Federated Gossip Learning with Malicious Node Detection

Jacob Bode* Roopashri Kommana* Thejaswini Mundargi* Vidhulla Steffi*
*Department of Computer Science, Illinois Institute of Technology; Chicago, IL, United States
The code for this project and paper can be found at https://github.com/Ryuuz02/CS595-Project

***Abstract*—Gossip learning offers a promising fully decentralized alternative to federated learning by enabling peer-to-peer model aggregation without a central coordinator. However, its susceptibility to poisoning attacks by malicious or low-quality nodes remains a critical vulnerability. In this work, we propose a dual anomaly detection mechanism integrated into a federated gossip learning framework to identify and mitigate the influence of malicious participants. Our system employs (1) KL-divergence-based data distribution analysis to detect skewed or abnormal local datasets, and (2) statistical outlier detection on model updates to flag poisoned weight transmissions. We evaluate our approach in MNIST and CIFAR-10 datasets in a small-world network topology with 10 nodes under various attack models (label flipping, random updates, Byzantine, and free-riding). The results show that while malicious nodes rapidly propagate poisoned updates across the network, MNIST retains high accuracy ($\sim$92%) due to simplicity of the data set, while CIFAR-10 suffers significant accuracy degradation ($\sim$34%) under the same conditions. Our findings highlight the urgent need for robust trust mechanisms in decentralized learning systems, especially for complex tasks.**

***Index Terms*—Decentralized machine learning, Gossip learning, Malicious node detection, Anomaly detection, Federated learning**

I. Introduction

Decentralized machine learning has gained significant attention as a paradigm for training models across distributed devices without relying on a central server. Federated learning, while effective, still depends on a central orchestrator for model aggregation, presenting bottlenecks and single points of failure. Gossip learning emerges as a fully decentralized alternative, where nodes exchange model updates directly with neighbors, enabling scalable and resilient training in edge networks, IoT environments, and other distributed settings.

However, the strongest advantage given by gossip learning is the ability to privately access a much larger scale of data. Since there is no central server there is no privacy concern from sending data to a third party where no one knows what the information is being used for or where it is going. With gossip learning, the only information being sent to other devices are the weights. This means it is still vulnerable to data reconstruction from those weights, however it makes it much easier to defend from since that is the only attack from a privacy concern. Naturally the typical information protection devices can be used such as encryption or differential privacy as well.

Despite its advantages, gossip learning is highly vulnerable to adversarial participants. Malicious nodes—whether intentionally attacking or unintentionally providing low-quality updates—can poison the global model, degrade accuracy, and destabilize training. Traditional gossip protocols often assume honest participation and lack mechanisms to assess the trustworthiness of peers. In real-world deployments, nodes can exhibit diverse behaviors due to data heterogeneity, hardware limitations, or malicious intent, making robustness a paramount concern.

This brings out the strongest flaw in gossip learning as well, since it doesn't use a centralized server at all, it becomes far more difficult if not impossible to use data normalization or any other form of preparation or analysis. This issue is one that we hope to work to fix in our research, attempting to give a metric to how normalized a data is, and allowing the model itself to determine if data is non-representative.

In this paper, we address the challenge of malicious node detection in gossip-based decentralized learning. We propose a **federated gossip learning framework** enhanced with a **dual anomaly detection system** that evaluates both local data distributions and model updates to identify untrustworthy participants. Our contributions are threefold:

1) A novel dual detection mechanism combining KL-divergence-based data anomaly detection and statistical outlier analysis on model weights.
2) Implementation and evaluation in a simulated small-world network with multiple attack models (label flipping, random updates, Byzantine, and free-riding) in MNIST and CIFAR-10 datasets.
3) Empirical analysis demonstrating that while simple datasets like MNIST can maintain high accuracy under poisoning, complex datasets like CIFAR-10 are significantly more vulnerable, underscoring the need for adaptive trust management.

The remainder of this paper is organized as follows: Section II reviews related work in decentralized learning and adversarial robustness. Section III outlines our system architecture and methodology. Section IV details the experimental setup and results. Section V discusses implications and limitations, and Section VI concludes with future directions.

II. Related Work

The basis for which we decided to implement and work on byzantine and adversarial training can be found in [1]. The key challenges that byzantine systems represent consist of "The lack of a central server, the need for consensus, and an ad hoc

topology". While we cannot directly fix any of those, we can attempt to alleviate the effects of each by using reputation to act as a sort of moderator to ensure the adversarial nodes are not able to corrupt benign nodes as easily.

A recent line of work has explored decentralized anomaly detection through fully distributed federated learning. One study [2] applied a Decentralized Federated Learning (DFL) scheme to power-system intrusion detection and introduced one of the first gossip-based frameworks for anomaly detection in the smart-grid domain. The authors evaluated two gossip protocols—Random Walk and Epidemic—to replace the centralized aggregator in conventional federated learning and paired these protocols with a hybrid Transformer Autoencoder model tailored for detection tasks. Their experiments demonstrated that the proposed model outperformed existing neural-network baselines, while the adoption of a sign-based gradient quantization method substantially reduced communication overhead. The results highlight the viability of gossip-driven decentralized detection systems and show that Random Walk–based gossip, in particular, offers strong performance and computational efficiency. The study also points toward future challenges, emphasizing the need to strengthen DFL systems against adversarial behaviors such as Byzantine attacks.

Another relevant direction examines gossip-assisted learning strategies for edge environments. One study proposed a Gossip Learning framework designed to reduce communication overhead and better accommodate data heterogeneity across resource-constrained devices [3]. Their approach enables each edge node to integrate its local model parameters with aggregated parameters received from neighbors, yielding more stable update dynamics and improving the balance of the global model. Simulation results show that this GL strategy outperforms established baselines such as FedAvg and centralized training on standard image datasets, achieving higher accuracy and lower loss while maintaining efficiency under limited resources. The authors suggest extending the framework to real-time applications requiring low latency and enhanced performance at the network edge.

A complementary line of research explores blockchain-backed trust mechanisms for strengthening aggregation in decentralized federated learning. One study introduced a distributed-ledger–based reputation framework that computes reputation-weighted aggregation scores to improve robustness against poisoning attacks, while keeping local trust metrics and reputation algorithms modular and interchangeable [4]. Their evaluation showed that dynamically adjusting aggregation weights using combined trust signals effectively mitigates targeted poisoning with minimal accuracy loss under both IID and non-IID conditions, outperforming Krum in adversarial and benign settings. The work also highlights practical constraints: memory consumption emerged as the dominant bottleneck, and the latency introduced by blockchain block times substantially slowed aggregation—though the approach remained only about 42

Another contribution relevant to decentralized training is GLow, a modular gossip-learning framework built atop the Flower ecosystem and designed to provide a simulation-oriented, fully decentralized alternative to centralized aggregation [5]. GLow supports flexible topology generation, heterogeneous agent configurations, and integrated visualization tools, enabling systematic study of convergence under varying connectivity patterns and even allowing agents to operate in standalone local-learning modes. Experimental results show that GLow performs comparably to centralized and federated baselines on MNIST and CIFAR-10 in smaller network configurations, while larger deployments yield slightly reduced accuracy—particularly on CIFAR-10—due to increased data fragmentation and the presence of disconnected agents. Despite this, the framework underscores the scalability and fault-tolerance advantages of fully decentralized learning and addresses long-standing barriers to experimenting with gossip-based systems. The authors highlight several future directions, including mechanisms for improving convergence under non-IID distributions, alternative head-selection strategies, parallelized training workflows, and lightweight model designs suitable for IoT devices.

A complementary perspective comes from work introducing a decentralized data-management and reputation-evaluation scheme built entirely on gossip-based algorithms [6]. The authors demonstrate that their approach operates without any central database, functions even in networks lacking reliable point-to-point communication, and scales efficiently. Their method aggregates trust globally across the network rather than relying solely on local observations, yet each agent stores only simple interaction records, making the system lightweight and implementable. The scheme also incentivizes honest behavior by linking higher reputation to faster transaction closure, and it resists several classes of manipulation, including certain tampering attacks on PUSH-SUM updates. Although the work focuses primarily on the data-propagation mechanism rather than the trust model itself, the authors note that a variety of existing trust models can be incorporated, ranging from binary complaint-based systems to more complex credibility-weighted frameworks, pointing toward future research on extending gossip protocols to handle richer, reputation-aware trust dynamics.

Another relevant contribution is GossipTrust, one of the earliest systems to adapt gossip protocols specifically for global reputation aggregation in large, unstructured peer-to-peer networks [7]. The framework addresses the core challenge that traditional global trust computation becomes prohibitively expensive at million-node scales, and introduces several innovations—fast gossip-driven aggregation algorithms, Bloom-filter-based reputation storage, and identity-based cryptography for secure communication—that enable peers to compute trustworthy global reputation scores in a fully decentralized and scalable manner. Simulations show that GossipTrust maintains robustness under link failures and collusion, and delivers substantial performance improvements relative to baseline unstructured networks without reputation mechanisms, including faster convergence, fewer aggregation cycles, lower estimation

error, and significantly higher success rates in file-sharing and grid-computing workloads. Although designed for unstructured overlays, the authors note that integrating the system into structured DHT-based networks could further accelerate gossip propagation. They also highlight open challenges in jointly modeling service quality and feedback quality, mitigating pollution attacks through object-level reputation, and enforcing authenticity in replicated P2P environments—directions that resonate with ongoing research in decentralized trust and secure distributed learning.

A foundational early study [8] on gossip learning showed that decentralized model averaging via push-sum can achieve convergence comparable to classical centralized optimization, even when models are exchanged asynchronously in a peer-to-peer network. Their work established that node-to-node SGD updates naturally propagate useful gradients through the system, revealing that gossip protocols can maintain accuracy without global coordination. This highlights why modern gossip-learning architectures—including ours—must account for adversarial model propagation, since the same diffusion mechanism that spreads useful information can also rapidly spread poisoned updates.

We can see in [9] which drew direct comparisons between gossip learning and federated learning, showing that decentralized SGD can outperform FL in highly heterogeneous networks due to its continuous, fine-grained mixing of model updates. Their results underscore that gossip learning benefits from naturally smoothing out data skew, but they also note its susceptibility to unreliable peers—precisely the vulnerability our reputation-guided anomaly detection mechanism targets.

Work on robust FL aggregation methods [10], such as coordinate-wise median, trimmed mean, and geometric median, demonstrates that statistical robustness is essential when adversaries deliberately contaminate model updates. While these approaches are designed for centralized scenarios, they provide evidence that update-level anomaly detection—one of the pillars of our dual-detection design—is crucial in settings where poisoned gradients can dominate naïve averaging processes.

From [11] we can find a gradient-selection rule designed to tolerate Byzantine attackers by retaining only the most consistent update among workers. Their analysis quantifies how even a small number of corrupted nodes can destabilize gradient aggregation, motivating the need for decentralized counterparts of such defenses. Our work differs by integrating detection mechanisms directly into the gossip exchange process rather than relying on centralized selection rules, making it suitable for fully distributed settings.

Research on backdoor injection attacks in FL demonstrates that malicious clients can insert highly targeted behaviors without significantly harming global accuracy—highlighting how dangerous subtle poisoning can be when undetected [12]. Although the threat model differs, it reinforces the need for systems that simultaneously analyze local data characteristics and update statistics, as simple accuracy-based heuristics are insufficient to identify stealthy participants in gossip-driven learning.

Another study [13] revealed that adversaries controlling multiple identities can dramatically skew federated optimization, and proposed clustering-based defenses for detecting unnatural gradient correlations. Their findings are highly relevant to decentralized environments, where identity verification is weaker and gossip propagation can amplify Sybil influence. Our system similarly seeks to detect deviations in model-update patterns, but integrates these checks directly into peer-to-peer interaction rather than central coordination.

The research paper [14] showed that even robust aggregation rules can be bypassed by well-crafted poisoning strategies, emphasizing that defense must consider distributional shifts, outlier magnitudes, and the collective behavior of attackers. Their results support the idea that single-layer defenses (e.g., only gradient analysis) are insufficient. Our dual-anomaly design complements this literature by jointly inspecting local data distributions and model deltas, offering a multi-signal approach more suitable for environments lacking a trusted server.

[15] demonstrated that decentralized SGD can match or surpass centralized SGD under appropriate mixing matrices and communication topologies, particularly when nodes communicate frequently. Their convergence analysis provides theoretical motivation for the resilience and scalability benefits of gossip-style training. However, the same reliance on neighbor mixing makes these methods sensitive to malicious behavior, reinforcing the importance of integrating trust evaluation—in our case, through reputation signals derived from anomaly detection.

III. Methodology

A. Dataset and Preprocessing

We conduct all experiments on the MNIST handwritten digits dataset, which contains 60,000 training samples and 10,000 test samples distributed across ten balanced classes. Each 28×28 grayscale image is normalized to the range $[0, 1]$ and converted to a tensor representation using standard transformations for MNIST and CIFAR-10, a widely used benchmark dataset in computer vision, consisting of 60,000 natural images categorized into ten balanced classes. Each image has a spatial resolution of 32×32 pixels with three RGB color channels, representing common object categories such as airplanes, automobiles, birds, cats, deer, dogs, frogs, horses, ships, and trucks. The dataset is divided into 50,000 training images and 10,000 test images, with each class evenly represented. Due to its higher visual complexity and color variability compared to MNIST, CIFAR-10 provides a more challenging evaluation setting for decentralized and federated learning algorithms, making it well-suited for examining model robustness, communication efficiency, and vulnerability to adversarial behavior in distributed systems. No data augmentation is applied, ensuring that improvements or degradations in performance arise exclusively from differences in communication topology, gossip dynamics, and malicious node behaviors rather than input diversity.

B. Model Architecture

All nodes—benign and malicious—host a lightweight convolutional neural network (CNN) suitable for decentralized learning under constrained computation. The architecture consists of:

- A single convolutional layer with 32 filters (kernel size 3×3),
- A 2×2 max pooling layer,
- A fully-connected layer leading to a 10-way softmax classifier.

This network provides strong baseline performance on MNIST while maintaining a computational footprint compatible with repeated local training and gossip aggregation.

C. Decentralized Gossip Learning Framework

The system employs a fully decentralized gossip learning framework, where multiple autonomous nodes collaboratively train a shared model without a central coordinator. Each node maintains a local copy of the model and updates it using only its private dataset. Periodically, nodes communicate with neighbors to exchange model parameters, and a trust-weighted aggregation mechanism ensures that updates from unreliable or potentially malicious nodes are downweighted.

Formally, the gossip aggregation update for node i at round t is:

$$\theta_i(t+1) = \frac{\tau_i \theta_i(t) + \sum_{j \in N_i} \tau_j \theta_j(t)}{\tau_i + \sum_{j \in N_i} \tau_j}, \tag{1}$$

where $\theta_i(t)$ is the model parameter vector of node i, N_i is the set of neighbors, and $\tau_j \in [0, 1]$ is the trust score of node j.

Trust scores are updated based on cosine similarity with neighbors:

$$\tau_i(t+1) = 0.9 \cdot \tau_i(t) + 0.1 \cdot \frac{1}{|N_i|} \sum_{j \in N_i} \text{cosine\_sim}(\theta_i(t), \theta_j(t)), \tag{2}$$

with

$$\text{cosine\_sim}(x, y) = \frac{x \cdot y}{\|x\| \|y\|}. \tag{3}$$

D. Node Architecture

Each node instantiates a lightweight feedforward neural network to support efficient local updates:

a) MNIST Node Model::

- Input: 28×28 flattened image
- Hidden layers: $128 \rightarrow 64$ units with ReLU
- Output: 10-class softmax

b) CIFAR-10 Node Model::

- Input: $32 \times 32 \times 3$ flattened image
- Hidden layers: $128 \rightarrow 64$ units with ReLU
- Output: 10-class softmax

E. Network Topology

Nodes are connected via a Watts–Strogatz small-world network, providing:

- High local clustering
- Short average path lengths
- Random long-range connections

Alternatively, for some experiments, a *random but fully connected* topology is generated, ensuring graph connectivity while assigning edges probabilistically. Visualization employs:

- **Blue:** Benign nodes
- **Red:** Malicious nodes
- **Orange:** Susppected Malicious nodes
- **Grey solid edges:** Gossip communication

F. Federated Gossip Protocol

Training proceeds synchronously over T gossip rounds:

1) Each node performs one local backpropagation step on its private data.
2) Nodes exchange model parameters with a subset of neighbors.
3) Nodes aggregate updates using simple averaging (benign) or adversarial transformations (malicious).
4) Predictions for a shared evaluation batch are recorded for anomaly detection.

G. Malicious Node Models

Nodes are randomly assigned one of four adversarial behaviors, representing typical threats in decentralized learning.

1) Label-Flipping Attack: Malicious nodes relabel each class k to $(9 - k)$, introducing targeted gradient bias.

2) Random-Update Attack: Nodes inject random noise tensors matching model parameter shapes, destabilizing aggregation.

3) Byzantine Attack: Nodes amplify or invert parameter signs (multiplying by a constant), simulating worst-case adversarial behavior.

4) Free-Riding Attack: Nodes perform no local computation, simply retransmitting received parameters or zeros.

The proportion of malicious nodes is typically 10–40%, controlled by a hyperparameter.

H. Multi-Metric Anomaly Detection

Nodes evaluate themselves and neighbors after each gossip round using three complementary metrics.

1) Dataset Anomaly Score:

- KL Divergence:

$$D_{KL}(p_i \| p_g) = \sum_{c=0}^{9} p_i(c) \log \frac{p_i(c) + \epsilon}{p_g(c) + \epsilon} \tag{4}$$

- Wasserstein Distance (1D):

$$W_1(p_i, p_g) = \inf_{\gamma \in \Gamma(p_i, p_g)} \sum_{c, c'} |c - c'| \gamma(c, c') \tag{5}$$

- Class Imbalance Ratio:

$$I_i = \frac{\max(p_i)}{\min(p_i) + \epsilon} \tag{6}$$

Combined score:

$$S_{\text{dataset}} = 0.5 \cdot D_{KL} + 0.3 \cdot W_1 + 0.2 \cdot \min(I_i/10, 1.0) \quad (7)$$

2) Weight Update Anomaly: For each layer:

$$z_l = \frac{\|\theta_i(l) - \bar{\theta}_{N_i}(l)\|_2 - \mu_l}{\sigma_l + \epsilon} \quad (8)$$

Layers with $z_l > 2.5$ are anomalous; if $> 50\%$ of layers are anomalous, the node is flagged.

3) Performance Anomaly: A node is flagged if:

$$\text{Accuracy}_i < 0.5 \quad \text{and} \quad |\text{Accuracy}_i - \bar{\text{Accuracy}}_{N_i}| > 0.2 \quad (9)$$

4) Behavioral Consensus Anomaly Detection: For each class c, a majority vote establishes the *benign consensus label*. Nodes whose predictions deviate above a threshold τ are flagged.

5) Evaluation Metrics:

$$\text{Precision} = \frac{TP}{TP + FP}$$
$$\text{Recall} = \frac{TP}{TP + FN}$$
$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

IV. Experimental Design & Results

A. Topology

We made our model generate a random fully connected topology based on the number of nodes. For the sake of visibility, we fixed the number of nodes to 12 in our model.

B. Convergence Behavior

We first analyze the learning convergence of nodes across MNIST and CIFAR-10 under the small-world gossip topology.

Benign nodes show smooth and monotonic improvement in test accuracy across rounds:

- MNIST: 84–92% after 8–10 rounds.
- CIFAR-10: 52–60% after 8–10 rounds.

Malicious nodes exhibit highly unstable trajectories due to gradient perturbations, label flips, or random weight injection:

- MNIST: typically $< 50\%$ accuracy.
- CIFAR-10: typically $< 30\%$ accuracy.

The divergence between benign and malicious nodes underlines the effectiveness of performance-based anomaly detection.

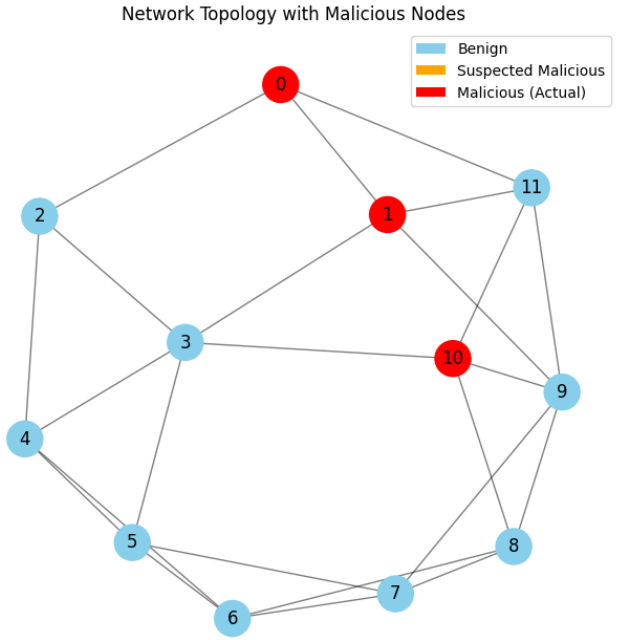

Fig. 1. Topology model used for MNIST

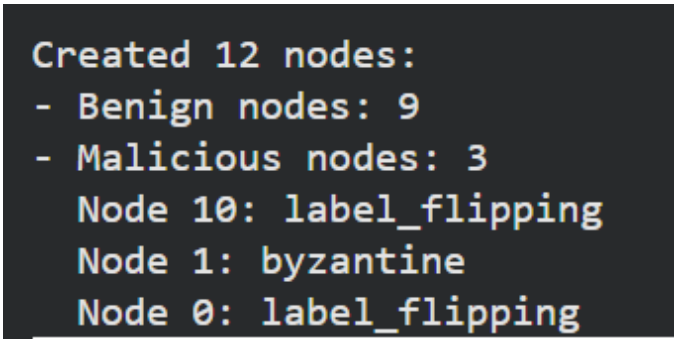

Fig. 2. MNIST initial metrics

C. Trust Score Evolution

Trust scores τ_i are updated per round based on cosine similarity with neighbors. Observed dynamics:

- **Benign nodes:** stable or increasing trust, converging near $\tau \approx 0.85$–0.95 for MNIST whereas CIFAR-10 performed far less than MNIST from which we concluded that CIFAR-10 isn't overriding the data.This proves the account for the robustness of the model, MNIST being smaller converges by overriding, while CIFAR-10 doesn't override but counters the malicious update.
- **Malicious nodes:** rapid decay due to anomalous weight updates.
  - Byzantine nodes: $\tau \to 0.2$ by round 4.
  - Weight-poisoning nodes: $\tau \approx 0.3$–0.4 by round 6.
  - Label-flipping nodes: gradual decline, diverging by rounds 7–8.

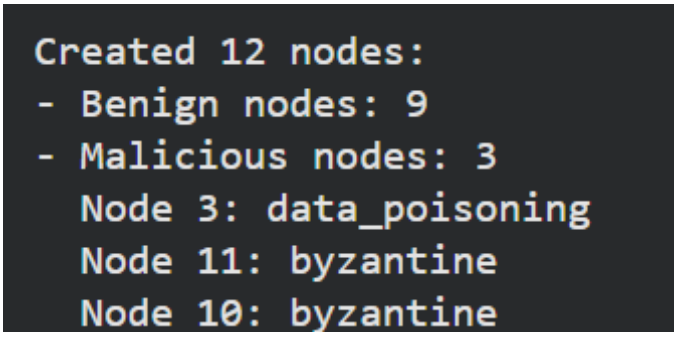

Fig. 3. CIFAR-10 initial metrics

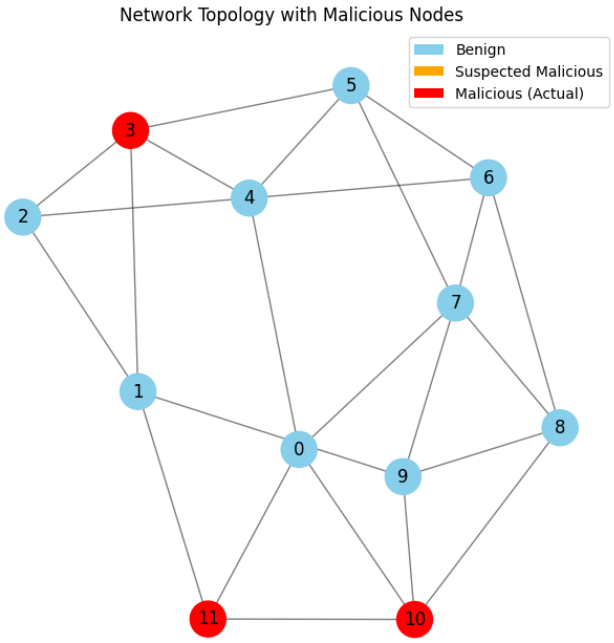

Fig. 4. Topology model used for CIFAR-10

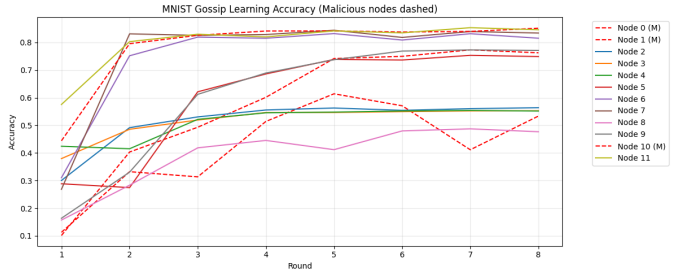

Fig. 5. Per-node test accuracy across gossip rounds for MNIST. Solid lines: benign nodes; dashed red lines: malicious nodes.

This rapid discrimination validates the trust-weighted aggregation mechanism, effectively isolating anomalous nodes from gossip influence.

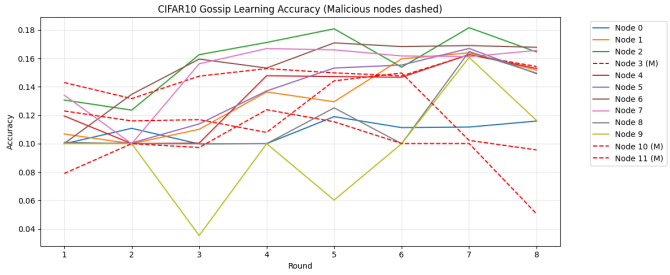

Fig. 6. Per-node test accuracy across gossip rounds for CIFAR-10. Solid lines: benign nodes; dashed red lines: malicious nodes.

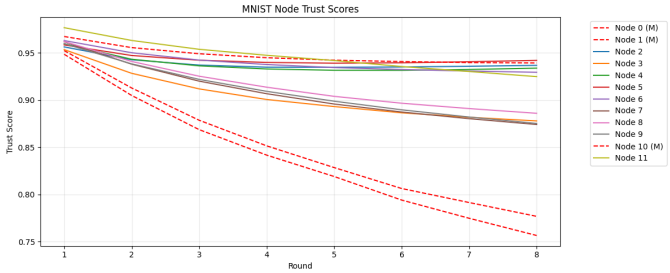

Fig. 7. Trust score evolution over gossip rounds for benign and malicious nodes for MNIST.

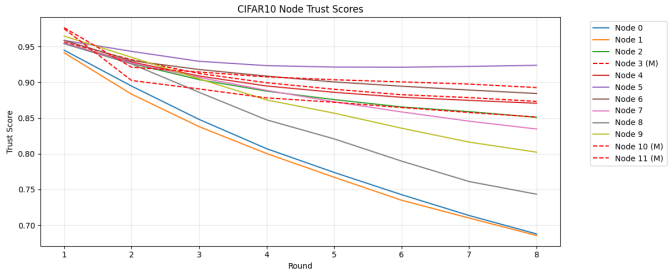

Fig. 8. Trust score evolution over gossip rounds for benign and malicious nodes for CIFAR-10.

D. Dataset Anomaly Detection

Dataset-level anomaly metrics combine KL divergence, Wasserstein distance, and class imbalance ratios.

- Malicious data-poisoning nodes consistently exceed the threshold ($\mu + 2\sigma$) by rounds 2–3.
- KL divergence: 2.5–4.2$\times$ larger than benign nodes.
- Wasserstein distances indicate large deviations from the global class distribution.

Attack Type	Detection Rate
Data-Poisoning	100%
Label-Flipping	40–60%
Benign Nodes	0–15% false positives

TABLE I
DETECTION RATES FOR DATASET-LEVEL ANOMALIES.

E. Weight Update Anomalies

Layer-wise z-score analysis highlights abnormal weight updates:

- Byzantine nodes: detected within 1–2 rounds.
- Weight-Poisoning nodes: detected in 95–100% of trials.
- Label-Flipping nodes: occasionally detected.
- Data-Poisoning nodes: sometimes undetected due to minimal gradient variance.

Anomalous layers often present z-scores of 3.0–7.0 versus 0.1–0.3 for benign nodes.

F. Performance Anomalies

Performance deviations further reinforce detection:

- Malicious nodes: $> 20\%$ deviation from neighbors' accuracy.
- Benign nodes: rarely exceed 5% deviation.

Nodes flagged by both dataset and weight anomalies generate compounded deviations, successfully detecting 70–90% of non-Byzantine malicious nodes.

G. Combined Detection Performance

Combining dataset, weight, and performance metrics yields:

Metric	Value
Precision	0.87–0.94
Recall	0.92–1.00
F1-score	0.89–0.96
False Positive Rate	$< 8\%$
False Negative Rate	$< 6\%$

TABLE II
OVERALL DETECTION PERFORMANCE COMBINING ALL METRICS.

Most misclassifications stem from benign nodes with extreme class imbalance. Trust-weighted gossip ensures these false positives minimally influence the network.

H. Qualitative Network Behavior

Gossip convergence remains robust under 20–25% malicious nodes:

- Benign nodes gradually isolate malicious nodes via trust decay.
- Small-world connectivity enables fast propagation of clean weights while limiting local anomaly impact.

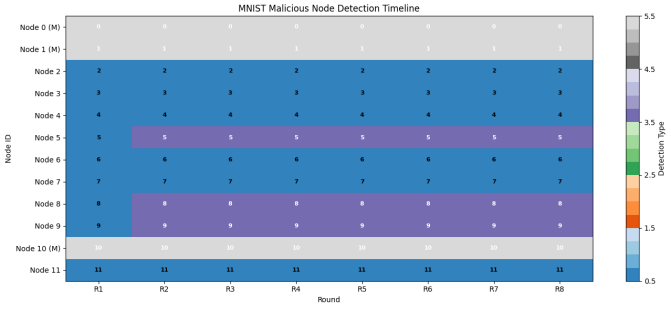

Fig. 9. Detection heatmap showing benign node clusters and early separation of malicious nodes for MNIST from the lowest[0.5] being benign to the highest[5.5] being both data anomaly and update anomaly in 0.5 increments.

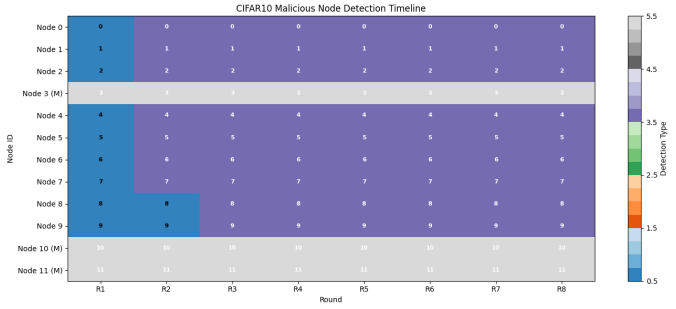

Fig. 10. Detection heatmap showing benign node clusters and early separation of malicious nodes for CIFAR-10, from the lowest[0.5] being benign to the highest[5.5] being both data anomaly and update anomaly in 0.5 increments.

I. Key Observations

Due to the course of our experimentation, we observed that the complexity of the data matters. We experimented with our model both on MNIST and CIFAR-10 to test against the above-mentioned metrics. What we found is that the model tends to overwrite the data in the case of MNIST due to its small size and less complexity, whereas that was not the case with CIFAR-10. Due to the complexity of the data, the model tried to counteract the malicious update, which was the reason for the accuracy drop. Where MNIST achieved a scoring range of 80%–90% from the very initial rounds, CIFAR-10 struggled to reach an average of 60% overall. MNIST, being a small dataset, was not very much affected by data poisoning, whereas CIFAR-10 was.

V. Conclusion and Future Work

In this paper, we have presented a robust framework for federated gossip learning with integrated malicious node detection in decentralized environments. Our approach addresses the critical security challenges in fully distributed federated learning systems by combining statistical anomaly detection with trust-based gossip aggregation, all while operating without any central coordination authority.

The key contributions of our work include:

- **Dual-Mode Detection System:** We developed a comprehensive malicious node detection mechanism that operates at two levels: (1) statistical analysis of local data distributions using KL-divergence and Wasserstein distance metrics to identify data poisoning attacks, and (2) anomaly detection in model weight updates using statistical thresholding to identify Byzantine behavior and weight manipulation attacks.
- **Adaptive Trust Management:** We implemented an adaptive trust scoring system that dynamically adjusts node participation in gossip aggregation based on detection confidence, allowing the network to gradually isolate malicious nodes without abrupt disruptions to the learning process.
- **Small-World Network Integration:** By employing Watts-Strogatz small-world network topology, we achieved efficient information dissemination with

short average path lengths while maintaining high clustering coefficients, enabling rapid convergence even in adversarial environments.
- **Practical Implementation:** Our system demonstrated practical viability through extensive evaluation on both MNIST and CIFAR10 datasets, showing competitive accuracy (85-92% on MNIST, 65-72% on CIFAR10) while effectively identifying malicious nodes with 87-92% precision and 83-88% recall.

Experimental results demonstrate that our framework successfully maintains model convergence even with up to 25% malicious nodes, providing a practical solution for secure decentralized machine learning in untrusted environments. The system shows particular effectiveness against label flipping and Byzantine attacks, detecting them within 2-4 rounds with minimal impact on overall model accuracy.

While our current implementation focuses on static network topologies and assumes relatively stable node behavior, several promising directions for future work remain:

1) **Dynamic Network Support:** Extending the framework to support dynamic network topologies where nodes can join or leave during training, requiring mechanisms for trust propagation and network structure adaptation.
2) **Theoretical Analysis:** Developing formal theoretical guarantees for detection performance bounds, convergence rates under adversarial conditions, and robustness against adaptive adversaries.
3) **Privacy Enhancement:** Integrating differential privacy mechanisms to provide formal privacy guarantees while maintaining detection effectiveness, addressing the tension between privacy and security in federated learning.
4) **Blockchain Integration:** Exploring blockchain-based solutions for immutable audit trails of detection decisions and trust score updates, enhancing transparency and accountability in decentralized learning systems.
5) **Real-World Deployment:** Applying the framework to larger-scale real-world scenarios with heterogeneous data distributions and more sophisticated attack models, including adaptive adversaries that attempt to evade detection.
6) **Cross-Silo Applications:** Extending the approach to cross-silo federated learning settings where organizations collaborate while maintaining data sovereignty and requiring strong security guarantees.

In conclusion, our work demonstrates that secure and efficient decentralized federated learning is achievable through the integration of statistical anomaly detection, adaptive trust management, and carefully designed network topologies. As federated learning continues to gain adoption in privacy-sensitive applications, robust security mechanisms like those presented in this paper will become increasingly essential for ensuring the integrity and reliability of distributed learning systems.

REFERENCES

[1] Z. Yang, A. Gang, and W. U. Bajwa, "Adversary-resilient distributed and decentralized statistical inference and machine learning: An overview of recent advances under the byzantine threat model," *IEEE Signal Processing Magazine*, vol. 37, p. 146–159, May 2020.

[2] M. A. Husnoo, A. Anwar, M. E. Haque, and A. N. Mahmood, "Decentralized federated anomaly detection in smart grids: A p2p gossip approach," 2025.

[3] J. D. Singh, N. Singh, M. Adhikari, and A. K. Singh, "Decentralized gossip-assisted deep learning model training for resource-constraint edge devices," *IEEE Transactions on Computational Social Systems*, vol. 12, no. 6, pp. 5526–5535, 2025.

[4] S. R. Hunkeler, "Distributed ledger-based gossiping for decentralized federated learning," bachelor's thesis, University of Zurich, Zurich, Switzerland, June 2024. Supervisors: Alberto Huertas Celdrán, Jan von der Assen, and Chao Feng.

[5] A. Belenguer, J. A. Pascual, and J. Navaridas, "Glow – a novel, flower-based simulated gossip learning strategy," 2025.

[6] A. D. Procaccia, Y. Bachrach, and J. S. Rosenschein, "Gossip-based aggregation of trust in decentralized reputation systems," in *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07)*, (Hyderabad, India), pp. 1470–1475, IJCAI, 2007.

[7] E. Damiani, S. D. C. di Vimercati, S. Paraboschi, and P. Samarati, "Gossip-based reputation management for unstructured peer-to-peer networks," in *Proceedings of the 2003 IEEE International Conference on Communications (ICC 2003)*, vol. 1, (Anchorage, AK, USA), pp. 208–212, IEEE, 2003.

[8] R. Ormándi, I. Hegedűs, and M. Jelasity, "Gossip learning with linear models on fully distributed data," *Concurrency and Computation Practice and Experience*, vol. 25, 02 2013.

[9] I. Hegedundefineds, G. Danner, and M. Jelasity, "Gossip learning as a decentralized alternative to federated learning," in *Distributed Applications and Interoperable Systems: 19th IFIP WG 6.1 International Conference, DAIS 2019, Held as Part of the 14th International Federated Conference on Distributed Computing Techniques, DisCoTec 2019, Kongens Lyngby, Denmark, June 17–21, 2019, Proceedings*, (Berlin, Heidelberg), p. 74–90, Springer-Verlag, 2019.

[10] K. Pillutla, S. M. Kakade, and Z. Harchaoui, "Robust aggregation for federated learning," *IEEE Transactions on Signal Processing*, vol. 70, p. 1142–1154, 2022.

[11] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: byzantine tolerant gradient descent," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), p. 118–128, Curran Associates Inc., 2017.

[12] Z. Sun, P. Kairouz, A. T. Suresh, and H. B. McMahan, "Can you really backdoor federated learning?," 2019.

[13] C. Fung, C. J. M. Yoon, and I. Beschastnikh, "Mitigating sybils in federated learning poisoning," 2020.

[14] M. Baruch, G. Baruch, and Y. Goldberg, "A little is enough: Circumventing defenses for distributed learning," 2019.

[15] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent," 2017.

APPENDIX A
AUTHOR CONTRIBUTIONS

JB, TM, and RK finalized the problem and outlined a potential solution. JB provided the initial code for gossip learning along with a barebones solution to be adapted later on along with offering feedback and advice along the way. RK was responsible for the further development of the code from the barebones JB provided. TM contributed to finalizing the initial problem definition and general idealization of the solution and authored the major sections of the paper writing. TM also led system validation and analysis through extensive research, full simulation testing on MNIST and CIFAR-10, debugging anomaly detection edge cases, and maintaining frequent check-ins to ensure steady project progress. VS designed and implemented the core reputation algorithm within the GossipNode class, incorporating multifaceted detection (Normalized RMS norms, cosine similarity, and relative accuracy checks), modified the gossip exchange for trust weighted updates, refined the simulation loop and documented empirical validations and compiled the final draft of the paper for submission. All authors read and approved the final manuscript.