Differential Privacy Analysis of Decentralized Gossip Averaging under Varying Threat Models

Antti Koskela and Tejas Kulkarni
Nokia Bell Labs

Abstract

Fully decentralized training of machine learning models offers significant advantages in scalability, robustness, and fault tolerance. However, achieving differential privacy (DP) in such settings is challenging due to the absence of a central aggregator and varying trust assumptions among nodes. In this work, we present a novel privacy analysis of decentralized gossip-based averaging algorithms with additive node-level noise, both with and without secure summation over each node's direct neighbors. Our main contribution is a new analytical framework based on a linear systems formulation that accurately characterizes privacy leakage across these scenarios. This framework significantly improves upon prior analyses, for example, reducing the Rényi DP parameter growth from $O(T^2)$ to $O(T)$, where T is the number of training rounds. We validate our analysis with numerical results demonstrating superior DP bounds compared to existing approaches. We further illustrate our analysis with a logistic regression experiment on MNIST image classification in a fully decentralized setting, demonstrating utility comparable to central aggregation methods.

1 Introduction

Common federated learning (FL) scenarios assume the presence of a central parameter server for coordinating model updates. In contrast, fully decentralized setups operate without a central orchestrator: compute nodes, each holding a private dataset, directly exchange model states or updates with a subset of peers. Such decentralized architectures offer advantages in scalability, fault tolerance, and robustness, but also introduce new algorithmic and privacy challenges.

Fully decentralized gradient-based optimization methods are typically distributed variants of gradient descent and can be broadly classified into two categories.

In random walk-based methods (Lopes and Sayed, 2007; Johansson et al., 2009; Mao et al., 2020), a node computes local gradients and sends its model state to a randomly selected neighbor, sampled according to a doubly stochastic mixing matrix. The neighbor then updates its local parameters and passes the updated model to another randomly chosen neighbor. Over time, and under mild assumptions on the transition matrix and the network graph (e.g., connectedness, symmetry), the random walk ensures uniform coverage of all nodes. However, such sequential update dynamics inherently limits parallelism and scalability.

In gossip averaging-based methods (Boyd et al., 2006; Dimakis et al., 2010), all nodes simultaneously communicate with their neighbors in synchronous rounds, sharing either gradients or full model states. Each node updates its parameters by averaging over the received messages. These operations are repeated iteratively until the network reaches approximate consensus on the model parameters. Gossip protocols are naturally parallelizable and more scalable than random walk approaches, making them attractive for large-scale decentralized learning. This work focuses on such gossip-based protocols for DP decentralized learning.

DP in Centralized FL. In centralized FL (Kairouz et al., 2021b), distributed DP (Ullah et al., 2023; Hartmann and Kairouz, 2023; Kairouz et al., 2021a) strengthens statistical privacy guarantees by combining local noise injection with cryptographic tools such as secure summation (Truex et al., 2019; Erlingsson et al., 2019; Bell et al., 2020). This not only eliminates the need to trust individual nodes but also enables lower per-user noise by aggregating over multiple contributions.

DP in Decentralized Optimization. The increased communication and lack of central control in decentralized optimization expands the attack surface, exposing the system to more potential privacy breaches (Dekker et al., 2025; Zhu et al., 2019; Mrini et al., 2024; Pasquini et al., 2023). Early decentralized DP methods achieve privacy by locally perturbing gradients or models (Huang et al., 2015; Bellet et al., 2018; Xu et al., 2022), relying on local DP guarantees. This often imposes a poor utility-privacy tradeoff compared to centralized DP (Chan et al., 2012).

To address this, recent works propose pairwise network DP (Cyffers et al., 2024, 2022; Cyffers and Bellet, 2022), where privacy loss is analyzed between pairs of nodes. These relaxations yield better utility while accounting for the structure of decentralized communication. However, for the practical case of gossip averaging, there does not yet exist a satisfactory composition analysis for the pairwise network DP guarantees. To the best of our knowledge, the best bounds in this case are the Rényi differential privacy (RDP) bounds by Cyffers et al. (2022) which exhibit a T^2-growth of the RDP parameters, where T is the number of training iterations, making them unsuitable for practical ML model training with reasonable privacy guarantees.

Our Approach. In this work, we develop techniques to analyze network DP guarantees under various threat models, showing that the total sensitivity grows empirically as $O(\sqrt{T})$, resulting in $O(T)$ growth of the RDP parameters, for example, representing a significant improvement over prior analyses. This network DP accounting is obtained by interpreting the dynamics of the gossip averaging as a linear state-space system and the view of individual nodes or subsets of nodes as projected Gaussian mechanisms (GMs). This also provides tools to analyze the network DP guarantees in different threat scenarios. In addition to considering the threat model of (Cyffers et al., 2022), where the neighboring nodes see the plain messages sent by their neighbors, we also analyze algorithms that incorporate a secure summation protocol between neighboring nodes.

Secure Aggregation in Decentralized Learning. While the amplification effect of summation is well-understood in centralized FL, its impact in fully decentralized settings remains underexplored. In current gossip-based systems (Cyffers et al., 2022), each node receives messages directly from its neighbors and can, in principle, view their full content. As we show, summation over a node's neighborhood can yield a meaningful privacy amplification. Several studies have proposed integrating MPC primitives into decentralized learning workflows (Jayaraman et al., 2018; Lian et al., 2018; Jeon et al., 2021). Concrete protocols for decentralized secure aggregation have been developed in recent works (Sabater et al., 2022; Pereira et al., 2024; Biswas

et al., 2024). The use of secure summation for private decentralized algorithms has also been considered in specific applications such as differentially private PCA (Nicolas et al., 2024).

Our Contributions:

- We develop a novel analytical framework for evaluating differential privacy guarantees in decentralized gossip-based averaging algorithms. Our framework leverages a linear dynamical systems perspective to track sensitivity propagation through iterative updates.
- We demonstrate that the DP guarantees, for both non-adaptive and adaptive compositions-with or without secure summation-can be directly characterized using the Gaussian mechanism, with sensitivity and noise scale computed with the help of our linear system model.
- Our empirical results show that the squared sensitivity scales as T over T training rounds, significantly improving over the $O(T^2)$ growth in state-of-the-art analyses based on Rényi differential privacy (RDP).
- We validate our findings on a decentralized logistic regression task using the MNIST dataset, showing privacy-utility trade-offs comparable to those of central FL.

2 Background

We first shortly review the required technicalities on differential privacy. We then discuss the projected Gaussian mechanism central to our analysis and define the model for decentralized learning and the network DP guarantees.

2.1 Differential Privacy

An input dataset containing n data points is denoted as $D = (x_1, \dots, x_n) \in \mathcal{D}$, where $\mathcal{D}$ denotes the set of datasets of all sizes. We say that two datasets D and D' are neighbors if we get one by adding or removing one element to/from the other (denoted $D \sim D'$). We say that a mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{O}$ is (ε, δ)-DP if the output distributions for neighboring datasets are always (ε, δ)-indistinguishable.

Definition 1 (Dwork et al. 2006)**.** *Let $\varepsilon \geq 0$ and $\delta \in [0, 1]$. A randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathcal{P}(\mathcal{O})$, where $\mathcal{P}(\mathcal{O})$ denotes the set of probability distributions over the output space $\mathcal{O}$, is (ε, δ)-DP if for every pair of neighboring datasets D, D', every measurable set $O \subset \mathcal{O}$,*

$$\mathbb{P}(\mathcal{M}(D) \in O) \leq \mathrm{e}^{\epsilon}\mathbb{P}(\mathcal{M}(D') \in O) + \delta.$$

Given two distributions (P, Q), if the conditions $\mathbb{P}(P \in O) \leq \mathrm{e}^{\epsilon}\mathbb{P}(Q \in O) + \delta$ and $\mathbb{P}(Q \in O) \leq \mathrm{e}^{\epsilon}\mathbb{P}(P \in O) + \delta$ hold for every measurable set $O \subset \mathcal{O}$, we also denote $P \simeq_{(\varepsilon,\delta)} Q$.

With the hockey-stick divergence we can equivalently measure the (ε, δ)-distance of distributions. For probability distributions P and Q, and for $\varepsilon \in \mathbb{R}$, it is defined as $H_{\mathrm{e}^{\varepsilon}}(P||Q) = \int [P(t) - \mathrm{e}^{\varepsilon}Q(t)]_+$, where $[z]_+ = \max\{0, z\}$. The (ε, δ)-DP guarantees can be then given as follows.

Lemma 2 (Balle et al. 2018). *A mechanism $\mathcal{M}$ satisfies (ϵ, δ)-DP if and only if,* $\max_{D \sim D'} H_{\mathrm{e}^\varepsilon}(\mathcal{M}(X)||\mathcal{M}(X')) \leq \delta$.

The Gaussian Mechanism is a common way to achieve (ϵ, δ)-differential privacy by adding Gaussian noise to a functions output.

Definition 3 (Gaussian Mechanism). *Let $f : \mathcal{D} \to \mathbb{R}^d$ be a function with ℓ_2-sensitivity defined as $\Delta_2(f) = \max_{D,D'} \|f(D) - f(D')\|_2$, where the maximum is over all adjacent datasets D, D' (i.e., datasets differing in at most one individual's data). The Gaussian mechanism outputs*

$$\mathcal{M}(D) = f(D) + Z,$$

where $Z \sim \mathcal{N}(0, \sigma^2 I_d)$.

Running a DP training algorithm for T iterations is commonly modeled as an adaptive composition of T mechanisms such that the adversary has a view on the output of all intermediate outputs. This means that we then analyze mechanisms of the form

$$\mathcal{M}^{(T)}(D) = \big(\mathcal{M}_1(D), \mathcal{M}_2(\mathcal{M}_1(D), D), \ldots, \mathcal{M}_T(\mathcal{M}_1(D), \ldots, \mathcal{M}_{T-1}(D), D)\big).$$

The results of (Balle and Wang, 2018) give the tight (ε, δ)-guarantees for the Gaussian mechanism. Bounds for compositions follow from the fact that the Gaussian mechanism is μ-Gaussian Differentially Private for $\mu = \Delta_2/\sigma$, and from the composition results for μ-GDP mechanisms by Dong et al. (2022).

Lemma 4 (Dong et al. 2022). *Consider an adaptive composition of T Gaussian mechanisms, each with L_2-sensitivity Δ and noise scale parameter σ. The adaptive composition is (ε, δ)-DP for*

$$\delta(\varepsilon) = \Phi\left(-\frac{\varepsilon\sigma}{\sqrt{T}\cdot\Delta} + \frac{\sqrt{T}\cdot\Delta}{2\sigma}\right) - e^{\varepsilon}\Phi\left(-\frac{\varepsilon\sigma}{\sqrt{T}\cdot\Delta} - \frac{\sqrt{T}\cdot\Delta}{2\sigma}\right).$$

2.2 Projected Gaussian Mechanism and Moore–Penrose Pseudoinverse

In our results, the network DP guarantees become those of a projected Gaussian mechanism of the form

$$\mathcal{M}(D) = f(D) + AZ, \tag{2.1}$$

where $f : \mathcal{D} \to \mathbb{R}^m$, $A \in \mathbb{R}^{m \times n}$ and $Z \sim \mathcal{N}(0, \sigma^2 I_n)$ for some $\sigma > 0$. To analyze the DP guarantees of these mechanism, we will use the Moore–Penrose pseudoinverse of A.

We consider the computationally tractable definition of the Moore–Penrose pseudoinverse based on the singular value decomposition (SVD) (Golub and Van Loan, 2013).

Definition 5 (Compact SVD and Moore–Penrose Pseudoinverse)**.** *Let $A \in \mathbb{R}^{m\times n}$ be a matrix of rank r. The compact SVD (also known as the economy SVD) of A is given by:*

$$A = U_r \Sigma_r V_r^\top$$

where $U_r \in \mathbb{R}^{m\times r}$ and $V_r \in \mathbb{R}^{n\times r}$ contain the left and right singular vectors corresponding to the non-zero singular values, respectively, and $\Sigma_r \in \mathbb{R}^{r\times r}$ is a diagonal matrix containing the non-zero singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$. The Moore–Penrose pseudoinverse of A, denoted A^+, is defined via compact SVD as

$$A^+ = V_r \Sigma_r^+ U_r^\top$$

where $\Sigma_r^+ \in \mathbb{R}^{r\times r}$ is a diagonal matrix with entries $\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_r}$, i.e., $\Sigma_r^+ = \operatorname{diag}\left(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \ldots, \frac{1}{\sigma_r}\right)$.

To analyze the gossip averaging algorithms, the following technical result will play a central role.

Lemma 6. *Let $\sigma > 0$ and $A \in \mathbb{R}^{m\times n}$. Suppose $\mathcal{M}(D)$ is a projected Gaussian mechanism of the form given in Eq. (2.1). Let D and D' be two datasets such that $f(D) - f(D') \in$ Range(A), where* Range(A) *denotes the subspace spanned by the columns of A. Then, for all $\alpha \geq 0$,*

$$H_\alpha\left(\mathcal{M}(D)||\mathcal{M}(D')\right) = H_\alpha\left(\mathcal{N}\left(\left\|A^+(f(D) - f(D'))\right\|_2, \sigma^2\right)||\mathcal{N}(0,\sigma^2)\right).$$

i.e., the (ε,δ)-distance between $\mathcal{M}(D)$ and $\mathcal{M}(D')$ is upper bounded by the (ε,δ)-DP guarantee of a gaussian mechanism with L_2-sensitivity $\|A^+(f(D) - f(D'))\|_2$ and noise parameter σ.

2.3 Decentralized Learning and Network DP

We consider a connected graph $G = (V, E)$, with V and E as the nodes and edges of graph, respectively. The graph structure is encoded into an adjacency matrix $A \in \{1, 0\}^{n\times n}$, with $A_{ij} = 1$ if nodes i and j are connected by an edge and 0 otherwise. We denote the nodes via indices, i.e., the n nodes of the graph are denoted via $[n] = \{1, \ldots, n\}$. The time is divided into discrete intervals. In each round $t \in [T]$, $T \in \mathbb{Z}_+$, each node i can exchange a message with its neighboring nodes $N_i = \{j|(i, j) \in E\}$. We denote the closed neighborhood of node i by $\bar{N}_i = N_i \cup \{i\}$.

The total dataset D is partitioned into n disjoint shards with each node i holding a local dataset D_i. In decentralized learning the goal is to learn a model θ that minimizes the loss function $f(\theta) = \sum_{i=1}^n f_i(\theta)$, where $f_i(\theta)$ denotes the empirical loss of node i, $i \in [n]$, and $\theta \in \mathbb{R}^d$ denotes the model parameters. At each round $t \in [T]$, the global state variable is denoted $\theta_t \in \mathbb{R}^n$ with node-wise variables $[\theta_t]_i$, $i \in [n]$. In our analysis, we consider univariate node-wise variables. Our results are generalizable to higher dimensions via tensorization using Kronecker products.

DP requires defining a neighborhood relation for datasets. In our experiments we adopt a record-level relation which describes the protection for an individual data element in a dataset D_i of a given node $i \in [n]$, however we note that our results are applicable to any neighborhood relation. Formally, $D = \cup_{i\in[n]} D_i$ and

$D = \cup_{i\in[n]} D_i$ are adjacent datasets, denoted $D \sim D'$, if there exists $i \in [n]$ such that only D_i and D'_i differ. We use $D \sim_i D'$ to denote that D and D' differ only in the data of node i.

We define a decentralized mechanism $\mathcal{M}$ as a randomized function that takes as input the total dataset $D = \cup_{i\in[n]} D_i$ and outputs all the shared messages. What is seen by whom, is represented by the view. We denote by $\text{View}_{\mathcal{M}(D)}(\mathcal{A})$ the part of $\mathcal{M}(D)$ visible to the set of nodes $\mathcal{A}$. For a set of nodes $\mathcal{A} = \{i_1, \dots, i_{|\mathcal{A}|}\}$ we define the selector matrix $S(\mathcal{A}) \in \mathbb{R}^{|\mathcal{A}|\times n}$ as $S(\mathcal{A}) = \sum_{k=1}^{|\mathcal{A}|} e_k e_{i_k}$, where e_k is the kth standard basis vector of $\mathbb{R}^{|\mathcal{A}|}$ and e_{i_k} the i_kth standard basis vector of $\mathbb{R}^n$.

We say that the node j is (ε, δ)-DP from point of view of a subset of nodes $\mathcal{A} \subset [n]$, if $D \simeq_j D'$ and $\text{View}_{\mathcal{M}(D)}(\mathcal{A}) \simeq_{(\varepsilon,\delta)} \text{View}_{\mathcal{M}(D')}(\mathcal{A})$. We also assume that nodes are honest but curious: they adhere to the protocol, but may attempt to infer additional information from the messages they observe. Similarly to (Cyffers et al., 2022), we first assume that the noise injected by the observing nodes contributes to the DP guarantees. In Section 3.5, we show how to remove this assumption.

The Gossip matrix determines the dynamics of the averaging. If the local messages are represented by a vector $x_t \in \mathbb{R}^n$ at time t, the averaging corresponds to the iteration $\theta_{t+1} = W_t x_t$, where the gossip matrix $W_t \in \mathbb{R}^{n\times n}$. We say that $W \in \mathbb{R}^{n\times n}$ is row-stochastic if $W\mathbf{1} = \mathbf{1}$, where $\mathbf{1} = \begin{bmatrix}1 & \dots & 1\end{bmatrix}^\top$ and doubly stochastic if also $W^\top \mathbf{1} = \mathbf{1}$. We note, however, that our privacy amplification results are applicable to an arbitrary $W \in \mathbb{R}^{n\times n}$. In our analysis, we focus on time-invariant graphs $W_t = W$, however as we show in Appendix, the presented results can be generalized to time-variant graphs.

3 Privacy Analysis

3.1 Gossip Averaging as Discrete-Time Linear State-Space Dynamics

We analyze gossip averaging algorithms by viewing their dynamics as those of discrete-time linear state-space systems (see, e.g., Antoulas, 2005). In particular, we consider the systems without the feedthrough term in which case their dynamics is are described by the equations

$$\begin{aligned}\theta_{t+1} &= A_t\theta_t + B_t u_t,\\ y_t &= C_t\theta_t,\end{aligned}$$

where at each time step t, θ_t represents the state vector, the matrix A_t the state transition matrix, B_t the input matrix which describes how the control input u_t at time step t affects the state, and the observation y_t is related to the state vector θ_t by the observation matrix C.

We utilize techniques of representing the sequence of observations $y_1, \dots, y_N$ as a large linear system, where u_t's are vectorized and the global dynamics is described by a large block-lower-triangular state transition matrix determined by A_t, B_t and C_t, $t \in [N]$ (see, e.g., Ch. 4 Antoulas, 2005). In our presentation, we focus on time-invariant graphs in which case A_t, B_t and C_t are fixed.

This state-space dynamics perspective can be used to analyze different threat models for both non-adaptive and adaptive compositions. To summarize our results, when measuring the DP guarantees for the data of node j, the coefficient matrices (A, B, C) are given as follows.

1. Non-adaptive analysis with secure summation from the view of node i:

$$(A, B, C) = (W, W, e_i^\top),$$

where e_i denotes the ith standard basis (one-hot) vector in $\mathbb{R}^n$.

2. Non-adaptive analysis with secure summation and a set of colluding nodes $\mathcal{C} = \{i_1, \ldots, i_{|\mathcal{C}|}\}$:

$$(A, B, C) = \big(W, W, S(\mathcal{C})\big)$$

with the selector matrix $S(\mathcal{C})$.

3. Non-adaptive and adaptive analysis without secure summation:

$$(A, B, C) = (W, I_n, e_j).$$

4. Adaptive analysis with secure summation:

$$(A, B, C) = \big(W, W, S(\bar{N}_j)\big)$$

with the selector matrix $S(\bar{N}_j)$.

We next derive these matrices and the corresponding DP guarantees one by one.

3.2 Analysis of Non-Adaptive Compositions with Secure Summation

We first show how to analyze DP gossip averaging with secure summation protocols for non-adaptive node-wise functions of data. To this end, without loss of generality, consider the task of globally averaging node-wise streams of data. I.e., each node i, $i \in [n]$, will have a stream $x_0^i, x_1^i, \ldots$. Suppose each node adds normally distributed noise with variance σ^2 to its data point at each round. For $t \in [T]$, denote also $x_t = \begin{bmatrix} x_t^1 & \ldots & x_t^n \end{bmatrix}^\top$. Denoting the global state variable at round t with θ_t, the gossip averaging with the gossip matrix W can be written as

$$\theta_{t+1} = W(\theta_t + x_t + u_t), \tag{3.1}$$

where $u_t \sim \mathcal{N}(0, \sigma^2 I_n)$. From the recursion (3.1) it follows that from the point-of-view of node i, the view is given by $\mathrm{View}_{\mathcal{M}(D)}(\{i\}) = \begin{bmatrix} y_1 & \ldots & y_n \end{bmatrix}^\top$, where

$$\begin{aligned}
y_1 &= e_i^\top (W x_0 + W u_0) \\
y_2 &= e_i^\top (W^2 x_0 + W x_1 + W^2 u_0 + W u_1) \\
&\vdots \\
y_T &= e_i^\top (W^T x_0 + \cdots + W x_{T-1} + W^T u_0 + \cdots + W u_{T-1}).
\end{aligned}$$

From this representation we get the following.

Lemma 7. *Consider the neighboring sets of data-streams D and D' that change at most by one node's contribution (let it be node $j \in [n]$), such that each for each $t \in [T]$, where T denotes the total number of iterations, it holds that $\left|x_t^j - x_t'^j\right| \leq 1$ (this without loss of generality). Then, the node j is (ε, δ)-DP from the point-of-view of node i, where (ε, δ) is bounded by the (ε, δ)-distance between the two multivariate mechanisms*

$$\mathcal{M}(D) = \widetilde{x}_T + H_T \widetilde{u}_T \quad \text{and} \quad \mathcal{M}(D') = H_T \widetilde{u}_T,$$

where $\widetilde{u}_T \sim \mathcal{N}(0, \sigma^2 I_T)$, and

$$\widetilde{x}_T = \begin{bmatrix} e_i^\top W e_j \\ e_i^\top (W^2 + W) e_j \\ \vdots \\ e_i^\top (W^{T-1} + \cdots + W) e_j \end{bmatrix} \quad \text{and} \quad H_T := \begin{bmatrix} e_i^\top W & 0 & 0 & \dots & 0 \\ e_i^\top W^2 & e_i^\top W & 0 & \dots & 0 \\ e_i^\top W^3 & e_i^\top W^2 & e_i^\top W & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e_i^\top W^T & e_i^\top W^{T-1} & \dots & e_i^\top W^2 & e_i^\top W \end{bmatrix} \tag{3.2}$$

Applying the DP analysis of the projected Gaussian mechanism (Lemma 6) to the pair of distributions given by Lemma 7 directly gives the following result.

Theorem 8. *Consider the Gossip averaging algorithm* (3.1) *and suppose D, D' are neighboring datasets such that $D \simeq_j D'$. Let $\widetilde{x}_T$ and H_T be given as in Eq.* (3.2)*. Denote $\Delta_{j\to i}^T = \left\|H_T^+ \widetilde{x}_T\right\|_2$ where H_T^+ denotes the Moore–Penrose pseudoinverse of the matrix H_T. Then, from the point-of-view of a node i, the node j is (ε, δ)-DP, where (ε, δ) is the privacy guarantee of the Gaussian mechanism with sensitivity $\Delta_{j\to i}^T$ and noise scale σ.*

3.3 Analysis of Adaptive Compositions with Secure Summation

When analyzing adaptive compositions, the gossip averaging algorithm can be written as

$$\theta_{t+1} = W(\theta_t + x_t(\theta_t) + u_t), \tag{3.3}$$

where $u_t \sim \mathcal{N}(0, \sigma^2 I_n)$ and θ_t denote the global state of the average estimate at round t. Naturally, each element of x_t depends only on the corresponding element of θ_t, i.e., for any $i \in [n]$, $[x_t(\theta_t)]_i = [x_t([\theta_t]_i)]_i$.

Central to our analysis is the observation that in case $D \simeq_j D'$, in case of both D and D', the view of each node i in the graph is post-processing of the view of $\bar{N}_j$, the closed neighborhood of the node j, i.e., for any node $i \in [n]$, the views $\text{View}_{\mathcal{M}(D)}(\{i\})$ and $\text{View}_{\mathcal{M}(D')}(\{i\})$ are obtained from post-processing of the views $\text{View}_{\mathcal{M}(D)}\big(\bar{N}_j\big)$ and $\text{View}_{\mathcal{M}(D')}\big(\bar{N}_j\big)$, respectively. As we show, this also allows analyzing adaptive compositions.

Notice that the selector matrix $S(\bar{N}_j)$ selects the rows of W corresponding to the indices in the set of closed neighborhood $\bar{N}_j$. More specifically, if $\left|\bar{N}_j\right| = d_j$ and $\bar{N}_j = \{i_1, \dots, i_{d_j}\}$, then

$$S(\bar{N}_j) W = \begin{bmatrix} e_{i_1}^\top W \\ \vdots \\ e_{i_{d_j}}^\top W \end{bmatrix} \in \mathbb{R}^{d_j \times n}.$$

Denoting $S_j := S(\bar{N}_j)$, from the point-of-view of $\bar{N}_j$, what is observed is then given by $\mathrm{View}_{\mathcal{M}(D)}(\bar{N}_j) = \begin{bmatrix} y_1 & \dots & y_T \end{bmatrix}^\top$, where

$$\begin{aligned} y_1 &= S_j\big(W(x_0+u_0)\big) \\ y_2 &= S_j\big((W^2(x_0+u_0)+W(x_1+u_1)\big) \\ &\vdots \\ y_T &= S_j\big(W^T(x_0+u_0)\cdots+W(x_{T-1}+u_{T-1})\big). \end{aligned} \tag{3.4}$$

With a similar derivation as in the case of non-adaptive compositions, this representation gives the following result.

Theorem 9. *Consider the neighboring datasets D and D' that change at the data of node $j \in [n]$), such that each for each $t \in [T]$, where T denotes the total number of iterations, it holds that $\left|x_t^j(\theta) - x'^j_t(\theta)\right| \leq 1$ for any auxiliary variable θ. Denote*

$$\widetilde{x}_T = \begin{bmatrix} S_jWe_j \\ S_j(W^2+W)e_j \\ \vdots \\ S_j(W^{T-1}+\cdots+W)e_j \end{bmatrix} \quad \textit{and} \quad H_T := \begin{bmatrix} S_jW & 0 & 0 & \dots & 0 \\ S_jW^2 & S_jW & 0 & \dots & 0 \\ S_jW^3 & S_jW^2 & S_jW & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ S_jW^T & S_jW^{T-1} & \dots & S_jW^2 & S_jW \end{bmatrix}$$

Then, the node j is (ε,δ)-DP from the point-of-view of any other node i, where (ε,δ) is the DP guarantee of the Gaussian mechanism with sensitivity $\Delta^T_{j\to i} = \left\|H_T^+\widetilde{x}_T\right\|_2$ and noise scale σ.

We now turn to the analysis of node-level differential privacy guarantees in the absence of secure summation, where nodes can observe the plain messages sent by their neighbors. In this setting, we analyze both non-adaptive and adaptive compositions simultaneously. However, we note that for the non-adaptive case, our techniques can yield tighter bounds by incorporating the distance between the observing node and the node whose privacy is being evaluated.

3.4 Gossip Averaging without Secure Summation

In case the nodes are allowed to see the plain locally perturbed messages of their neighboring nodes, the analysis changes. Now, crucial observation is that the view of any node in the graph is post-processing of the view of the differing node only. I.e., in case $D \simeq_j D'$, then for any node $i \in [n]$, the views $\mathrm{View}_{\mathcal{M}(D)}(\{i\})$ and $\mathrm{View}_{\mathcal{M}(D')}(\{i\})$ are obtained by applying the same jth node-independent post-processing to the views $\mathrm{View}_{\mathcal{M}(D)}(\{j\})$ and $\mathrm{View}_{\mathcal{M}(D')}(\{j\})$, respectively.

To analyze the view $\mathrm{View}_{\mathcal{M}(D)}(\{j\})$, instead of the iteration (3.3), we have

$$\mathrm{View}_{\mathcal{M}(D)}(\{j\}) = \begin{bmatrix} e_j^\top\theta_1 \\ \vdots \\ e_j^\top\theta_T \end{bmatrix}$$

where

$$\theta_{t+1} = W\theta_t + x_t + u_t, \quad \theta_0 = 0, \tag{3.5}$$

Writing out the recursion as in Eq. (3.4) and following the same steps, gives the following result.

Theorem 10. *Consider the gossip averaging* (3.5) *corresponding to gossip averaging without secure summation. Denote*

$$\widetilde{x}_T = \begin{bmatrix} 1 \\ e_j^\top (W+I) e_j \\ \vdots \\ e_j^\top (W^{T-1} + \cdots + W + I) e_j \end{bmatrix} \quad \textit{and} \quad H_T := \begin{bmatrix} e_j^\top & 0 & 0 & \dots & 0 \\ e_j^\top W & e_j^\top & 0 & \dots & 0 \\ e_j^\top W^2 & e_j^\top W & e_j^\top & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e_j^\top W^{T-1} & e_j^\top W^{T-2} & \dots & e_j^\top W & e_j^\top \end{bmatrix}$$

Then, the node i is (ε, δ)-DP from the point-of-view of any node i, where (ε, δ) is the DP guarantee of the Gaussian mechanism with sensitivity $\Delta_{j\to i}^T = \left\| H_T^+ \widetilde{x}_T \right\|_2$ and noise scale σ.

3.5 Accounting for Nodes' Knowledge of Injected Noise

As stated in Section 2, we have so far assumed that the noise injected by the observing node (or subset of nodes) contributes to the DP guarantees of the other nodes. In practice, this likely has a small effect on the the DP guarantees. However, we can obtain rigorous guarantees by removing the noise terms corresponding to the observing nodes. This corresponds to removing suitable columns from the matrix H_T appearing in the DP guarantees.

For example, in case of non-adaptive compositions and secure summation, instead of Thm. 8, we get the DP guarantees for node j from the point of view node i using the sensitivity $\Delta_{j\to i}^T = \left\| \widehat{H}_T^+ \widetilde{x}_T \right\|_2$ where $\widehat{H}_T$ is the $T \times \big(T \cdot (n-1)\big)$ matrix corresponding to the $T \times (T \cdot n)$ matrix H_T of Eq. (3.2) with the ith column vector of each $T \times n$-block column removed. Similar correction can be carried out also in all the other cases. We do this correction in all our experiments.

4 Experiments

Illustration of the DP Bounds for Non-Adaptive Compositions. We first consider numerical evaluation of the DP bounds applicable for non-adaptive compositions (Thm. 8 and Thm. 10).

First, consider a synthetic Erdős–Rényi graph $G(n, p)$ with $n = 100$ and $p = 0.2$, i.e., each of the n users is connected to each other with probability p. We take the gossip matrix W to be a doubly stochastic matrix obtained via a max degree normalization, where we first scale the off-diagonal elements as $W_{ij} = A_{ij} / \max\{d_i, d_j\}$, where A denotes the symmetric adjacency matrix and d_i the degree of node i, $i \in [n]$. We then set the diagonal elements as $W_{ii} = 1 - \sum_{j, i\neq j} W_{ij}$. The left figure of Fig. 1 illustrates that in case of the secure summation, the squared sensitivity $(\Delta_{j\to i}^T)^2$ for a randomly chosen j grows linearly w.r.t. T, and in particular, approaches $\frac{T}{n}$ which corresponds to the sensitivity in the case of centralized aggregation.

Second, consider the Facebook Social Circle dataset (Leskovec and Mcauley, 2012) available in the Stanford Large Network Dataset Collection (Leskovec and Sosič, 2016), which describes an directed graph of $n = 4036$ nodes connected via binary edge weights. We consider an undirected version of this graph that we obtain via symmetrization of the adjacency matrix. We obtain a doubly stochastic gossip matrix W using the max degree normalization similarly to the synthetic case. The right figure of Fig. 1 illustrates the scaled sensitivity $(\Delta^T_{j\to i})^2/T$ for observing the node j from node i, where j is the node index 300 (an arbitrary node) and i is either one of the first two indices in the neighborhood of node j or one of the first two indices outside of the neighborhood of node j.

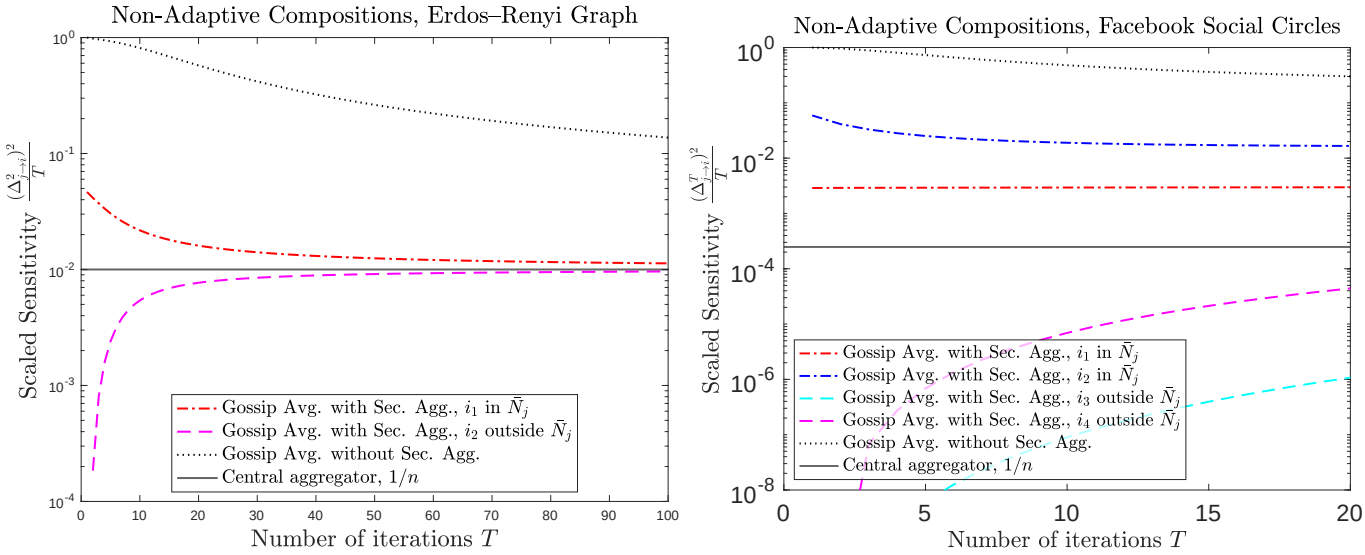

Figure 1: Left: Erdős–Rényi graph $G(n, p)$ with $n = 100$ and $p = 0.2$, Right: Facebook Social Circle dataset (Leskovec and Mcauley, 2012), and the scaled sensitivity $(\Delta^T_{j\to i})^2/T$ as a function of T. We consider separately the cases when random nodes i and j are neighbors and not. In case of the synthetic Erdős–Rényi graph, the scaled sensitivity approaches $\frac{1}{n}$ which corresponds to the scaled sensitivity in the centralized case using secure aggregation.

Illustration of the DP Bounds for Adaptive Compositions. We also consider comparing the bounds in case of adaptive compositions (Thm. 9 and Thm. 10), which corresponds to gossip averaging based ML model training via private gradient descent, for example. We consciously exclude the composition bound of (Cyffers et al., 2022, Thm. 6) in our comparisons. Their bound exhibits an $O(T^2)$ dependency in the Rényi DP parameters and becomes even looser than the standard LDP bound in our setting (when the number of Muffliato averaging steps in their algorithm is $K = 1$). We compute average the ε-values of the pair-wise DP guarantees, i.e., the average of the DP-ε's between all pairs (i, j) over the whole graph, when $\delta = 10^{-5}$. Interestingly, the bound we obtain using Thm. 10 for the averaging without secure summation becomes tighter than the bound for the averaging with secure summation (Thm. 9) as T grows. This suggests that the bound

of Thm. 9 can be improved further. Notably, both bounds represent clear improvements over the LDP bound.

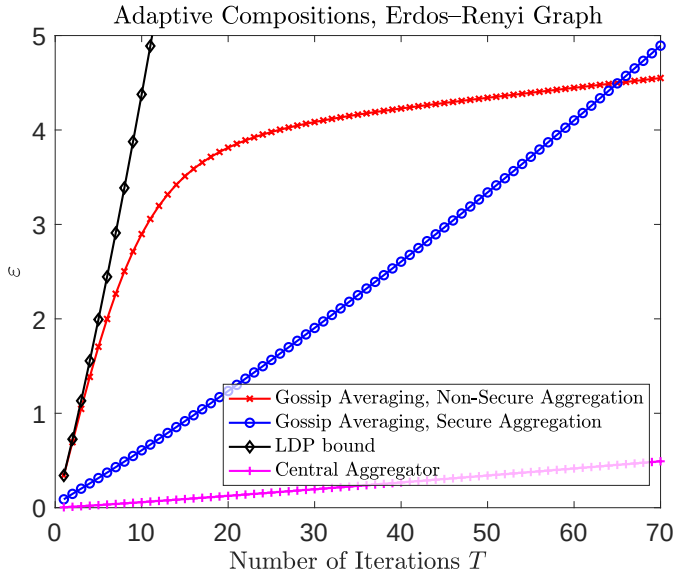

Figure 2: DP amplification for adaptive compositions over a synthetic Erdős–Rényi graph $G(n, p)$ with $n = 50$ and $p = 0.2$.

Decentralized Learning for Logistic Regression. We distribute the MNIST (LeCun, 1998) dataset IID to 100 nodes and each node is minimizing the cross-entropy loss for the logistic regression model using gradient descent. The nodes are connected based on an adjacency matrix $A \in \{0,1\}^{n\times n}$ that corresponds to a randomly drawn Erdős–Rényi graph $G(n, p)$ with $n = 100$ and $p = 0.2$. We compute the average DP-ε values of all pairs (i, j) (DP of node j from the point of view of node i) over the whole graph, when $\delta = 10^{-5}$. The messages of the nodes are locally DP updated models and nodes carry out gossip averaging over the closed neighborhoods, i.e., the row-stochastic gossip matrix $W = \Lambda^{-1}(A + I_n)$, where Λ is a diagonal matrix with $\Lambda_{ii} = \left|\bar{N}_i\right|^{-1}$, $i \in [n]$.

5 Conclusions and Outlook

In this work, we have given the first privacy amplification results for DP gossip averaging algorithms on graphs that are able to accurately capture the privacy amplification arising from all the noise injected in the system. As a future work, it will be interesting to see, whether the strong results for non-adaptive compositions that use secure summation can be translated to bounds for adaptive compositions. From computational perspective, a necessary future task will be to speed up linear algebraic subroutines for evaluating the total sensitivities for large and possibly sparse graphs. Also, the proposed framework is likely applicable to other decentralized optimization methods exhibiting a similar linear structure as the gossip averaging considered

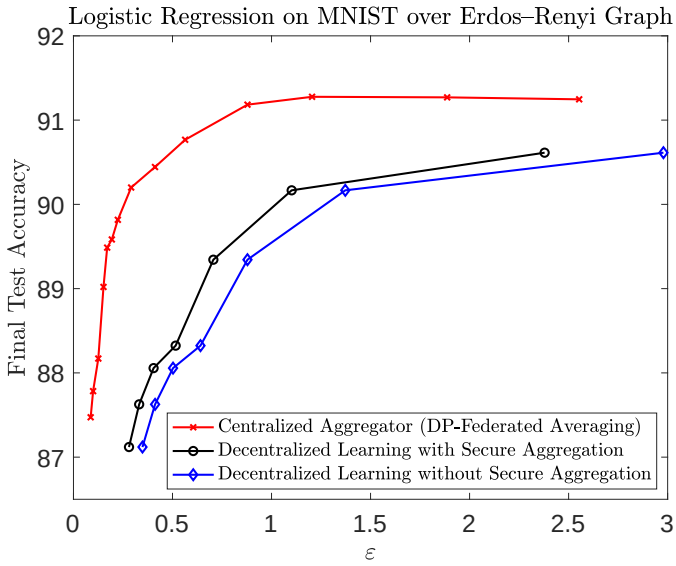

Figure 3: Decentralized learning for logistic regression on MNIST dataset over a synthetic Erdős–Rényi graph $G(n, p)$ with $n = 100$ and $p = 0.2$. The gossip matrix is the row-stochastic matrix corresponding to averaging over the closed neighborhood of each node.

here.

Bibliography

Antoulas, A. C. (2005). *Approximation of large-scale dynamical systems*. SIAM.

Balle, B., Barthe, G., and Gaboardi, M. (2018). Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *Advances in Neural Information Processing Systems*, pages 6277–6287.

Balle, B. and Wang, Y.-X. (2018). Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning*, pages 394–403.

Bell, J. H., Bonawitz, K. A., Gascón, A., Lepoint, T., and Raykova, M. (2020). Secure single-server aggregation with (poly)logarithmic overhead. *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 1253–1269.

Bellet, A., Guerraoui, R., Taziki, M., and Tommasi, M. (2018). Personalized and private peer-to-peer machine learning. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pages 473–481.

Biswas, S., Kermarrec, A., Pires, R., Sharma, R., and Vujasinovic, M. (2024). Secure aggregation meets sparsification in decentralized learning. *arXiv preprint arXiv:2405.07708*.

Boyd, S., Ghosh, A., Prabhakar, B., and Shah, D. (2006). Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530.

Chan, T. H., Shi, E., and Song, D. (2012). Optimal lower bound for differentially private multi-party aggregation. In *Algorithms - ESA 2012 - 20th Annual European Symposium*, volume 7501, pages 277–288. Springer.

Cyffers, E. and Bellet, A. (2022). Privacy amplification by decentralization. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *International Conference on Artificial Intelligence and Statistics, AISTATS 2022*.

Cyffers, E., Bellet, A., and Upadhyay, J. (2024). Differentially private decentralized learning with random walks. In *Forty-first International Conference on Machine Learning, ICML 2024*.

Cyffers, E., Even, M., Bellet, A., and Massoulié, L. (2022). Muffliato: peer-to-peer privacy amplification for decentralized optimization and averaging. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, https://arxiv.org/pdf/2206.05091*.

Dekker, F. W., Erkin, Z., and Conti, M. (2025). Topology-based reconstruction prevention for decentralised learning. *Proc. Priv. Enhancing Technol.*, 2025(1).

Dimakis, A. G., Kar, S., Moura, J. M. F., Rabbat, M. G., and Scaglione, A. (2010). Gossip algorithms for distributed signal processing. *Proceedings of the IEEE*, 98(11):1847–1864.

Dong, J., Roth, A., and Su, W. J. (2022). Gaussian differential privacy. *Journal of the Royal Statistical Society Series B*, 84(1):3–37.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Proc. TCC 2006*, pages 265–284.

Erlingsson, Ú., Feldman, V., Mironov, I., Raghunathan, A., Talwar, K., and Thakurta, A. (2019). Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2468–2479. SIAM.

Golub, G. H. and Van Loan, C. F. (2013). *Matrix computations*. JHU press.

Hartmann, F. and Kairouz, P. (2023). Distributed differential privacy for large-scale data analysis. Blog.

Huang, Z., Mitra, S., and Vaidya, N. (2015). Differentially private distributed optimization. In *Proceedings of the 16th International Conference on Distributed Computing and Networking*, ICDCN '15. Association for Computing Machinery.

Jayaraman, B., Wang, L., Evans, D., and Gu, Q. (2018). Distributed learning without distress: Privacy-preserving empirical risk minimization. In *Advances in Neural Information Processing Systems*, pages 6343–6354.

Jeon, B., Ferdous, S., Rahman, M. R., and Walid, A. (2021). Privacy-preserving decentralized aggregation for federated learning. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–6. IEEE.

Johansson, B., Rabi, M., and Johansson, M. (2009). A randomized incremental subgradient method for distributed optimization in networked systems. *SIAM J. Optim.*, 20(3):1157–1170.

Kairouz, P., Liu, Z., and Steinke, T. (2021a). The distributed discrete gaussian mechanism for federated learning with secure aggregation. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, pages 5201–5212. PMLR.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021b). Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210.

LeCun, Y. (1998). The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

Leskovec, J. and Mcauley, J. (2012). Learning to discover social circles in ego networks. *Advances in neural information processing systems*, 25.

Leskovec, J. and Sosič, R. (2016). Snap: A general-purpose network analysis and graph-mining library. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(1):1–20.

Lian, X., Zhang, W., Zhang, C., and Liu, J. (2018). Asynchronous decentralized parallel stochastic gradient descent. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3043–3052. PMLR.

Lopes, C. G. and Sayed, A. H. (2007). Incremental adaptive strategies over distributed networks. *IEEE Transactions on Signal Processing*, 55(8):4064–4077.

Mao, X., Yuan, K., Hu, Y., Gu, Y., Sayed, A. H., and Yin, W. (2020). Walkman: A communication-efficient random-walk algorithm for decentralized optimization. *IEEE Transactions on Signal Processing*, 68:2513–2528.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In Singh, A. and Zhu, J., editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR.

Mironov, I. (2017). Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275.

Mrini, A. E., Cyffers, E., and Bellet, A. (2024). Privacy attacks in decentralized learning. *Proceedings of the 41 st International Conference on Machine Learning. PMLR 235, 2024*.

Nicolas, J., Sabater, C., Maouche, M., Mokhtar, S. B., and Coates, M. (2024). Differentially private and decentralized randomized power method. *arXiv preprint arXiv:2411.01931*.

Pasquini, D., Raynal, M., and Troncoso, C. (2023). On the (In)security of Peer-to-Peer Decentralized Machine Learning . In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 418–436.

Pereira, D., Ricardo Reis, P., and Borges, F. (2024). Secure aggregation protocol based on dc-nets and secret sharing for decentralized federated learning. *Sensors*, 24(4):1299.

Ponomareva, N., Hazimeh, H., Kurakin, A., Xu, Z., Denison, C., McMahan, H. B., Vassilvitskii, S., Chien, S., and Thakurta, A. G. (2023). How to dp-fy ml: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research*, 77:1113–1201.

Sabater, C., Bellet, A., and Ramon, J. (2022). An accurate, scalable and verifiable protocol for federated differentially private averaging. *Machine Learning*, 111:4249–4293.

Truex, S., Baracaldo, N., Anwar, A., Steinke, T., Ludwig, H., Zhang, R., and Zhou, Y. (2019). A hybrid approach to privacy-preserving federated learning. In *Proceedings of the 12th ACM workshop on artificial intelligence and security*, pages 1–11.

Ullah, E., Choquette-Choo, C. A., Kairouz, P., and Oh, S. (2023). Private federated learning with autotuned compression. In *International Conference on Machine Learning, ICML 2023*, volume 202 of *Proceedings of Machine Learning Research*, pages 34668–34708. PMLR.

Xu, J., Zhang, W., and Wang, F. (2022). A(dp)^{2}2sgd: Asynchronous decentralized parallel stochastic gradient descent with differential privacy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8036–8047.

Zhu, L., Liu, Z., and Han, S. (2019). Deep leakage from gradients. In *Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*.

Zhu, Y., Dong, J., and Wang, Y.-X. (2022). Optimal accounting of differential privacy via characteristic function. *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*.

A Proof of Lemma 6

Lemma A.1. *Let $\sigma > 0$ and $A \in \mathbb{R}^{m\times n}$. Suppose $\mathcal{M}(D)$ is a projected Gaussian mechanism of the form given in Eq. (2.1). Let D and D' be two datasets such that $f(D) - f(D') \in$ Range(A), where* Range(A) *denotes the subspace spanned by the columns of A. Then, for all $\alpha \geq 0$,*

$$H_\alpha\left(\mathcal{M}(D)||\mathcal{M}(D')\right) = H_\alpha\left(\mathcal{N}\left(\left\|A^+(f(D) - f(D'))\right\|_2, \sigma^2\right)||\mathcal{N}(0,\sigma^2)\right).$$

i.e., the (ε,δ)-distance between $\mathcal{M}(D)$ and $\mathcal{M}(D')$ is upper bounded by the (ε,δ)-DP guarantee of a gaussian mechanism with L_2-sensitivity $\|A^+(f(D) - f(D'))\|_2$ and noise parameter σ.

Proof. Recall, the projected Gaussian mechanism is give as

$$\mathcal{M}(D) = f(D) + AZ,$$

where $f : \mathcal{D} \to \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$ and $Z \sim \mathcal{N}(0,\sigma^2 I_n)$ for some $\sigma > 0$.

Denote the compact SVD of A as $A = U_r\Sigma_r V_r^\top$, where r denotes the rank of A. Since the columns of U_r give a basis for the subspace Range(A) and U_r has orthonormal columns, $U_rU_r^\top$ gives a projector onto Range(A). Since $f(D) - f(D') \in$ Range(A), it holds that

$$f(D) - f(D') = U_rU_r^\top\left(f(D) - f(D')\right). \tag{A.1}$$

By using the compact SVD of A, we have that for all $\alpha \geq 0$:

$$\begin{aligned}
H_\alpha\left(\mathcal{M}(D)||\mathcal{M}(D')\right) &= H_\alpha\left(f(D) + AZ \,||\, f(D') + AZ\right)\\
&= H_\alpha\left(f(D) - f(D') + AZ \,||\, AZ\right)\\
&= H_\alpha\left(f(D) - f(D') + U_r\Sigma_r V_r^\top Z \,||\, U_r\Sigma_r V_r^\top Z\right)\\
&= H_\alpha\left(U_rU_r^\top\left(f(D) - f(D')\right) + U_r\Sigma_r V_r^\top Z \,||\, U_r\Sigma_r V_r^\top Z\right)\\
&= H_\alpha\left(U_r^\top\left(f(D) - f(D')\right) + \Sigma_r V_r^\top Z \,||\, \Sigma_r V_r^\top Z\right)\\
&= H_\alpha\left(\Sigma_r^{-1}U_r^\top\left(f(D) - f(D')\right) + V_r^\top Z \,||\, V_r^\top Z\right)\\
&= H_\alpha\left(\Sigma_r^{-1}U_r^\top\left(f(D) - f(D')\right) + \widetilde{Z} \,||\, \widetilde{Z}\right),
\end{aligned}$$

where $\widetilde{Z} \sim \mathcal{N}(0,\sigma^2 I_r)$ and where before the third last equality we have carried out multiplication from the left by $U_r^\top$ and before the second last equality we have carried out multiplication from the left by Σ_r^{-1}. The last step follows from the fact that V_r has orthonormal columns. We see that for all $\varepsilon \in \mathbb{R}$,

$$H_{\mathrm{e}^\varepsilon}\left(\mathcal{M}(D)||\mathcal{M}(D')\right) = H_{\mathrm{e}^\varepsilon}\left(\Sigma_r^{-1}U_r^\top\left(f(D) - f(D')\right) + \widetilde{Z} \,||\, \widetilde{Z}\right),$$

where the right-hand side gives the tight $\delta(\varepsilon)$ for the Gaussian mechanism with sensitivity $\left\|\Sigma_r^{-1}U_r^\top\big(f(D)-f(D')\big)\right\|_2$ and noise scale σ. Further, since V_r has orthonormal columns,

$$\begin{aligned}\left\|\Sigma_r^{-1}U_r^\top\big(f(D)-f(D')\big)\right\|_2 &= \left\|V_r\Sigma_r^{-1}U_r^\top\big(f(D)-f(D')\big)\right\|_2\\ &= \left\|A^+\big(f(D)-f(D')\big)\right\|_2\end{aligned}$$

and the claim follows. □

B Proof of Lemma 7

Lemma B.1. *Consider the neighboring sets of data-streams D and D' that change at most by one node's contribution (let it be node $j\in[n]$), such that each for each $t\in[T]$, where T denotes the total number of iterations, it holds that $\left|x_t^j - x_t'^j\right|\le 1$ (this without loss of generality). Then, the node j is (ε,δ)-DP from the point-of-view of node i, where (ε,δ) is bounded by the (ε,δ)-distance between the two multivariate mechanisms*

$$\mathcal{M}(D)=\widetilde{x}_T+H_T\widetilde{u}_T \quad \textit{and} \quad \mathcal{M}(D')=H_T\widetilde{u}_T,$$

where $\widetilde{u}_T\sim\mathcal{N}(0,\sigma^2 I_T)$, and

$$\widetilde{x}_T=\begin{bmatrix} e_i^\top W e_j\\ e_i^\top(W^2+W)e_j\\ \vdots\\ e_i^\top(W^{T-1}+\cdots+W)e_j\end{bmatrix} \quad \textit{and} \quad H_T:=\begin{bmatrix} e_i^\top W & 0 & 0 & \dots & 0\\ e_i^\top W^2 & e_i^\top W & 0 & \dots & 0\\ e_i^\top W^3 & e_i^\top W^2 & e_i^\top W & \dots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ e_i^\top W^T & e_i^\top W^{T-1} & \dots & e_i^\top W^2 & e_i^\top W\end{bmatrix}$$

Proof. We now need to bound the hockey-stick divergence

$$H_\alpha\left(\mathrm{View}_{\mathcal{M}(D)}(\{i\})\,||\,\mathrm{View}_{\mathcal{M}(D')}(\{i\})\right).$$

Recall that from the recursion (3.1) it follows that

$$\mathrm{View}_{\mathcal{M}(D)}(\{i\})=\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{bmatrix},$$

where

$$\begin{aligned} y_1 &= e_i^\top(Wx_0+Wu_0)\\ y_2 &= e_i^\top(W^2x_0+Wx_1+W^2u_0+Wu_1)\\ &\vdots\\ y_T &= e_i^\top(W^Tx_0+\cdots+Wx_{T-1}+W^Tu_0+\cdots+Wu_{T-1}).\end{aligned}$$

I.e.,

$$\mathrm{View}_{\mathcal{M}(D)}(\{i\}) = H_T\widehat{x}_T + H_T\widetilde{u}_T,$$

where $\widetilde{u}_T \sim \mathcal{N}(0, \sigma^2 I_{T\cdot n})$ and

$$\widehat{x}_T = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{T-1} \end{bmatrix}.$$

similarly,

$$\mathrm{View}_{\mathcal{M}(D')}(\{i\}) = H_T\widehat{x}'_T + H_T\widetilde{u}_T,$$

where $\widetilde{u}_T \sim \mathcal{N}(0, \sigma^2 I_{T\cdot n})$ and

$$\widehat{x}'_T = \begin{bmatrix} x'_0 \\ x'_1 \\ \vdots \\ x'_{T-1} \end{bmatrix}$$

for some $\widehat{x}'_T \in \mathbb{R}^{T\cdot n}$ such that

$$\widehat{x}_T - \widehat{x}'_T = \begin{bmatrix} e_j\Delta_1 \\ e_j\Delta_2 \\ \vdots \\ e_j\Delta_n \end{bmatrix},$$

where e_j denotes the jth standard basis vector in $\mathbb{R}^n$ and $|\Delta_i| \leq 1$ for all $i \in [T]$.

From the translational invariance of the hockey-stick divergence, we have

$$\begin{aligned} H_\alpha\left(\mathrm{View}_{\mathcal{M}(D)}(\{i\}) \,||\, \mathrm{View}_{\mathcal{M}(D')}(\{i\}))\right) &= H_\alpha\left(H_T\widehat{x}_T + H_T\widetilde{u}_T \,||\, H_T\widehat{x}'_T + H_T\widetilde{u}_T\right) \\ &= H_\alpha\left(H_T\big(\widehat{x}_T - \widehat{x}'_T\big) + H_T\widetilde{u}_T \,||\, H_T\widetilde{u}_T\right). \end{aligned}$$

By the compact SVD $H_T = U_r\Sigma_r V_r^\top$, where r denotes the rank of H_T, we have that for all $\alpha \geq 0$,

$$\begin{aligned} &H_\alpha\left(U_r\Sigma_r V_r^\top\big(\widehat{x}_T - \widehat{x}'_T\big) + U_r\Sigma_r V_r^\top\widetilde{u}_T \,||\, U_r\Sigma_r V_r^\top\widetilde{u}_T\right) \\ =&H_\alpha\left(U_r\Sigma_r V_r^\top\big(\widehat{x}_T - \widehat{x}'_T\big) + U_r\Sigma_r V_r^\top\widetilde{u}_T \,||\, U_r\Sigma_r V_r^\top\widetilde{u}_T\right) \\ =&H_\alpha\left(\Sigma_r V_r^\top\big(\widehat{x}_T - \widehat{x}'_T\big) + \Sigma_r V_r^\top\widetilde{u}_T \,||\, \Sigma_r V_r^\top\widetilde{u}_T\right) \\ =&H_\alpha\left(V_r^\top\big(\widehat{x}_T - \widehat{x}'_T\big) + V_r^\top\widetilde{u}_T \,||\, V_r^\top\widetilde{u}_T\right) \\ =&H_\alpha\left(V_r^\top\big(\widehat{x}_T - \widehat{x}'_T\big) + \widehat{u} \,||\, \widehat{u}\right) \end{aligned}$$

where $\widehat{u} \sim \mathcal{N}(0, \sigma^2 I_T)$.

Denoting $\Delta x := \widehat{x}_T - \widehat{x}'_T$, we see that $-1 \leq (\Delta x)_i \leq 1$ for all $i \in [T]$. For an upper bound of the sensitivity, we need to find the maximum of $\left\|V_r^\top \Delta x\right\|_2^2$ when Δx is in the convex and compact domain $[0,1]^T$. We see that maximizing $\left\|V_r^\top \Delta x\right\|_2^2$ is equivalent to finding the maximum of the quadratic form $(\Delta x)^\top P \Delta x$ in the convex and compact domain $[0,1]^T$ where the projector matrix $P = V_r V_r^\top$ is positive semi-definite and a projection on to the row space of H_T. Thus, the optimum is found from somewhere at the boundary, i.e., it holds that $|(\Delta x)_i| = 1$ for all $i \in [T]$. Moreover, since the rows of H_T are vectors with non-negative entries, $\left\|V_r^\top \Delta x\right\|_2^2$ is maximized when $\Delta x = \begin{bmatrix} 1 & \dots & 1 \end{bmatrix}^\top$. □

C Proof of Thm. 9

Theorem C.1. *Consider the neighboring datasets D and D' that change at the data of node $j \in [n]$), such that each for each $t \in [T]$, where T denotes the total number of iterations, it holds that $\left|x_t^j(\theta) - x'^j_t(\theta)\right| \leq 1$ for any auxiliary variable θ. Denote*

$$\widetilde{x}_T = \begin{bmatrix} S_j W e_j \\ S_j(W^2 + W)e_j \\ \vdots \\ S_j(W^{T-1} + \cdots + W)e_j \end{bmatrix} \quad \textit{and} \quad H_T := \begin{bmatrix} S_j W & 0 & 0 & \dots & 0 \\ S_j W^2 & S_j W & 0 & \dots & 0 \\ S_j W^3 & S_j W^2 & S_j W & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ S_j W^T & S_j W^{T-1} & \dots & S_j W^2 & S_j W \end{bmatrix}$$

Then, the node j is (ε, δ)-DP from the point-of-view of any other node i, where (ε, δ) is the DP guarantee of the Gaussian mechanism with sensitivity $\Delta_{j\to i}^T = \left\|H_T^+ \widetilde{x}_T\right\|_2$ and noise scale σ.

Proof. Suppose $D \simeq_j D'$. Denote $S_j = S(\bar{N}_j)$. We need to carry out the analysis for the view $\text{View}_{\mathcal{M}(D)}(\bar{N}_j) = \begin{bmatrix} y_1 & \dots & y_T \end{bmatrix}^\top$, where

$$\begin{aligned} y_1 &= S_j\big(W(x_0 + u_0)\big) \\ y_2 &= S_j\big((W^2(x_0 + u_0) + W(x_1 + u_1)\big) \\ &\vdots \\ y_T &= S_j\big(W^T(x_0 + u_0) + \cdots + W(x_{T-1} + u_{T-1})\big), \end{aligned}$$

where $[x_i]_k$ depends on the state $[\theta_i]_k$ (or could depend also on $[\theta_{i-1}]_k, [\theta_{i-2}]_k, \dots$).

The only source of randomness in $\text{View}_{\mathcal{M}(D)}(\bar{N}_j)$ is that of $u_0, \dots, u_{T-1}$. To carry out the DP analysis, we "roll out" by starting the integration of the hockey-stick divergence from the last noise vector u_{T-1}. This is similar to the composition analysis using RDP (Mironov, 2017) and using dominating pairs of distributions (Zhu et al., 2022). In particular, the proof is similar to the proof of (Thm. 10, Zhu et al., 2022).

The general case can be illustrated using with the case $T = 2$. When $T = 2$, we need to analyze the mechanism

$$\mathcal{M}(D) = \begin{bmatrix} S_j\big(W(x_0 + u_0)\big) \\ S_j\big(W^2(x_0 + u_0) + W(x_1 + u_1)\big) \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix},$$

where $[x_1]_k$ depends on the state $[\theta_1]_k$ for all $k \in [n]$.

The essential observation here is that when $D \simeq_j D'$, the the states of the nodes outside of the closed neighborhood $\bar{N}_j$ are obtained by applying the same post-processing to the views $\mathrm{View}_{\mathcal{M}(D)}(\bar{N}_j)$ and $\mathrm{View}_{\mathcal{M}(D')}(\bar{N}_j)$. More specifically, given the view of the nodes inside the closed neighborhood $\bar{N}_j$, in case the output of the first component is y_1 and when only considering the randomness of the vector $W(x_0 + u_0)$ in the nodes outside of $\bar{N}_j$, the vectors $S_j\big(W^2(x_0 + u_0)$ and $S_j\big(W^2(x'_0 + u_0)$ have the same distribution, since they consist of linear combinations of the entries of y_1 and of entries of θ_1 outside of $\bar{N}_j$ which are distributed similarly in case of D and D'.

Therefore, when starting to integrate with respect to the last source of randomness, u_1, since y_1 is observed, we can fix both $S_j\big(W^2(x_0 + u_0)$ and $S_j\big(W^2(x'_0 + u_0)$ to same value, and we can also fix all the terms in Wx_1 that are evaluated using the states of the nodes outside of $\bar{N}_j$. What remains in Wx_1 are the terms evaluated using y_1, and thus we can also carry out an adaptive analysis. We then roll out the last term to become state-independent, i.e., we have that for all $\alpha \geq 0$,

$$H_\alpha\big(\mathcal{M}(D)||\mathcal{M}(D')\big) \leq H_\alpha\big(\mathcal{M}_1(D)||\mathcal{M}_1(D')\big),$$

where

$$\mathcal{M}_1(D) = \begin{bmatrix} S_j\big(W(x_0 + u_0)\big) \\ S_j\big(W^2(x_0 + u_0) + W(e_j + u_1)\big) \end{bmatrix}$$

and

$$\mathcal{M}_1(D') = \begin{bmatrix} S_j\big(W(x_0 + u_0)\big) \\ S_j\big(W^2(x_0 + u_0) + Wu_1\big) \end{bmatrix}.$$

Switching then the order of integration, we integrate out u_0 and, by the reasoning of the proof of Lemma 7, we have that

$$H_\alpha\big(\mathcal{M}_1(D)||\mathcal{M}_1(D')\big) \leq H_\alpha\big(\mathcal{M}_2(D)||\mathcal{M}_2(D')\big),$$

where

$$\mathcal{M}_2(D) = \begin{bmatrix} S_j\big(W(e_j + u_0)\big) \\ S_j\big(W^2(e_j + u_0) + W(e_j + u_1)\big) \end{bmatrix} = \begin{bmatrix} S_jWe_j \\ S_j\big(W^2 + W\big)e_j \end{bmatrix} + \begin{bmatrix} S_jW & 0 \\ S_jW^2 & S_jW \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \end{bmatrix}$$

and

$$\mathcal{M}_2(D') = \begin{bmatrix} S_j\big(Wu_0\big) \\ S_j\big(W^2u_0 + Wu_1\big) \end{bmatrix} = \begin{bmatrix} S_jW & 0 \\ S_jW^2 & S_jW \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \end{bmatrix}.$$

In case $T = 3$, we can use the same reasoning, and similarly, given the observations y_1 and y_2, we can freeze the terms $S_j\big(W^3(x_0 + u_0)$ and $S_jW^2(x_1 + u_1)$ when integrating w.r.t. u_2. The general case also follows from this reasoning.

□

D Proof of Thm. 10

Theorem D.1. *Consider the gossip averaging* (3.5) *corresponding to gossip averaging without secure summation. Denote*

$$\widetilde{x}_T = \begin{bmatrix} 1 \\ e_j^\top (W+I)e_j \\ \vdots \\ e_j^\top (W^{T-1} + \cdots + W + I)e_j \end{bmatrix} \quad \textit{and} \quad H_T := \begin{bmatrix} e_j^\top & 0 & 0 & \dots & 0 \\ e_j^\top W & e_j^\top & 0 & \dots & 0 \\ e_j^\top W^2 & e_j^\top W & e_j^\top & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e_j^\top W^{T-1} & e_j^\top W^{T-2} & \dots & e_j^\top W & e_j^\top \end{bmatrix}$$

Then, the node i is (ε, δ)-DP from the point-of-view of any node i, where (ε, δ) is the DP guarantee of the Gaussian mechanism with sensitivity $\Delta^T_{j\to i} = \left\| H_T^+ \widetilde{x}_T \right\|_2$ and noise scale σ.

Proof. The proof goes similarly as the proof of Thm. 9 and can be illustrated using the case $T = 2$.

Given the view of the node j, in case the output of the first component is y_1 and when only considering the randomness of the vector $W(x_0+u_0)$ in the rest of the nodes, the vectors $e_j\big(W^2(x_0+u_0)$ and $e_j\big(W^2(x_0'+u_0)$ have the same distribution.

Therefore, when starting to integrate w.r.t. to the last source of randomness, u_1, since y_1 is observed, we can fix both $e_j\big(W^2(x_0+u_0)$ and $e_j\big(W^2(x_0'+u_0)$ to same value, and roll out the last term to become state-independent, i.e., we have that for all $\alpha \geq 0$,

$$H_\alpha\big(\mathcal{M}(D)||\mathcal{M}(D')\big) \leq H_\alpha\big(\mathcal{M}_1(D)||\mathcal{M}_1(D')\big),$$

where

$$\mathcal{M}_1(D) = \begin{bmatrix} e_j^\top (x_0+u_0) \\ e_j^\top \big(W(x_0+u_0) + e_j + u_1\big) \end{bmatrix}$$

and

$$\mathcal{M}_1(D') = \begin{bmatrix} e_j^\top (x_0+u_0) \\ e_j^\top \big(W(x_0+u_0) + u_1\big) \end{bmatrix}$$

Switching then the order of integration, we integrate out u_0 and, by the reasoning of the proof of Lemma 7, we have that

$$H_\alpha\big(\mathcal{M}_1(D)||\mathcal{M}_1(D')\big) \leq H_\alpha\big(\mathcal{M}_2(D)||\mathcal{M}_2(D')\big),$$

where

$$\mathcal{M}_2(D) = \begin{bmatrix} e_j^\top (e_j+u_0) \\ e_j^\top \big(W(e_j+u_0) + e_j + u_1\big) \end{bmatrix} = \begin{bmatrix} 1 \\ e_j^\top \big(W+I\big)e_j \end{bmatrix} + \begin{bmatrix} e_j^\top & 0 \\ e_j^\top W & e_j \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \end{bmatrix}$$

and

$$\mathcal{M}_2(D') = \begin{bmatrix} e_j^\top W u_0 \\ e_j^\top \big(W u_0 + u_1\big) \end{bmatrix} = \begin{bmatrix} e_j^\top & 0 \\ e_j^\top W & e_j \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \end{bmatrix}.$$

The general case follows from the same reasoning, similarly as in the proof of Thm. 9.

□

E Results for Time-Varying Graphs

We see from the proofs that all the results can be straightforwardly generalized to time-varying gossip matrices W_t, $t \in [T]$. For example, in case of the non-adaptive compositions with secure summation, we would have Theorem 7 with

$$\widetilde{x}_T = \begin{bmatrix} e_i^\top W_0 e_j \\ e_i^\top (W_1 \cdot W_0 + W_0) e_j \\ \vdots \\ e_i^\top \left(\prod_{i=0}^{T-1} W_i + e_i^\top \prod_{i=0}^{T-2} W_i + \cdots + W_0 \right) e_j \end{bmatrix}$$

and

$$H_T := \begin{bmatrix} e_i^\top W_0 & 0 & 0 & \dots & 0 \\ e_i^\top W_0 \cdot W_1 & e_i^\top W_1 & 0 & \dots & 0 \\ e_i^\top W_0 \cdot W_1 \cdot W_2 & e_i^\top W_1 \cdot W_2 & e_i^\top W_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e_i^\top \prod_{i=0}^{T-1} W_i & e_i^\top \prod_{i=0}^{T-2} W_i & \dots & e_i^\top W_{T-2} \cdot W_{T-1} & e_i^\top W_{T-1} \end{bmatrix}$$

The vector $\widetilde{x}_T$ and the matrix H_T of Theorems 9 and 10 can be modified analogously.

F Details for Experiments

F.1 Additional Details

- For each model, both when using the DP Decentralized Averaging and DP-FedAvg, we optimize the learning rate using the grid $\{10^{-i/2}\}$, $i \in \mathbb{Z}$.
- In case of the gossip averaging, the test accuracies are evaluated on a model of a randomly chosen node.
- The experiments on the logistic regression including all the learning rate tuning took around 48 hours using 8 RTX 3080 GPUs.

F.2 Baseline Method for Experiments

As a baseline algorithm, we consider the full batch version of the DP Federated Averaging algorithm depicted in Alg. 11. (Ponomareva et al., 2023; McMahan et al., 2017)

F.3 Gossip Averaging Algorithm

In the experiments, we consider the full decentralized DP gradient descent shown in Alg. 10.

Algorithm 1 Differentially Private Federated Averaging (DP-FedAvg) with Record-Level DP

Inputs: number of clients n, number of rounds T, local learning rate η, clipping norm C, noise parameter σ, initial global model w^0
for each round $t = 0$ to $T - 1$ **do**
  for each client i in parallel **do**
    Initialize local model: $w_i^t \leftarrow w^t$
    Perform DP gradient descent step on D_i, with clipping constant C, learning rate η and noise multiplier $\sigma/\sqrt{n}$ obtaining updated model w_i^t
    Compute model update: $\Delta_i^t \leftarrow w_i^t - w^t$
  end for
  Server aggregates updates:

$$\tilde{\Delta}^t \leftarrow \frac{1}{n} \sum_{i=1}^{n} \Delta_i^t.$$

  Update global model:

$$w^{t+1} \leftarrow w^t + \tilde{\Delta}^t$$

  and distribute it to clients.
end for
Output: Final global model w^T

Algorithm 2 Differentially Private Gossip Averaging (DP-GossipAvg)

Inputs: number of clients n, number of rounds T, local learning rate η, clipping norm C, noise multiplier σ, gossip matrix $W \in \mathbb{R}^{n \times n}$, initial local models $\{w_i^0\}_{i=1}^n$
for each round $t = 0$ to $T - 1$ **do**
  for each client i in parallel **do**
    Perform local DP gradient descent on D_i with clipping constant C, learning rate η and noise multiplier σ to obtain updated local model w_i^t
  end for
  for each client i in parallel **do**
    Gossip averaging:

$$w_i^{t+1} \leftarrow \sum_{j=1}^{n} W_{ij} w_j^t$$

  end for
end for
Output: Final local models $\{w_i^T\}_{i=1}^n$