SCIENCE CHINA
Information Sciences

• LETTER •

September 2026, Vol. 69, Iss. 9, 199102:1–199102:2
https://doi.org/10.1007/s11432-025-4812-9

TraceDFL: towards practical model confirmation and traitor tracking for decentralized federated learning

Shuai WANG[1], Youliang TIAN[2*], Jinbo XIONG[3*], Axin XIANG[2*], Renwan BI[4] & Jianfeng MA[5]

[1]*State Key Laboratory of Public Big Data, College of Computer Science and Technology, Guizhou University, Guiyang 550025, China*
[2]*College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China*
[3]*School of Computer Science and Mathematics, Fujian University of Technology, Fuzhou 350118, China*
[4]*School of Cyber Science and Engineering, Minjiang University, Fuzhou 350108, China*
[5]*School of Cyber Engineering, Xidian University, Xi'an 710071, China*

Received 16 September 2025/Revised 26 December 2025/Accepted 24 February 2026/Published online 13 August 2026

Citation Wang S, Tian Y L, Xiong J B, et al. TraceDFL: towards practical model confirmation and traitor tracking for decentralized federated learning. Sci China Inf Sci, 2026, 69(9): 199102, https://doi.org/10.1007/s11432-025-4812-9

Federated learning (FL) enables multiple parties to collaboratively train high-performance models without sharing raw data, yet the trained models are accessible to multiple participants, which introduces significant risks of unauthorized redistribution and intellectual property (IP) disputes. This risk has been further amplified by the rapid growth of large-scale models, where competitive pressure and high training costs have driven model development toward distributed and multi-party training architectures. In such settings, model knowledge is inherently dispersed across participants, enabling malicious parties to re-deploy or reproduce similar models after training and thereby trigger ownership conflicts. Model watermarking has therefore been widely recognized as an effective and practical solution for resolving model ownership ambiguity [1], since it embeds verifiable ownership evidence directly into the model and allows copyright claims to be validated even after deployment or leakage. Nevertheless, most existing watermarking schemes [2, 3] are developed for centralized or server-coordinated FL settings and implicitly assume the presence of a trusted authority to manage watermark embedding and verification. This assumption no longer holds in large-scale distributed training scenarios, where no single entity can be fully trusted to arbitrate ownership claims. Under such conditions, existing approaches are insufficient to prevent selfish participants from falsely asserting exclusive ownership or reselling collaboratively trained models; we refer to such behavior as a model exclusivity claim attack (MEA). Moreover, current FL watermarking frameworks typically rely on centralized control or trusted servers to distribute verification material, and they lack effective mechanisms for decentralized co-ownership verification and traitor tracing. Consequently, when the server is absent or cannot be trusted, watermark generation, commitment, and verification cannot be securely coordinated among mutually distrustful clients, rendering these methods ineffective in fully decentralized environments [4].

We propose TraceDFL, the first framework specifically designed for decentralized federated learning (DFL) that simultaneously supports model ownership verification and traitor tracing in a fully serverless setting, where all participants are co-authors and can independently verify ownership claims. TraceDFL addresses decentralized co-ownership verification by introducing a Verkle tree-based white-box watermarking mechanism that constructs compact, publicly verifiable ownership proofs with reduced verification complexity, making large-scale DFL verification feasible for resource-constrained participants. To complement this capability, TraceDFL further designs a client-driven distributed backdoor watermarking scheme that aligns naturally with decentralized training, enabling black-box ownership verification and effective traitor tracing without centralized coordination. Together, these mechanisms establish a unified and scalable IP protection framework for DFL, with detailed theoretical and experimental analysis provided in Appendixes D and E. TraceDFL is abstracted into four phases: *Gen*(), *Embed*(), *Verify*(), and *Trace*(), with Figure 1 illustrating the high-level execution process of TraceDFL.

Gen(). TraceDFL is initialized by a trusted authority (TA), which is taken offline after setup. For each client $u \in \mathcal{U}$, the generation procedure is defined as $Gen(1^\lambda) \to (\mathbf{T}_u, \mathbf{K}_u)$. TA partitions a global trigger into client-specific sub-triggers $\{\mathbf{T}_u\}$ according to client identities, enabling distributed backdoor embedding. In parallel, TA samples a secret matrix $\mathbf{A}_u \in \mathbb{R}^{M\times N}$ from a standard Gaussian distribution and derives the client identity key as $\mathbf{K}_u = \text{sign}(\mathbf{A}_u \mathbf{w}^\gamma) \in \{0,1\}^N$, where $\mathbf{w}^\gamma$ denotes the selected embedding-layer weights.

Embed(). Given $(\mathbf{T}_u, \mathbf{K}_u, \mathbf{A}_u)$, each client $u \in \mathcal{U}$ embeds dual-layer watermarks during local training. For white-box embedding, the client regularizes the selected layer weights $\mathbf{w}^\gamma$ toward its identity bits $\mathbf{K}_u$ by minimizing a hinge-like loss $\mathcal{L}_{\mathbf{K}_u}(\mathbf{w}^\gamma)$ (defined in Appendix C) jointly with the main-task loss. For black-box embedding, the client injects its local trigger set $\mathbf{T}_u$ into minibatches and optimizes the standard cross-entropy loss $\mathcal{L}_{\mathbf{T}_u} = l(y_{\mathbf{T}_u}, \mathbf{w}_u(x_{\mathbf{T}_u}))$, so that the global model preserves clean accuracy while exhibiting the target behavior on triggered in-

* Corresponding author (email: youliangtian@163.com, jbxiong@fjut.edu.cn, axxiang@gzu.edu.cn)

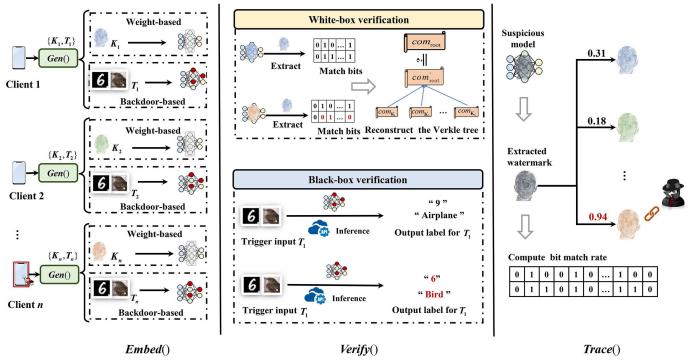

Figure 1 (Color online) The illustration of TraceDFL.

puts after aggregation. After embedding, each client publishes an identity commitment $com_{\mathbf{K}_u} = \mathrm{Commit}(\mathbf{K}_u; r_u)$ to the bulletin board, enabling the construction of a Verkle tree root commitment com_{root} as the public identifier of the co-author set.

Verify(). Given a suspicious model $\hat{\mathbf{w}}_{wm}$ with the embedding-layer weights $\hat{\mathbf{w}}^{\gamma}$, TraceDFL performs dual-layer ownership verification. For white-box verification, each client $u \in \mathcal{U}$ extracts an identity bitstring $\hat{\mathbf{K}}_u = \mathrm{sign}(\mathbf{A}_u \hat{\mathbf{w}}^{\gamma})$. Instead of using Hamming distance, which may yield ambiguous ties, we compute a continuous identity matching rate by softening each bit with $p_i = \big(1 + \exp(-\eta(\hat{\mathbf{K}}_{u,i} - \xi))\big)^{-1}$ and aggregating

$$\tau_u^K = \frac{1}{N} \sum_{i=1}^{N} \Big(p_i \cdot \mathbb{1}\{\mathbf{K}_{u,i} = 1\} + (1 - p_i) \cdot \mathbb{1}\{\mathbf{K}_{u,i} = 0\} \Big). \quad (1)$$

For black-box verification, client u queries $\hat{\mathbf{w}}_{wm}$ with its trigger set $\mathbf{X}_u$ and computes the trigger-label matching rate $\tau_u^T = \Pr\big(\mathrm{Pred}_{\hat{\mathbf{w}}_{wm}}(\mathbf{X}_u) = \mathbf{Y}_u\big)$ by comparing predicted labels with the expected backdoor labels. The suspicious model passes verification only if $\tau_u^K \geqslant \epsilon_K$ and $\tau_u^T \geqslant \epsilon_T$. Upon acceptance, each client commits the extracted watermark $com_{\hat{\mathbf{K}}_u} = \mathrm{Commit}(\hat{\mathbf{K}}_u; r_u)$, retrieves $\{com_{\mathbf{K}_v}\}_{v \in \mathcal{U}}$ from the bulletin board, and deterministically reconstructs the Verkle tree; co-ownership holds only if the locally derived root commitment matches the published com_{root}.

Trace(). To identify a malicious participant who leaked the model, TraceDFL performs client-level traitor tracing using the same identity-watermark scoring as in *Verify*(). Given a suspicious model $\hat{\mathbf{w}}_{wm}$, each client $u \in \mathcal{U}$ extracts $\hat{\mathbf{K}}_u = \mathrm{sign}(\mathbf{A}_u \hat{\mathbf{w}}^{\gamma})$ and computes its identity matching rate τ_u^K using the bit-wise confidence p_i and the aggregation rule in (1). Each client publishes τ_u^K to the public bulletin board, and the authority collects $\vec{\tau} = \{\tau_u^K\}_{u \in \mathcal{U}}$ to output the most likely leaker as $v_{\mathrm{mal}} = \arg\max_{u \in \mathcal{U}} \tau_u^K$. The detailed algorithm and steps are provided in Appendix C.

Proof sketch. We prove TraceDFL from the following properties. **Validity:** *Verify*() accepts iff $\tau_u^K \geqslant \epsilon_K$, $\tau_u^T \geqslant \epsilon_T$, and the reconstructed Verkle root matches the public com_{root}. **Fidelity:** joint optimization of the main-task loss with watermark losses preserves model utility under bounded poisoning and regularization. **Robustness:** watermark evidence remains verifiable after common model perturbations, otherwise a probabilistic polynomial-time (PPT) adversary would break commitment security. **Unforgeability:** forging a different identity key that passes verification succeeds only with negligible probability under hiding and binding commitments. **Exclusivity:** MEA fails since any exclusive-ownership claim not in the Verkle-committed co-author set requires forging membership or changing the public root. Full proofs are provided in Appendix D.

Numerical results. We benchmark TraceDFL in a simulated DFL setting across both IID and Non-IID data distributions, comparing against vanilla FedAvg (without watermarking) and representative FL watermarking baselines such as FedIPR [2] and FedTracker [3]. Results show that TraceDFL preserves main-task utility with negligible accuracy loss (within about 1%) while achieving reliable ownership verification and traitor tracing. Moreover, TraceDFL remains effective under common model modification attacks (e.g., pruning, fine-tuning, and overwriting), maintaining high watermark verifiability and leakage attribution accuracy. Scalability tests further indicate that Verkle-based verification provides lower verification latency than Merkle-based designs, enabling lightweight verification for larger client populations. Numerical results are presented in Appendix E.

Conclusion and future work. This study proposes TraceDFL, a decentralized IP protection framework for DFL that enables model ownership verification and traitor tracing without relying on a trusted server. TraceDFL combines a Verkle tree-based white-box weight watermark verification mechanism to prevent model exclusivity claim attacks with a distributed backdoor watermarking strategy for efficient embedding and practical black-box verification. Extensive experiments and formal analyses in Appendix C validate its effectiveness in ownership confirmation, traitor tracing, and robustness. Future work will explore removing or further minimizing the trusted-authority bootstrap via fully decentralized setup and on-chain commitment management, and strengthening robustness against adaptive collusion and watermark-removal attacks in large-scale, heterogeneous DFL deployments.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62272123, 62272102), Project of High-level Innovative Talents of Guizhou Province (Grant No. [2020]6008), Science and Technology Program of Guizhou Province (Grant Nos. [2020]5017, [2022]065), Natural Science Foundation Key Program of Fujian Province (Grant No. 2023J02014), Project of Science and Technology Innovation Platform of Guizhou Province (Grant No. CXPTXM[2025]024), and Minjiang University Scientific Research Startup Fund for Introduced Talents (Grant No. MJY25019).

Supporting information Appendixes A–E. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Cao X, Jia J, Gong N Z. IPGuard: protecting intellectual property of deep neural networks via fingerprinting the classification boundary. In: Proceedings of ACM Asia Conference on Computer and Communications Security (AsiaCCS), 2021
2 Li B, Fan L, Gu H, et al. FedIPR: ownership verification for federated deep neural network models. IEEE Trans Pattern Anal Mach Intell, 2022, 45: 4521–4536
3 Shao S, Yang W, Gu H, et al. FedTracker: furnishing ownership verification and traceability for federated learning model. IEEE Trans Dependable Secure Comput, 2024, 22: 114–131
4 Martínez Beltrán E T, Pérez M Q, Sánchez P M S, et al. Decentralized federated learning: fundamentals, state of the art, frameworks, trends, and challenges. IEEE Commun Surv Tut, 2023, 25: 2983–3013