ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10094739

ROBUST NETWORK TOPOLOGIES FOR DISTRIBUTED LEARNING

Chutian Wang and Stefan Vlaski

Department of Electrical and Electronic Engineering, Imperial College London

ABSTRACT

The robustness of networks against malicious agents is a critical issue for their reliability in distributed learning. While a significant number of works in recent years have investigated the development of robust algorithms for distributed learning, few have examined the influence and design of the underlying network topology on robustness. Robust schemes for distributed learning typically require certain conditions on the arrangement of malicious agents in the network. In particular, the majority of neighbors of any benign agent must be benign, and the subgraph of benign agents must be connected. In this work, we propose a scheme for the design of such topologies based on prior information of the risk profile of participating agents. We show that the resulting topology is asymptotically almost surely connected and benign agents have majority benign neighborhoods. At the same time, the proposed design asymptotically tolerates a fraction of malicious agents arbitrarily close to one, while risk agnostic designs, such as complete graphs, break down as soon as the majority of agents is malicious.

***Index Terms*—** Networked systems, decentralized learning, federated learning, diffusion algorithms, robustness.

1. INTRODUCTION

Distributed learning architectures allow a collection of intelligent agents to collaborate on solving a learning task by exchanging processed models in lieu of raw data. Formally, most learning problems can be formulated as stochastic optimization problems, taking the form [1–4]:

$$\min_{w} \sum_{k=1}^{K} p_k J_k(w) \tag{1}$$

Here, $J_k(w)$ represents the risk function of agent k, taking the form $J_k(w) = \mathbb{E}Q(w; \boldsymbol{x}_k)$ and $p_k > 0$ denote scalar weights, commonly chosen as $p_k = \frac{1}{K}$. In this sense, problem (1) defines a best average model over a collection of potentially heterogeneous agents with local data $\boldsymbol{x}_k$. Distributed architectures can broadly be decomposed into fusion-center based approaches, which involve communication of each individual agent with a single coordinating unit [4], and decentralized approaches relying instead on peer-to-peer interactions over local neighborhoods as defined by some network topology [2, 3, 5, 6].

Most algorithms for distributed learning involve a combination of local self-learning, where individual agents utilize local data and computational resources, and social learning, where these local individual updates are aggregated either across neighborhoods or at a central fusion center. As a representative example of a distributed learning algorithm, we may consider the ATC-diffusion strategy [2]:

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu \widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) \tag{2}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \boldsymbol{\psi}_{\ell,i} \tag{3}$$

Here, $\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1})$ represents a stochastic approximation of the gradient $\nabla J_k(\boldsymbol{w}_{k,i-1})$ based on data available to agent k at time i. One common choice is $\widehat{\nabla J}_k(\boldsymbol{w}_{k,i-1}) \triangleq Q(\boldsymbol{w}_{k,i-1}; \boldsymbol{x}_{k,i})$, although other constructions such as mini-batch, asynchronous, or noisy approximations are possible as well [3]. The set $\mathcal{N}_k$ denotes the neighborhood of agent k and includes the agent itself. The matrix $[A]_{\ell k} = a_{\ell k}$ denotes a left-stochastic, weighted adjacency matrix, satisfying $\sum_{\ell=1}^{K} a_{\ell k} = 1$, $a_{\ell k} > 0$ when $\ell \in \mathcal{N}_k$ and $a_{\ell k} = 0$ otherwise.

Many variations of (2)–(3) exist in the literature, including implementations employing subgradients [6], primal-dual and bias-corrected constructions [5, 7–9], random or time-varying network topologies to model asynchronous behavior and multiple local updates [10, 11] and multitasking [12]. Fusion-center based architectures, as in the case of federated learning [4], can then be obtained from decentralized algorithms by specializing them to a fully-connected or star-shaped topology.

1.1. Robust Aggregation for Distributed Learning

All of the aforementioned algorithms involve linear aggregation steps of the form (3), where intermediate estimates are combined in a linear fashion across neighborhoods $\mathcal{N}_k$. Linear aggregation ensures collaboration, asymptotic consensus and ultimately improved performance. However, allowing for linear contributions from all agents in a neighborhood also make the resulting schemes susceptible to disproportionate influence by faulty, or potentially malicious, actors. This observation has motivated the need to develop *robust* variations of distributed algorithms, which limit the negative effect that a subset of agents cause by deviating from the prescribed learning protocol (e.g., (2)–(3)). Most robust strategies for distributed learning are derived by replacing the linear, but non-robust aggregation (3) by a non-linear, robust aggregator. For example, the work [13] proposes a robust aggregator based on the geometric median:

$$\boldsymbol{w}_{k,i} = \arg\min_{\boldsymbol{w}} \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \|\boldsymbol{\psi}_{\ell,i} - \boldsymbol{w}\| \tag{4}$$

Alternatives based on other robust estimators of the mean [14–20] are possible as well, as are strategies based on robust penalization and primal-dual arguments [21–23].

1.2. Robust Network Topologies

While different schemes for robust distributed learning employ different aggregation schemes, they generally rely on a common set of

Emails: {chutian.wang21, s.vlaski}@imperial.ac.uk

conditions on the network topology to ensure robustness, both in deriving analytical performance guarantees, and in practice [13, 17, 18, 20–22, 22, 24, 25]. To make this precise, we denote the set of all agents $\mathcal{N}$ and decompose $\mathcal{N} = \mathcal{N}^b \cup \mathcal{N}^m$. The set $\mathcal{N}^b$ denotes the set of *benign* agents, which have an interest in solving a cooperative learning problem of the form (1), and faithfully follow the prescribed learning protocol, whether (2) followed by (3) or a robust alternative such as (4). *Malicious* agents are collected in the set $\mathcal{N}^m$ and are modelled as Byzantine agents [15], meaning that they are allowed to deviate arbitrarily from the prescribed learning rule. We can then introduce the following two conditions:

Assumption 1 (**Benign agents are connected.**). *The subnetwork of benign agents $\mathcal{N}^b$ obtained by removing malicious agents $\mathcal{N}^m$ and their associated edges from $\mathcal{N}$ is connected.* □

Assumption 2 (**Majority benign neighborhoods**). *For every benign agent, the majority of their neighbors are benign, i.e.:*

$$\frac{|\mathcal{N}_k \cap \mathcal{N}^b|}{|\mathcal{N}_k|} > \frac{1}{2}, \quad \forall k \in \mathcal{N}^b. \tag{5}$$

Here, $|\cdot|$ denotes the cardinality of the set. We may also define the short-hand notation $\mathcal{N}_k^b \triangleq \mathcal{N}_k \cap \mathcal{N}^b$ and $\mathcal{N}_k^m \triangleq \mathcal{N}_k \cap \mathcal{N}^m$ to denote the benign and malicious neighbors of agent k respectively. □

Assumption 1 ensures that if we remove malicious agents, the remaining network will remain connected, ensuring that information can be diffused through the system without relying on malicious agents as relays. Of course, in practice, one does not know which agents are malicious, and so removing the influence of malicious agents is a non-trivial task. To this end, Assumption 2 ensures that a majority of neighbors of any benign agent are benign. Robust estimators of the mean such as (4) and alternatives employed in the literature [14–20] have a breakdown point of 50% [26], and are hence able to limit the influence of malicious agents on any neighborhood so long as Assumption 2 holds. Several works in the literature have established that Assumptions 1 and 2, together with standard conditions on the cost functions $J_k(\cdot)$ and gradient approximations $\widehat{\nabla J_k}(\cdot)$ are sufficient to ensure provable cooperative learning of benign agents despite the interference by malicious agents [14–22].

Definition 1 (**Robust Network Topology**). *We say that a network topology for a set of nodes $\mathcal{N} = \mathcal{N}^b \cup \mathcal{N}^m$, described by adjacency matrix A, is robust if it satisfies Assumptions 1 and 2, implying that robust distributed learning over the topology is possible using established techniques [14–20].* □

For a fully connected network of agents, or a federated architecture employing a star topology, it is straightforward to verify that the topology is robust if, and only if, the majority of all agents $\mathcal{N}$ is benign, i.e., $\frac{|\mathcal{N}^b|}{|\mathcal{N}|} > 0.5$. In this sense, the network inherits the breakdown point of the underlying robust aggregator. At the same time, given perfect information on the status of agents (benign or malicious), one can construct a robust topology by connecting only benign agents $\mathcal{N}^b$, tolerating an arbitrary number of malicious agents. In this work, we develop a construction for robust network topologies given *imperfect* information on the status of agents. In particular, we model the status of an agent as a Bernoulli random variable, and assume knowledge of the *risk* of an agents, rather than its actual status. As we demonstrate both analytically and in numerical simulations, this limited information is sufficient to construct, asymptotically and with high probability, robust topologies which tolerate fractions of malicious agents arbitrarily close to one.

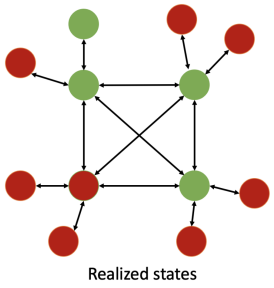


Fig. 1. An example of a fully-connected robust network topology where a majority of agents are malicious (in red).

2. CONSTRUCTION

2.1. Problem formulation

We denote the state of agent k through the random variable $\boldsymbol{q}_k$, where $\boldsymbol{q}_k = 1$ if $k \in \mathcal{N}^b$ and $\boldsymbol{q}_k = 0$ if $k \in \mathcal{N}^m$. The state $\boldsymbol{q}_k$ is a Bernoulli random variable with:

$$\boldsymbol{q}_k = \begin{cases} 1, & \text{w.p. } \overline{q}_k, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Hence the sets of benign and malicious agents $\mathcal{N}^b$ and $\mathcal{N}^m$ respectively are now random. The probability $0 \leq \overline{q}_k \leq 1$ can be viewed as a measure of the risk of agent k. For the purposes of constructing robust topologies, we will assume knowledge of the risk $\overline{q}_k$, but not of the actual state $\boldsymbol{q}_k$. Values of $\overline{q}_k$ close to zero indicate that agent k is very likely malicious, while values close to one indicate that agent k is very likely benign. In practice, values of the risk profile $\overline{q}_k$ will be provided depending on the application at hand, and can be estimated for example via the age or location of a sensor or drone, the type of device, firewall, or past bahavior. We may then introduce the set of likely benign agents as $\widehat{\mathcal{N}^b} \triangleq \{k : q_k > 0.5\}$ and the set of likely malicious agents as $\widehat{\mathcal{N}^m} \triangleq \{k : q_k \leq 0.5\}$, which are deterministic.

Based on this decomposition, we may attempt to construct a network topology by removing likely malicious agents from the network, and connecting likely benign agents $\widehat{\mathcal{N}^b}$. However, we note that such a construction would in general not yield a robust topology asymptotically as $|\mathcal{N}| \to \infty$, since some subset of malicious agents $\mathcal{N}^m$ with $\overline{q}_k > 0$ will almost surely turn out to be benign, yet be disconnected from the set of likely benign agents $\widehat{\mathcal{N}^b}$, hence violating Assumption 1. On the other hand, a fully connected topology will fail almost surely as $|\mathcal{N}| \to \infty$ as soon as $\frac{1}{|\mathcal{N}|}\sum_{k=1}^{|\mathcal{N}|} \overline{q}_k \leq 0.5$, as a majority of agents will almost surely be malicious, hence violating Assumption 2. To overcome this breakdown point of 50%, we need to develop a more nuanced construction. On a high level, we do this by closely connecting likely benign agents to form a resilient cluster, while attaching likely malicious to the periphery, hence limiting their impact (Assumption 2) while ensuring connectivity (Assumption 1).

As illustrated by the discussion so far, and pointed out in [27,28], networks can be simultaneously well-connected and non-robust. To study the conflicting forces behind Assumptions 1 and 2, we now formulate measures of connectivity and robustness in terms of the risk profiles $\boldsymbol{q}_k$ introduced above.

Definition 2 (**Components of benign agents.**). *For a given set of agents $\mathcal{N}$, risk profiles $\overline{q}_k$, and network construction, we denote by*

$\boldsymbol{C}^b_{\mathcal{N}}$ *the number of components of the graph of benign agents* $\mathcal{N}^b$. □

Clearly, since the status of agents, and hence the sets $\mathcal{N}^b$ and $\mathcal{N}^m$ are random, the number of connected components $\boldsymbol{C}^b_{\mathcal{N}}$ will be random as well. Hence, we will study $\mathbb{E}\boldsymbol{C}^b_{\mathcal{N}}$, and our aim will be to show that $\lim_{|\mathcal{N}|\to\infty} \mathbb{E}\boldsymbol{C}^b_{\mathcal{N}} = 1$. Since $\boldsymbol{C}^b_{\mathcal{N}} \geq 1$ with probability one, this would imply that $\lim_{|\mathcal{N}|\to\infty} \boldsymbol{C}^b_{\mathcal{N}} = 1$ almost surely, and hence the construction satisfies Assumption 1 almost surely as $|\mathcal{N}| \to \infty$.

2.2. Proposed Scheme

We will construct the robust network via a random graph, where a link between nodes k and ℓ is established with probability $b_{\ell k}$. Rather than determine $|\mathcal{N}|^2$ separate linkage probabilities, we parametrize $[B]_{\ell k} \triangleq b_{\ell k}$ in terms of scalar penalty parameters d^b and d^m as follows:

Algorithm 1 Proposed design scheme

Require: Risk profiles $\{\overline{q}_k\}_{k=1}^{|\mathcal{N}|}$

1: Compute number of likely benign agents $|\widehat{\mathcal{N}^b}|$ and the smallest risks $\overline{q}^b \triangleq \min_{k\in\widehat{\mathcal{N}^b}} \overline{q}_k$. Analogously, compute $|\widehat{\mathcal{N}^m}|$ and $\overline{q}^m \triangleq \min_{k\in\widehat{\mathcal{N}^m}} \overline{q}_k$ for likely malicious agents.
2: Solve optimization problem (11) to determine the penalty coefficients d^b, d^m.
3: Define a random variable $\boldsymbol{c}$ as:

$$\boldsymbol{c}_k \triangleq \begin{cases} 1, & \text{w.p. } \overline{q}^b \quad \text{if } k \in \widehat{\mathcal{N}^b}, \\ d^b, & \text{w.p. } 1-\overline{q}^b \quad \text{if } k \in \widehat{\mathcal{N}^b}, \\ 1, & \text{w.p. } \overline{q}^m \quad \text{if } k \in \widehat{\mathcal{N}^m}, \\ d^m, & \text{w.p. } 1-\overline{q}^m \quad \text{if } k \in \widehat{\mathcal{N}^m}. \end{cases} \tag{7}$$

4: Compute matrix of linkage probabilities $B = \mathbb{E}\boldsymbol{c}\boldsymbol{c}^{\mathsf{T}}$.
5: Set $b_{kk} = 1$ for all k, ensuring self-loops.
6: Generate the adjacency matrix A according to linkage probabilities B.

The construction (7) ensures that nodes which are likely to be benign are connected with high probability, while likely malicious nodes less likely to be connected. Notably, this construction ensures that pairs of the set of likely benign agents form an Erdős–Rényi graph with uniform edge probability:

$$b_{k\ell} \triangleq b^{bb} = [d^b + (1-d^b)\overline{q}^b]^2 \quad \forall\, k,\ell \in \widehat{\mathcal{N}^b} \tag{8}$$

Similarly, the set of likely malicious nodes form an Erdős–Rényi graph with uniform edge probability:

$$b_{k\ell} \triangleq b^{mm} = [d^m + (1-d^m)\overline{q}^m]^2 \quad \forall\, k,\ell \in \widehat{\mathcal{N}^m} \tag{9}$$

while pairs of likely benign and likely malicious nodes are connected with probability:

$$b_{k\ell} \triangleq b^{bm} \triangleq b^{mb} = [d^b + (1-d^b)\overline{q}^b][d^m + (1-d^m)q^m] \quad \forall\, k \in \widehat{\mathcal{N}^b}, \ell \in \widehat{\mathcal{N}^m} \tag{10}$$

Choosing $d^b = d^m = 1$ results in a fully connected network, while $d^b = d^m = 0$ yields linkage probabilities in relation to the probability that both agents in a pair are benign. The level of penalization of likely malicious nodes is determined by the scalars d^b, d^m. These parameters are critical to the connectivity-robustness trade-off, as sparse connections increase the likelihood of benign agents being disconnected from the network (violating Assumption 1), while dense connections increase the likelihood that malicious agents are able to dominate benign agents (violating Assumption 2). To determine an optimal choice of penalization parameters d^b, d^m, we formulate the following problem which quantifies this trade-off.

$$\max_{d^b,d^m} \quad \sum_{t\in\{b,m\}} \overline{q}^t \frac{\sum_{t'=\{b,m\}} b^{tt'}|\widehat{\mathcal{N}^{t'}}|\overline{q}^{t'}}{\sum_{t'\in\{b,m\}} b^{tt'}|\widehat{\mathcal{N}^{t'}}|} \tag{11a}$$

$$\text{s.t.} \quad d^t \leq 1, \quad t \in \{b,m\} \tag{11b}$$

$$0 \leq b^{tt'} \leq 1, \quad t,t' \in \{b,m\} \tag{11c}$$

$$b^{bb} > \frac{log(|\widehat{\mathcal{N}^b}|)}{|\widehat{\mathcal{N}^b}|} \tag{11d}$$

$$\left(\left\lceil b^{bb}|\widehat{\mathcal{N}^b}|\frac{2\overline{q}^b-1}{1-2\overline{q}^m}\right\rceil - 1\right)(1-\overline{q}^b) < b^{bm}|\widehat{\mathcal{N}^m}| \tag{11e}$$

$$b^{bm}|\widehat{\mathcal{N}^m}| < \left\lceil |\widehat{\mathcal{N}^b}|(2\overline{q}^b b^{bb}-1)\right\rceil - 1 \tag{11f}$$

The objective (11a) is a weighted sum of the expected ratio of benign neighbors of the two types of agents. We are looking to maximize it in order to maximize the probability of success of the network. Constraints (11b) and (11c) ensure the resulting quantities are valid probabilities. Constraint (11d) is a sharp threshold from percolation theory [29] that ensures that the Erdős–Rényi random graph (of likely benign agents) is almost surely connected for sufficiently large network sizes. Constraint (11e) imposes a lower bound on the linkage probability between likely malicious and likely benign agents, ensuring that likely malicious agents are sparsely connected to the set of likely benign agents. The final constraint (11f) ensures robustness of the network by imposing that, even if all likely malicious agents turn out to be malicious, the neighborhood of any likely benign agent will continue to be majority benign.

Theorem 1 (**Asymptotic Robustness**). *Suppose the ratio of likely benign to likely malicious agents* $\frac{|\widehat{\mathcal{N}^b}|}{|\widehat{\mathcal{N}^m}|} = \epsilon$ *is fixed. Then, for any* $\epsilon > 0$*, there exists a network size* $|\mathcal{N}|$ *large enough, such that the topology obtained via Algorithm 1 is robust with probability arbitrarily close to one.*

Proof. The proof follows essentially by construction of the optimization problem (11). Due to space limitations, we only provide a sketch. First, we note that due to the structure of the linkage probability matrix B as defined in (7), the subset of likely benign agents forms an Erdős–Rényi graph, and in light of (11d) this graph is connected almost surely. To ensure the connectivity condition in Assumption 1, it is then sufficient to further show that each agent in $|\widehat{\mathcal{N}^m}|$, connects to at least one node in $|\widehat{\mathcal{N}^b}|$. This is ensured by condition (11e). It can be verified that a sufficient condition for the feasibility of (11e)–(11f) is that:

$$|\widehat{\mathcal{N}^m}| \leq \left(\left\lceil |\widehat{\mathcal{N}^b}|\frac{2\overline{q}^b-1}{1-2\overline{q}^m}\right\rceil - 1\right)(1-\overline{q}^b)|\widehat{\mathcal{N}^b}| \tag{12}$$

Since the right-hand side of (12) grows quadratically with $|\widehat{\mathcal{N}^b}|$, it follows that for any ϵ, there will be an $|\mathcal{N}|$ large enough, such that (12) holds, and hence the optimization problem (11) is feasible, yielding a connected graph. Constraint (11f), on the other hand ensures robustness. □

3. NUMERICAL RESULTS

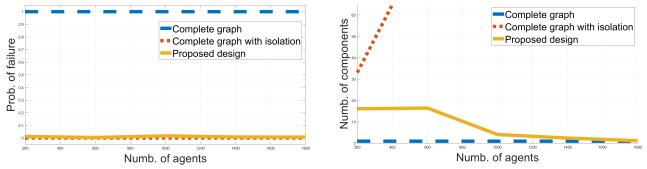
3.1. Asymptotic behavior for growing network sizes

We examine the asymptotic behavior of the number of connected components (quantifying Assumption 1) and probability of being majority benign (quantifying Assumption 2) for the proposed design as well as two baselines, namely complete graphs of all agents $\mathcal{N}$ [27, 30] and complete graphs of likely benign agents $\widehat{\mathcal{N}^b}$ with likely malicious agents isolated[1].

The setting is detailed in Table 3.1. Since the expected ratio of benign agents is $0.28 < 0.5$, fully-connected graphs, while satisfying Assumption 1, are expected to not be majority benign (Assumption 2). A fully-connected graph of only likely benign agents, on the other hand, is expected to be majority benign (Assumption 2), but not include all actual benign agents (failing Assumption 1). The proposed design is expected to exhibit a failure rate that approaches zero (indicating majority benign neighborhoods) and a number of components that approaches one (indicating a connected set of benign agents). This can be verified in Fig. 2. In Fig. 2(a) we observe that beginning at a network size of 200, the generated network topology has failure rate close to zero. In Fig. 2(b) on the other hand, we observe that larger networks are necessary to ensure that the number of components decays down to one. This means that some benign agents are disconnected incorrectly (based on their high likelihood of being malicious) from the cluster of likely benign agents. Nevertheless, as the size of the network grows, even those agents are attached, resulting in a single connected component as the size of the network approaches 2000 while preserving failure rate near zero.

Size of network	200 to 1800
Number of trials at each size	1000
Ratio of likely benign agents	20%
Ratio of likely malicious agents	80%
Benign chance of l.b. agents	0.6
Benign chance of l.m. agents	0.2
Expected ratio of benign agents	0.28

Table 1. Settings for asymptotic performance examination



(a) Statistical failure rate (b) Number of components

Fig. 2. Asymptotic behaviour of network topologies.

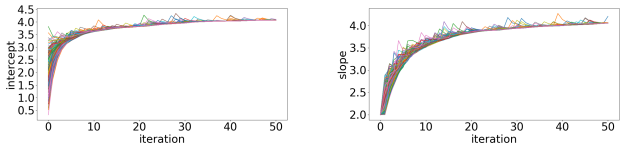
3.2. Distributed linear regression with malicious agents

We consider a collection of agents with the aim of solving a distributed linear regression problem by employing with a weighted median aggregation scheme as in [15]. Malicious agent pass a faulty signal following a normal distribution $\mathcal{N}(10, 5)$. Data is generated according to the linear model $\boldsymbol{\gamma} = 2\boldsymbol{h} + 2$. The setting is detailed in the Table 3.2. As shown in Fig. 3, a fully-connected graph of all agents $\mathcal{N}$ results in consensus on incorrect slope and intercept around 4. This bias is the result of the fact that malicious agents are able to influence benign agents over the non-robust network topology. The proposed design shields a majority of benign agents from the influence of malicious agents. In this particular example, illustrated in Fig. 4, 260 (93.9%) of 277 benign agents are included in the large connected component likely benign agents and come to consensus on accurate estimates of the slope and intercept. The remaining 17 (6.1%) of 277 benign agents are corrupted and/or disconnected. Notably, despite the fact that a majority of agents in the network (323 out of 600) are malicious, the vast majority of benign agents (260 out of 277) are unaffected. As the size of the network grows, the number of corrupted benign agents will approach zero, as shown in the asymptotic studies in Section 3.1.

[1]Early versions of the simulations in Sec. 3 were carried out on a Graphcore Intelligence Processing Unit (IPU) made available through its Academics Program.

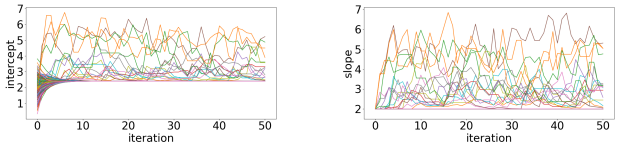
Size of network	600
Total number of samples	10000
Number of samples drawn	50
Ratio of likely benign agents	20%
Ratio of likely malicious agents	80%
Benign chance of l.b. agents	0.8
Benign chance of l.m. agents	0.4
Number of iterations	50
Expected ratio of benign agents	0.48
Number of realized benign agents	277

Table 2. Settings for distributed linear regression under Byzantine attack



(a) Intercept (b) Slope

Fig. 3. Biased regression over a complete graph.



(a) Intercept (b) Slope

Fig. 4. Distributed linear regression over proposed topology.

4. CONCLUSION

Robust schemes for distributed learning rely on certain conditions on the arrangement of malicious agents in a network in order to guarantee successful learning. We have proposed a strategy for designing robust network topologies based on limited information of agents' risk profiles and demonstrated both analytically and empirically that the resulting topologies are able to asymptotically tolerate fraction of malicious agents arbitrarily close one, while ensuring connectivity and robustness.

5. REFERENCES

[1] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Automatic Control*, vol. 31, no. 9, pp. 803–812, 1986.

[2] A. H. Sayed, "Adaptation, learning, and optimization over networks," *Foundations and Trends in Machine Learning*, vol. 7, no. 4-5, pp. 311–801, July 2014.

[3] S. Vlaski, S. Kar, A. H. Sayed, and J. M. F. Moura, "Networked signal and information processing," *available as arXiv:2210.13767*, Oct. 2022.

[4] P. Kairouz, H. B. McMahan, et al, "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, pp. 1–210, 2021.

[5] D. Jakovetić, D. Bajović, J. Xavier, and J. M. F. Moura, "Primal–dual methods for large-scale and distributed convex optimization and data analytics," *Proc. IEEE*, vol. 108, no. 11, pp. 1923–1938, 2020.

[6] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Automatic Control*, vol. 54, no. 1, pp. 48–61, Jan 2009.

[7] W. Shi, Q. Ling, G. Wu, and W. Yin, "EXTRA: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.

[8] P. Di Lorenzo and G. Scutari, "NEXT: In-network nonconvex optimization," *IEEE Trans. Signal and Information Processing over Networks*, vol. 2, no. 2, pp. 120–136, 2016.

[9] K. Yuan, B. Ying, X. Zhao, and A. H. Sayed, "Exact diffusion for distributed optimization and learning – Part II: Convergence analysis," *IEEE Trans. Signal Processing*, vol. 67, no. 3, pp. 724–739, Feb 2019.

[10] X. Zhao and A. H. Sayed, "Asynchronous adaptation and learning over networks—Part I: modeling and stability analysis," *IEEE Trans. Signal Processing*, vol. 63, no. 4, pp. 811–826, 2015.

[11] A. Nedić and A. Olshevsky, "Distributed optimization over time-varying directed graphs," *IEEE Trans. Automatic Control*, vol. 60, no. 3, pp. 601–615, 2015.

[12] R. Nassif, S. Vlaski, C. Richard, J. Chen, and A. H. Sayed, "Multitask learning over graphs: An approach for distributed, streaming machine learning," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 14–25, 2020.

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

[14] D. Yin, Y. Chen, K. Ramchandran, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," *available as arXiv:1803.01498*, March 2018.

[15] P. Blanchard, E. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[16] 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, no. 3, pp. 146–159, 2020.

[17] S. Li, E. Ngai, and T. Voigt, "Byzantine-robust aggregation in federated learning empowered industrial iot," *IEEE Transactions on Industrial Informatics*, pp. 1–1, 2021.

[18] B. Zhu, J. Jiao, and J. Steinhardt, "When does the tukey median work?," in *2020 IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 1201–1206.

[19] C. Fang, Z. Yang, and W. U. Bajwa, "BRIDGE: Byzantine-resilient decentralized gradient descent," *available as arXiv:1908.08098*, Aug 2022.

[20] S. Vlaski, C. Schroth, M. Muma, and A. M. Zoubir, "Robust and efficient aggregation for distributed learning," *arXiv preprint arXiv:2204.00586*, Apr. 2022.

[21] L. Li, W. Xu, T. Chen, G. B. Giannakis, and Q. Ling, "RSA: byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets," in *The Thirty-Third AAAI Conference on Artificial Intelligence*, 2019, pp. 1544–1551.

[22] J. Peng, W. Li, and Q. Ling, "Byzantine-robust decentralized stochastic optimization over static and time-varying networks," *Signal Processing*, vol. 183, pp. 108020, 2021.

[23] Y. SarcheshmehPour, Y. Tian, L. Zhang, and A. Jung, "Networked federated multi-task learning," *available as arXiv:2105.12769*, May 2021.

[24] G. Damaskinos, E. El Mhamdi, R. Guerraoui, A. Guirguis, and S. Rouault, "Aggregathor: Byzantine machine learning via robust gradient aggregation," in *Proceedings of Machine Learning and Systems*, A. Talwalkar, V. Smith, and M. Zaharia, Eds., 2019, vol. 1, pp. 81–106.

[25] E. El Mhamdi, R. Guerraoui, and S. Rouault, "The hidden vulnerability of distributed learning in byzantium," *available as arXiv:1802.07927*, Feb. 2018.

[26] R.A. Maronna, D.R. Martin, and V.J. Yohai, *Robust Statistics: Theory and Methods*, Wiley Series in Probability and Statistics. Wiley, 2006.

[27] H. Zhang and S. Sundaram, "Robustness of information diffusion algorithms to locally bounded adversaries," in *2012 American Control Conference (ACC)*, 2012, pp. 5855–5861.

[28] H. LeBlanc and X. Koutsoukos, "Consensus in networked multi-agent systems with adversaries," in *Proceedings of the 14th international conference on Hybrid systems: computation and control*, Jan. 2011, pp. 281–290.

[29] P. Erdos and A. Rényi, "On the evolution of random graphs," *Publ. Math. Inst. Hung. Acad. Sci*, vol. 5, no. 1, pp. 17–60, 1960.

[30] J. Usevitch and D. Panagou, "r-robustness and (r, s)-robustness of circulant graphs," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, 2017, pp. 4416–4421.