}, W, K$) [Boyd et al., 2006]

```
for all nodes $v$ in parallel do
  $x_v^0 \leftarrow x_v$
for $k = 0$ to $K - 1$ do
  for all nodes $v$ in parallel do
    $x_v^{k+1} \leftarrow \sum_{w \in \mathcal{N}_v} W_{v,w} x_w^k$, where $\mathcal{N}_v = \{w : W_{v,w} > 0\}$
return $x_1^K, \ldots, x_n^K$
```
• Consider again $F(\theta; D) = \frac{1}{n} \sum_{v=1}^{n} F_v(\theta; D_v)$ with $F_v(\theta; D_v) = \frac{1}{|D_v|} \sum_{(x_v, y_v) \in D_v} \ell(\theta; x_v, y_v)$

Algorithm  Gossip-based decentralized SGD [Lian et al., 2017, Koloskova et al., 2020]

Initialize $\theta_1^{(0)}, \ldots, \theta_n^{(0)} \in \mathbb{R}^p$

for $t = 0$ to $T - 1$ do
  for all nodes $v$ in parallel do
    $\hat{\theta}_v^t \leftarrow \theta_v^t - \gamma \nabla_\theta \ell(\theta_v^t; x_v^t, y_v^t)$ where $(x_v^t, y_v^t) \sim D_v$
    $\theta_v^{t+1} \leftarrow \text{GOSSIP}_\text{AVERAGING}(\{\hat{\theta}_v^t\}_{v \in \mathcal{V}}, W, K)$
  return $\theta_1^T, \ldots, \theta_n^T$

• Note: to improve the dependence on the topology in the convergence rate we actually use accelerated gossip [Berthier et al., 2020]
To make the algorithm private, we simply add Gaussian noise before gossiping.

**Algorithm** PRIVATE_GOSSIP_AVERAGING($\{x_v\}_{v \in V}, W, K, \sigma^2$)

for all nodes $v$ in parallel do

$\tilde{x}_v^0 \leftarrow x_v + \eta_v$ where $\eta_v \sim \mathcal{N}(0, \sigma^2)$

$x_1^K, \ldots, x_n^K \leftarrow \text{GOSSIP_AVERAGING}(\{\tilde{x}_v^0\}_{v \in V}, W, K)$

return $x_1^K, \ldots, x_n^K$

**Algorithm** Private gossip-based decentralized SGD [Cyffers et al., 2022]

Initialize $\theta_1^{(0)}, \ldots, \theta_n^{(0)} \in \mathbb{R}^p$

for $t = 0$ to $T - 1$ do

for all nodes $v$ in parallel do

$\hat{\theta}_v^t \leftarrow \theta_v^t - \gamma \nabla_{\theta} \ell(\theta_v^t; x_v^t, y_v^t)$ where $(x_v^t, y_v^t) \sim D_v$

$\theta_v^{t+1} \leftarrow \text{PRIVATE_GOSSIP_AVERAGING}(\{\hat{\theta}_v^t\}_{v \in V}, W, K, \gamma^2 \sigma^2 \Delta^2)$

return $\theta_1^T, \ldots, \theta_n^T$
Theorem ([Cyffers et al., 2022])

After $K$ iterations, Private Gossip Averaging is $(\alpha, f)$-PNDP with

$$f(u, v) = \frac{\alpha \Delta^2}{2\sigma^2} \sum_{k=0}^{K-1} \sum_{w: \{v, w\} \in \mathcal{E}} \frac{(W^k_{u,w})^2}{\|W^k_{w,:}\|^2}$$

$$\leq \frac{\alpha \Delta^2 n}{2\sigma^2} \max_{\{v, w\} \in \mathcal{E}} W^{-2}_{v,w} \sum_{k=1}^{K} \mathbb{P}(X^k = v | X^0 = u)^2,$$

where $(X^k)_k$ is the random walk on graph $G$, with transitions $W$.

- As desired, this exhibits the fact that, for two nodes $u$ and $v$, privacy guarantees improve with their “distance” in the graph
Recall central DP achieves $O\left(\frac{\alpha p \Delta^2}{n^2 \epsilon}\right)$ and local DP achieves $O\left(\frac{\alpha p \Delta^2}{n \epsilon}\right)$.

Setting the mean privacy loss $\bar{\epsilon}_v = \frac{1}{n} \sum_{u \in V \setminus \{v\}} f(u, v)$ to satisfy $\bar{\epsilon} = \max_{v \in V} \bar{\epsilon}_v \leq \epsilon$, for private gossip averaging we get (ignoring log terms):

<table>
<thead>
<tr>
<th>Graph</th>
<th>Arbitrary</th>
<th>Complete</th>
<th>Ring</th>
<th>Expander</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utility (MSE)</td>
<td>$\frac{\alpha p \Delta^2 d}{n^2 \epsilon \sqrt{\lambda_W}}$</td>
<td>$\frac{\alpha p \Delta^2}{n \epsilon}$</td>
<td>$\frac{\alpha p \Delta^2}{n \epsilon}$</td>
<td>$\frac{\alpha p \Delta^2}{n^2 \epsilon}$</td>
</tr>
</tbody>
</table>

We match the utility of central DP up to an additional $d/\sqrt{\lambda_W}$ factor, where $d$ is the max degree and $\lambda_W$ of the spectral gap of $W$.

Some graphs (e.g., expanders) make this constant: we get privacy and efficiency!

Note: we also have extensions to time-varying graphs and randomized gossip.
Theorem ([Cyffers et al., 2022])

Let $F$ be $\mu$-strongly convex, $F_v$ be $L$-smooth and $\mathbb{E}[\|\nabla \ell(\theta^*; x_v, y_v) - \nabla F(\theta^*)\|^2] \leq \rho^2_v$. Let $\bar{\rho}^2 = \frac{1}{n} \sum_{v \in V} \rho^2_v$. For any $\epsilon > 0$, and appropriate choices of $T$ and $K$, there exists $f$ such that the algorithm is $(\alpha, f)$-PNDP, with:

$$\forall v \in V, \quad \bar{\epsilon}_v = \frac{1}{n} \sum_{u \in V \setminus \{v\}} f(u, v) \leq \epsilon \quad \text{and} \quad \mathbb{E}[F(\bar{\theta}^{1:T}) - F(\theta^*)] \leq \mathcal{O} \left( \frac{\alpha p \Delta^2 d}{n^2 \mu \epsilon \sqrt{\lambda_W}} + \frac{\bar{\rho}^2}{nL} \right),$$

where $d_v$ is the degree of node $v$ and $\lambda_W$ is the spectral gap associated with $W$.

- The term $\frac{\bar{\rho}^2}{nL}$ is privacy-independent and dominated by the first term.
- The first term has the same form as before, so same conclusions apply!
- In particular, with an expander graph, we match the privacy-utility trade-off of centralized SGD with a trusted curator (up to log terms).
• Users get local DP guarantees w.r.t. their direct neighbors but stronger privacy w.r.t. to other users depending on their distance and the mixing properties of the graph

• This fits the privacy expectations of users in many use-cases (e.g., social networks)

• For learning, we can randomize the graph after each local computation step to make the privacy loss concentrate!
Take-home message

- Decentralized learning can amplify differential privacy guarantees, providing a new incentive for using such approaches beyond the usual motivation of scalability.

Perspectives

- Privacy and utility guarantees for random walk-based decentralized SGD on arbitrary graphs [Johansson et al., 2009], possibly with multiple parallel walks [Hendrikx, 2022].

- Capturing the redundancy in gossip-based communication (i.e., correlated noise) to further improve privacy guarantees (recall that even constants matter in DP!).
THANK YOU FOR YOUR ATTENTION!
QUESTIONS?
   Deep learning with differential privacy.
   In CCS.

   Algorithmic stability for adaptive data analysis.
   In STOC.

   In FOCS.

   Accelerated gossip in networks of given dimension using jacobi polynomial iterations.

   Randomized gossip algorithms.

   Membership inference attacks from first principles.
   In S&P.


A randomized incremental subgradient method for distributed optimization in networked systems.


A Unified Theory of Decentralized SGD with Changing Topology and Local Updates.
In ICML.

In NIPS.

Rényi Differential Privacy.
In CSF.

Reconstructing Genotypes in Private Genomic Databases from Genetic Risk Scores.
In International Conference on Research in Computational Molecular Biology RECOMB.