Learning Trust Over Directed Graphs in Multiagent Systems

Orhan Eren Akgün* erenakgun@g.harvard.edu
Arif Kerem Dayı* keremdayi@college.harvard.edu
Stephanie Gil* sgil@seas.harvard.edu
Angelia Nedić† Angelia.Nedich@asu.edu

Abstract

We address the problem of learning the legitimacy of other agents in a multiagent network when an unknown subset is comprised of malicious actors. We specifically derive results for the case of directed graphs and where stochastic side information, or observations of trust, is available. We refer to this as "learning trust" since agents must identify which neighbors in the network are reliable, and we derive a protocol to achieve this. We also provide analytical results showing that under this protocol i) agents can learn the legitimacy of all other agents almost surely, and that ii) the opinions of the agents converge in mean to the true legitimacy of all other agents in the network. Lastly, we provide numerical studies showing that our convergence results hold in practice for various network topologies and variations in the number of malicious agents in the network.

Keywords: Multiagent systems, adversarial learning, directed graphs, networked systems

1 Introduction

Learning the network topology in multiagent systems, what edges exist and are reliable, is critical because of the central role it plays in many multiagent collaboration tasks. This includes a wide range of tasks from estimation, to control, to machine learning, optimization and beyond [16, 18, 21]. Many times both the coordination protocols and achievable performance of the team is dictated by topology [2, 15, 17, 27]. Two aspects that can greatly complicate the learning however, are i) directed graphs, and ii) the presence of untrustworthy data. Directed graphs are more common in practice due to heterogeneity in sensing and communication capabilities in multiagent systems, but are often more difficult to analyze due to non-symmetric information flow. On the other hand, the presence of malicious agents are an important real-world consideration but lead to untrustworthy data in the system [6, 11, 23, 24]. Unfortunately, the compounded impact of both of these challenges is a very complex problem with sparse theory to date. ***Our objective in this paper is to develop a learning protocol and its related analysis, where agents learn over time the legitimacy of their neighbors in the presence of malicious agents over directed graphs.***

The class of problems over directed graphs pose a particular challenge to achieving resilience: many distributed algorithms on directed graphs require agents to have some information about their out-neighbors, but because of the asymmetric information flow, they cannot sense or obtain information directly from these agents. This makes detection of malicious out-neighbors particularly difficult. For instance, the distributed optimization algorithms presented in [13, 15, 20, 25, 26] and the distributed consensus algorithms [2, 4] all require that the agents know the number of out-neighbors they have. This assumption can break if

*School of Engineering and Applied Sciences (SEAS), Harvard University
†ECEE, Arizona State University

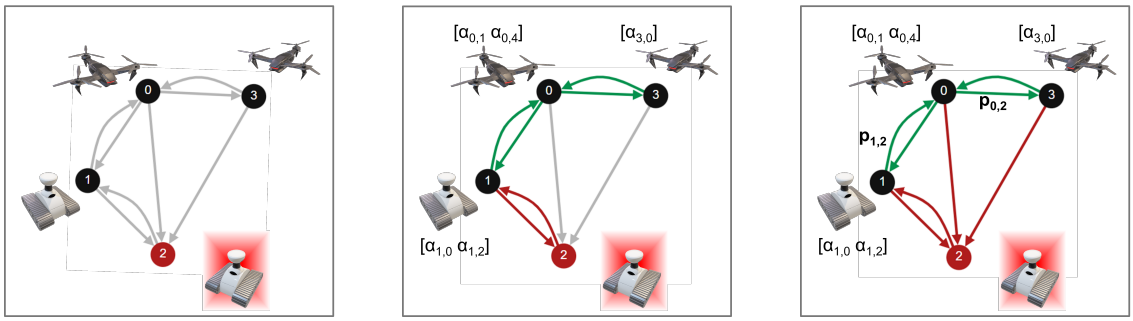
an agent designs the update rule considering an out-neighbor as legitimate, but that agent is malicious in reality. Hence, agents need to have some information about the trustworthiness of their out-neighbors. An interesting concept that has the potential to help this difficult problem is the use of"side information" or data in cyberphysical systems [3, 7–9, 12, 14, 19, 22, 28]. Recent work has shown that by leveraging physical channels of information in the system, agents can gain stochastic information about the trustworthiness of the other agents [8, 9, 12, 28]. We call these "stochastic observations of trust." It has been shown that exploiting these observations leads to stronger results in resilience for multiagent systems [7, 14, 29]. Unfortunately however, existing results do not immediately extend to the case of directed graphs.

In this work, we are interested in learning a trusted graph topology over a directed graph. Using stochastic information about trustworthy neighbors, agents can decide how they should process information that they receive from their in-neighbors, and with which out-neighbors they should share their information. Since agents cannot necessarily observe their out-neighbors, it is natural to think that they need to get information about their out-neighbors from the other agents. We investigate what sufficient information agents can share and how they should process this information to learn the trustworthiness of the other agents in the system in a robust way. This setup is particularly challenging since there might be malicious agents in the system sharing misinformation during this learning process. We present a learning protocol to enable each agent to learn the trustworthiness of all other agents in the system leveraging the opinion of their neighbors. Agents develop opinions in two ways: For their in-neighbors they can obtain a trust observation, they then use this information to form their own opinions. For the other agents, they use the opinions of their in-neighbors they trust to update their opinions. Under the assumption that the subgraph of legitimate agents is strongly connected and each malicious agent is observed by at least one legitimate agent, we show that all legitimate agents can almost surely learn the trustworthiness of all other agents.

Our contributions can be summarized as follows: i) We present a novel learning protocol that enables the legitimate agents in the system to learn the trustworthiness of the other agents where the underlying communication network is a directed graph; ii) We prove that using our learning protocol, legitimate agents can learn the identities of the other agents almost surely; iii) We show that opinions of the agents converge in mean to the true identity of the agents; iv) We provide extensive numerical studies to show that the convergence results hold in practice for various network topologies and the number of malicious agents.

2 Problem Formulation

We consider a distributed multi-agent system where agents need to collaborate in order to achieve a common task such as solving an optimization problem. We represent the communication graph among agents with a directed graph $G = (V, E)$ where the set V represents the set of agents communicating over G with a set E of directed links. Moreover, we let $N = |V|$ be the number of agents. If there is an edge $(i, j) \in E$, then agent i can send information to j, and we say that j is an out-neighbor of i and i is an in-neighbor of j. We assume each agent i has a self-loop $(i, i) \in E$. Moreover, for an agent $i \in V$, we define its in-neighborhood $\mathcal{N}_i^{\text{in}} = \{j \in V \mid (j, i) \in E\}$ and out-neighborhood $\mathcal{N}_i^{\text{out}} = \{j \in V \mid (i, j) \in E\}$. We assume that agents in the system communicate at every time step t. Moreover, we assume that there might be a set $\mathcal{M} \subsetneq V$, called *malicious agents*, of non-cooperative agents in the system that are either adversarial or malfunctioning. We assume that malicious agents can act arbitrarily. We call the set of cooperative agents, that is, the set of agents outside the set $\mathcal{M}$, legitimate agents, denoted by $\mathcal{L}$. We have $\mathcal{L} \cap \mathcal{M} = \emptyset$ and $\mathcal{L} \cup \mathcal{M} = V$. We say that malicious agents are untrustworthy and legitimate agents are trustworthy. We assume that the set of malicious agents $\mathcal{M}$ is unknown. We wish to learn the *trustworthiness* of agents in the network. We are interested in the problems where every agent receives a stochastic observation of trust from an agent that sends information during each communication round. We note that stochastic observations of trust have been developed in previous works [8, 29] and we use a similar definition here:



(a) (b) (c)

Figure 1: This schematic shows our problem setup with one malicious agent shown as a red node. Various stages of learning are depicted: (a) initial state (b) agents use their direct observations to learn the trustworthiness of other agents (c) agents indirectly learn the trustworthiness of the entire network by propagating their opinions.

Definition 1 (Stochastic Observation of Trust α_{ij}) *We denote stochastic observations of trust with $\alpha_{ij}(t)$ if agent j sends information to agent i at time t, and we assume that $\alpha_{ij}(t) \in [0, 1]$. Here, $\alpha_{ij}(t)$ represents the stochastic value of trust of agent j as observed by agent i.*

Agents can develop opinions about trustworthiness of their in-neighbors using these stochastic trust observations over time. However, it is not straightforward how they can develop opinions about their out-neighbors since they have no direct observations of their trustworthiness. Next, we formalize the notion of opinion and then we discuss how to construct opinions of agents.

Definition 2 (Opinion of Trust) *We denote agent i's opinion of trust about agent j at time t with $p_{ij}(t) \in [0, 1]$. We say agent i trusts agent j if $p_{ij}(t) \geq 1/2$ and does not trust agent j otherwise.*

We want to find a learning protocol to enable the legitimate agents to develop accurate opinions $p_{ij}(t)$ about their neighbors in directed graphs, including their out-neighbors. An example case is shown in Figure 1. Next, we state our assumptions under which we develop our protocol.

Assumption 1 (Connectivity of Network) *1. Sufficiently connected graph: The subgraph $G_{\mathcal{L}}$ induced by the legitimate agents is strongly connected.*

2. Observation of malicious agents: For any malicious agent $j \in \mathcal{M}$, there exists some legitimate agent $i \in \mathcal{L}$ that observes j, i.e., $j \in \mathcal{N}_i^{\text{in}}$ for some $i \in \mathcal{L}$.

Assumption 2 (Trust Observations) *Suppose that the following hold:*

1. *Homogeneity of trust variables: The expectation of the variables $\alpha_{ij}(t)$ are constant for the case of malicious transmissions and legitimate transmissions, respectively, i.e., for some scalars c, d with $c < 0$ and $d > 0$, $c = \mathbb{E}[\alpha_{ij}(t)] - 1/2$ for all $i \in \mathcal{L}$, $j \in \mathcal{N}_i^{\text{in}} \cap \mathcal{M}$, and $d = \mathbb{E}[\alpha_{ij}(t)] - 1/2$ for all $i \in \mathcal{L}$, $j \in \mathcal{N}_i^{\text{in}} \cap \mathcal{L}$.*

2. *Independence of trust observations: The observations $\alpha_{ij}(t)$ are independent for all t and all pairs of agents i and j, with $i \in \mathcal{L}$, $j \in \mathcal{N}_i^{\text{in}}$. Moreover, for any $i \in \mathcal{L}$ and $j \in \mathcal{N}_i^{\text{in}}$, the observation sequence $\{\alpha_{ij}(t)\}_{t \in \mathbb{N}}$ is identically distributed.*

Note that stochastic observations of trust satisfying Assumption 2.1 were derived in [8]. Additionally, we make the same Assumptions 1.1, 2.1, and 2.2 as in the work [29], except for the first assumption, where we require the graph to be strongly connected instead of connected since we deal with directed graphs. Assumption 1.2 is new and necessary since it is not possible to learn the legitimacy of an agent if no other agent is observing that agent. This requirement shows up in the analysis later on. We formalize the problem that we are aiming to solve in this paper as follows:

Problem 1 *Let i be a legitimate agent and let q be an arbitrary agent in the system. Assume that stochastic observations of trust are available and Assumption 1 and Assumption 2 hold. We want to find a learning protocol such that for all legitimate agent $i \in \mathcal{L}$ and for all agents $q \in V$, $p_{ij}(t)$ converges to 1 if $q \in \mathcal{L}$ and 0 if $q \in \mathcal{M}$ almost surely.*

3 Learning Protocol

In this section we introduce our learning protocol. Let each agent i store a vector of trust $p_i(t)$ at time t, where $p_i(t)$ is an $N \times 1$ column vector. Let $p_{ij}(t)$ denote the jth component of $p_i(t)$. The value $p_{ij}(t)$ represents agent i's opinion about the node j where a higher $p_{ij}(t)$ indicates that agent i trusts agent j more. Let $\beta_{ij}(t)$ represent an aggregate trust value for the link (j, i) at time t. Following [29], we define $\beta_{ij}(t)$ as

$$\beta_{ij}(t) = \sum_{k=0}^{t} (\alpha_{ij}(k) - 1/2) \tag{1}$$

for all $j \in \mathcal{N}_i^{\text{in}}$ and we define $\beta_{ii}(t) = 1$ for all t. Using the aggregated stochastic trust value $\beta_{ij}(t)$, a legitimate agent i decides on its trusted in-neighbor set by defining $\mathcal{N}_i^{\text{in}}(t) = \{j \in \mathcal{N}_i^{\text{in}} \mid \beta_{ij}(t) \geq 0\}$. In our learning protocol, an agent i shares $p_i(t)$ with its out-neighbors. A legitimate agent i determines its vector of $p_i(t)$ after receiving $p_j(t-1)$ from all of its in-neighbors $j \in \mathcal{N}_i^{\text{in}}$ using the following update rule:

$$p_{iq}(t) = \begin{cases} 1 & \text{if } q \in \mathcal{N}_i^{\text{in}} \text{ and } \beta_{iq}(t) \geq 0 \\ 0 & \text{if } q \in \mathcal{N}_i^{\text{in}} \text{ and } \beta_{iq}(t) < 0. \\ \sum_{j \in \mathcal{N}_i^{\text{in}}(t)} \frac{p_{jq}(t-1)}{|\mathcal{N}_i^{\text{in}}(t)|} & \text{if } q \notin \mathcal{N}_i^{\text{in}} \end{cases} \tag{2}$$

Every legitimate agent i initializes its opinion vector with vector $p_i(0)$ with all ones, meaning that in the beginning, they trust everyone in the network. However, this choice of initialization is arbitrary and as it does not affect our results. A legitimate agent i decides on its trusted out-neighbor set by defining $\mathcal{N}_i^{\text{out}}(t) = \{j \in \mathcal{N}_i^{\text{out}} \mid p_{ij}(t) \geq 1/2\}$.

Notice that the trust vector $p_i(t)$ is in $[0, 1]^N$ by definition. We assume that malicious agents can decide its trust vector $p_i(t)$ arbitrarily. With this protocol, legitimate agents use only the stochastic observations of trust α_{ij} to determine the legitimacy of their in-neighbors. For the other nodes, they use the opinions of their trusted in-neighbors to form their opinion.

4 Analysis

Recall that agents either directly observe an agent and develop their own opinions using their observations, or they use the opinions of others to generate an opinion about an agent. In our analysis, we first show that all legitimate agents learn their in-neighbors such that their trusted in-neighbors are the same as their legitimate in-neighbors. Learning in-neighbors allow agents to propagate this information to others and also stop the inflow of information from any malicious agent. Then, we analyze the propagation of information after

legitimate agents learned their in-neighbors. To do this, we write the update rule of trustworthiness about an agent in matrix form, and show that the effect of the error introduced by malicious agents is asymptotically eliminated. More precisely, we show that estimated trust values converge in mean and almost surely to true trust values (1 for legitimate, 0 for malicious agents).

4.1 Notation

Let $|S|$ denote the cardinality of set S. Let $[W]_{ij}$ denote entry in row i and column j of matrix W. For some agent j and set S, define the indicator function $\mathbf{1}_{\{j\in\mathcal{S}\}}$ as:

$$\mathbf{1}_{\{j\in\mathcal{S}\}} = \begin{cases} 1 & \text{if } j \in \mathcal{S} \\ 0 & \text{otherwise} \end{cases}.$$

We also use the same notation for indicator vectors when the size of the vector is clear from the context.

4.2 Learning Trustworthiness

Since agents use their trusted in neighbors in their updates, we start by showing that agents learn the legitimacy of their in-neighbors. This will be useful later to show that the protocol converges to the desired state.

Lemma 1 *There exists a random finite time T_f such that the following holds almost surely*

$$\begin{aligned} \beta_{ij}(t) \geq 0 \text{ for all } t \geq T_f \text{ and } i \in \mathcal{L}, j \in \mathcal{N}_i^{\text{in}} \cap \mathcal{L} \\ \beta_{ij}(t) < 0 \text{ for all } t \geq T_f \text{ and } i \in \mathcal{L}, j \in \mathcal{N}_i^{\text{in}} \cap \mathcal{M} \end{aligned} \tag{3}$$

Proof: Follows directly from [29, Proposition 1] □

Corollary 1 *There exists a random finite time T_f such that for all $t \geq T_f$ and for all legitimate agents i, trusted in-neighbor set consist of all legitimate neighbors of the agent i, that is:*

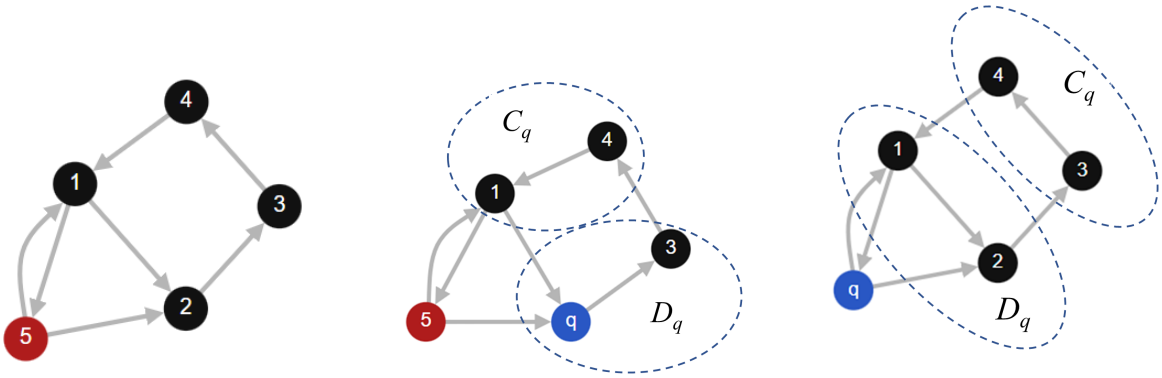
$$\mathcal{N}_i^{\text{in}}(t) = \mathcal{N}_i^{\text{in}} \cap \mathcal{L}.$$

Proof: Follows directly from Lemma 1 and the update rule of the learning protocol given by (2). □ Notice that corollary 1 shows that every legitimate agent can learn its in-neighbors correctly. Now, let $q \in V$ be a an arbitrary but fixed agent in the network. Our goal is to show that all legitimate agents learn the identity of q. This process requires information to propagate from agents receiving trust information directly from q to other agents in the network, which motivates the following definitions:

Define $\mathcal{D}_q \subseteq \mathcal{L}$ to be the subset of legitimate agents directly observing q, i.e. $\mathcal{D}_q \triangleq \mathcal{N}_q^{\text{out}} \cap \mathcal{L}$. Similarly, define $\mathcal{C}_q \subseteq \mathcal{L}$ be the subset of legitimate agents not observing q, i.e. $\mathcal{C}_q \triangleq \mathcal{L} \backslash D_q$. These sets are illustrated in Fig. 2.

These sets are defined for the sake of analysis and are not assumed to be known in practice. Notice that the set $\mathcal{D}_q$ of observing agents is non-empty. This follows directly from Assumption 1.2, since there exists at least one legitimate agent $i \in \mathcal{N}_q^{\text{out}}$. On the other hand, the set $\mathcal{C}_q$ can be empty if all agents are directly observing q. In that case, all legitimate agents will eventually learn the identity of q by Corollary 1.

Now, we analyze the evolution of $p_{iq}(t)$ by writing the evolution of opinions about agent q in matrix form. Let $u_q = |\mathcal{C}_q|$, i.e. the number of agents not observing q. Without loss of generality, reorder the indices of agents such that $\mathcal{C}_q = \{1, 2, \ldots, u_q\}$, and $\mathcal{D}_q = \{u_q + 1, \ldots, |\mathcal{L}|\}$. We denote the vector of trust



(a) Example network (b) $q = 2$ (c) $q = 5$

Figure 2: (a) Network with four legitimate and one malicious nodes. Legitimate nodes are black and the malicious node is red. (b) Learning dynamics for agent $q = 2$: Both agents 2 and 3 directly observe agent 2, so $2, 3 \in \mathcal{D}_q$. 1 and 4 are in the set $\mathcal{C}_q$ since they do not directly observe agent 2. (c) Learning dynamics for agent $q = 5$: Both agents 1 and 2 directly observe agent 5, so $1, 2 \in \mathcal{D}_q$. 3 and 4 are in $\mathcal{C}_q$ since they do not observe 5.

estimates of legitimate agents about the agent q by $p_{\mathcal{C}_q}(t) = \begin{bmatrix} p_{1,q}(t) & \dots & p_{u_q,q}(t) \end{bmatrix}^T$ for the agents in the set $\mathcal{C}_q$ and $p_{\mathcal{D}_q}(t) = \begin{bmatrix} p_{u_q+1,q}(t) & \dots & p_{|\mathcal{L}|,q}(t) \end{bmatrix}^T$ for the agents in the set $\mathcal{D}_q$. Finally, we denote the vector of trust estimates of malicious agents about the agent q by $p_{\mathcal{M},q}(t) = \begin{bmatrix} p_{|\mathcal{L}|+1,q}(t) & \dots & p_{N,q}(t) \end{bmatrix}^T$. Take an arbitrary agent $i \in \mathcal{C}_q$. Using these reordered indices, we can rewrite the learning protocol as:

$$p_{iq}(t) = \sum_{j \in \mathcal{N}_i^{\text{in}}(t)} \frac{p_{jq}(t-1)}{|\mathcal{N}_i^{\text{in}}(t)|} \tag{4}$$

$$= \sum_{j \in \mathcal{C}_q} [W_q(t)]_{ij} p_{jq}(t-1) + \sum_{j \in \mathcal{D}_q} [W_q(t)]_{ij} p_{jq}(t-1) + \sum_{j \in \mathcal{M}} [W_q(t)]_{ij} p_{jq}(t-1), \tag{5}$$

where $[W_q(t)]_{ij} = \frac{1}{|\mathcal{N}_i^{\text{in}}(t)|}$ if $j \in \mathcal{N}_i^{\text{in}}(t)$ and $[W_q(t)]_{ij} = 0$ otherwise. Here, $W_q(t)$ is a row-stochastic matrix with size $u_q \times N$. Then, we can divide $W_q(t)$ into three parts based on the sets $\mathcal{C}_q$, $\mathcal{D}_q$, $\mathcal{M}$ as $W_q(t) = [W_{\mathcal{C}_q}(t) \; W_{\mathcal{D}_q}(t) \; W_{\mathcal{M}_q}(t)]$ where the matrices $W_{\mathcal{C}_q}(t)$, $W_{\mathcal{D}_q}(t)$, $W_{\mathcal{M}_q}(t)$ have sizes $u_q \times u_q$, $u_q \times |\mathcal{D}_q|$, and $u_q \times |\mathcal{M}|$ respectively. With this representation, we can express the update rule (4) in the matrix form as:

$$p_{\mathcal{C}_q}(t) = \begin{bmatrix} W_{\mathcal{C}_q}(t) & W_{\mathcal{D}_q}(t) & W_{\mathcal{M}_q}(t) \end{bmatrix} \begin{bmatrix} p_{\mathcal{C}_q}(t-1) \\ p_{\mathcal{D}_q}(t-1) \\ p_{\mathcal{M}_q}(t-1) \end{bmatrix}, \tag{6}$$

Recall that there exists some random finite time T_f such that all legitimate agents learn their in-neighbors correctly. Until the system reaches time T_f, malicious agents can affect the learning dynamics. Nevertheless, we will show that the legitimate agents can recover from that effect after reaching time T_f. Now, we focus our analysis on the system dynamics after time T_f.

Lemma 2 *For $t \geq T_f$, the following hold almost surely:*

L2.1 The matrix representing the contribution of malicious agents $W_{\mathcal{M}_q}(t) = 0$

L2.2 $W_{\mathcal{C}_q}(t) = \overline{W}_{\mathcal{C}_q}$ *for some constant matrix* $\overline{W}_{\mathcal{C}_q}$

L2.3 $W_{\mathcal{D}_q}(t) = \overline{W}_{\mathcal{D}_q}$ *for some constant matrix* $\overline{W}_{\mathcal{D}_q}$

L2.4 $p_{\mathcal{D}_q}(t-1) = \mathbf{1}_{\{q\in\mathcal{L}\}}$.

Proof: Assuming that $t \geq T_f$, by Corollary 1, we have that $\mathcal{N}_i^{\text{in}}(t) \cap \mathcal{M} = \emptyset$ for $i \in \mathcal{L}$. Therefore, if $i \in \mathcal{C}_q$ and j is a malicious agent, then $j \notin \mathcal{N}_i^{\text{in}}(t)$. By the definition of W_q, we have $[W_q(t)]_{ij} = 0$ for all malicious j as desired. Similarly, $\mathcal{N}_i^{\text{in}}(t) = \mathcal{N}_i^{\text{in}} \cap \mathcal{L}$, so the update matrices $W_{\mathcal{C}_q}(t)$ and $W_{\mathcal{D}_q(t)}$ are constant for $t \geq T_f$. Finally, L2.4 follows directly from Lemma 1. □

Remark 1 *The matrix* $[\overline{W}_{\mathcal{C}_q}\ \overline{W}_{\mathcal{D}_q}]$ *is row stochastic.*

This follows from the fact that $W_q(t)$ is a row-stochastic matrix and that $W_{\mathcal{M}_q}(t)$ is zero. Since agents in $\mathcal{D}_q$ have already learned the trust of agent q after time T_f, we now focus on the agents in $\mathcal{C}_q$. For all $t \geq T_f + 1$, we can describe the evolution of $p_{\mathcal{C}_q}(t)$ as follows:

$$p_{\mathcal{C}_q}(t) = \overline{W}_{\mathcal{C}_q} p_{\mathcal{C}_q}(t-1) + \overline{W}_{\mathcal{D}_q} p_{D_q}(t-1) \tag{7}$$

We want $p_{\mathcal{C}_q}(t) = \mathbf{1}_{\{q\in\mathcal{L}\}}$, i.e., $p_{\mathcal{C}_q}(t)$ should be equal to a vector of ones if $q \in \mathcal{L}$ and a vector of zeros if $q \in \mathcal{M}$. We can define the error in the estimation of legitimate agents in $\mathcal{C}_q(t)$ about the identity of the agent q at time t as:

$$\Delta_{\mathcal{C}_q}(t) = p_{\mathcal{C}_q}(t) - \mathbf{1}_{\{q\in\mathcal{L}\}} \tag{8}$$

We want to show that $\|\Delta_{\mathcal{C}_q}(t)\| \to 0$ as t goes to infinity. Using (7) we can represent $\Delta_{\mathcal{C}_q}(t)$ as

$$\begin{aligned}
\Delta_{\mathcal{C}_q}(t) &= \overline{W}_{\mathcal{C}_q} p_{\mathcal{C}_q}(t-1) + \overline{W}_{\mathcal{D}_q} p_{\mathcal{D}_q}(t-1) - \mathbf{1}_{\{q\in\mathcal{L}\}} \\
&\overset{(a)}{=} \overline{W}_{\mathcal{C}_q} p_{\mathcal{C}_q}(t-1) + \overline{W}_{\mathcal{D}_q} p_{\mathcal{D}_q}(t-1) - (\overline{W}_{\mathcal{C}_q} \mathbf{1}_{\{q\in\mathcal{L}\}} + \overline{W}_{\mathcal{D}_q} \mathbf{1}_{\{q\in\mathcal{L}\}}) \\
&= \overline{W}_{\mathcal{C}_q} (p_{\mathcal{C}_q}(t-1) - \mathbf{1}_{\{q\in\mathcal{L}\}}) + \overline{W}_{\mathcal{D}_q} (p_{\mathcal{D}_q}(t-1) - \mathbf{1}_{\{q\in\mathcal{L}\}}) \\
&\overset{(b)}{=} \overline{W}_{\mathcal{C}_q} (p_{\mathcal{C}_q}(t-1) - \mathbf{1}_{\{q\in\mathcal{L}\}}) \\
&= \overline{W}_{\mathcal{C}_q} \Delta_{\mathcal{C}_q}(t-1),
\end{aligned} \tag{9}$$

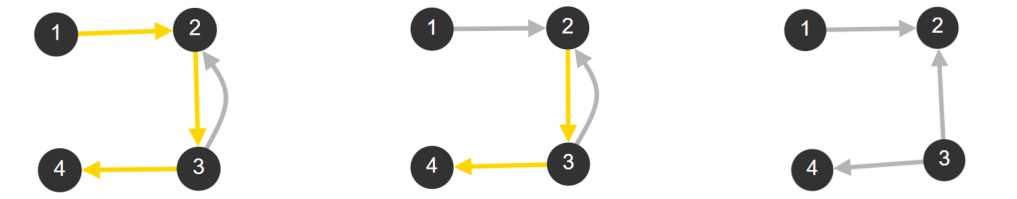
where (a) follows from the fact that $[\overline{W}_{\mathcal{C}_q}\ \overline{W}_{\mathcal{D}_q}]$ is row stochastic and (b) follows from $p_{\mathcal{D}_q}(t-1) = \mathbf{1}_{\{q\in\mathcal{L}\}}$. By using (9) recursively, we obtain

$$\Delta_{\mathcal{C}_q}(t) = \overline{W}_{\mathcal{C}_q}^{t-T_f} \Delta_{\mathcal{C}_q}(T_f) \tag{10}$$

Now, we can bound the error norm:

$$\|\Delta_{\mathcal{C}_q}(t)\| \leq \|\overline{W}_{\mathcal{C}_q}^{t-T_f}\| \|\Delta_{\mathcal{C}_q}(T_f)\|. \tag{11}$$

Here, $\|\Delta_{\mathcal{C}_q}(T_f)\|$ includes the error introduced by malicious agents before all agents learn their in-neighbors. Since the convergence of the error term $\|\Delta_{\mathcal{C}_q}(t)\|$ depends on the convergence of $\overline{W}_{\mathcal{C}_q}$, we analyze the matrix $\overline{W}_{\mathcal{C}_q}$ next.



$$\begin{bmatrix} 1/3 & 1/3 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/2 & 1/2 \end{bmatrix} \qquad \begin{bmatrix} 1/3 & 1/3 & 0 & 0 \\ 0 & 1/4 & 1/4 & 0 \\ 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/2 & 1/2 \end{bmatrix} \qquad \begin{bmatrix} 1/3 & 1/3 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/2 & 1/2 \end{bmatrix}$$

(a) $\widehat{con}W = 3$ (b) $\widehat{con}W = 2$ (c) $\widehat{con}W = \infty$

Figure 3: Three matrices with different contraction indices and the corresponding graphs. The path achieving the contraction index is given in yellow. (a) The only row that sums to less than one is row 1. Therefore, the path with the maximum length from agent 1 to another agent has a length of 3. (b) Row 2 also sums to less than one. Therefore, the longest path is the one from agent 2 to 4. (c) The only row that sums to less than one is row 1. Since there is no path from agent 1 to agent 3 or 4, the index of contraction is ∞.

4.3 Convergence of Weakly Chained Substochastic Matrices

Now, we aim to show that $\overline{W}_{\mathcal{C}_q}$ is *convergent*, i.e. $\|\overline{W}^t_{\mathcal{C}_q}\| \to 0$ as $t \to \infty$. In this part, we will show that $\overline{W}_{\mathcal{C}_q}$ belongs to a family of *convergent* substochastic matrices called *weakly chained substochastic* matrices. This will conclude that the error term goes to 0. First, we give some definitions.

Definition 3 *Digraph of matrix: Let the square matrix $W \in \mathbb{R}^{n\times n}$ be non-negative, i.e. $W_{ij} \geq 0$ for all i, j. Hence, the graph of W, denoted by $G(W) = (V(W), E(W))$ is the graph such that $V(W) = \{1, \ldots, n\}$ and for all $i, j \in \{1, \ldots, n\}$, $(i, j) \in E(W)$ if and only if $W_{ij} > 0$.*

To analyze the convergence properties of $\overline{W}_{\mathcal{C}_q}$, we define the index of contraction following [1]

Definition 4 *Index of contraction: Let the matrix $W \in \mathbb{R}^{n\times n}$ be substochastic. Define the set $\hat{J}(W) \triangleq \{1 \leq i \leq n : \sum_{j=1}^{n} W_{ij} < 1\}$, and let the set $\hat{K}_i(W)$ be the set of all paths[1] in the digraph of W from i to all $j \in \hat{J}(W)$. The index of contraction $\widehat{con}W$ associated with matrix W is defined as:*

$$\widehat{con}W \triangleq \max\left\{0,\ \sup_{i\notin\hat{J}(W)}\left\{\inf_{\omega\in\hat{K}_i(W)}\{|\omega|\}\right\}\right\}, \tag{12}$$

where $|\omega|$ denotes the length of the path ω. Also, we follow the conventions that $\inf\emptyset = \infty$ and $\sup\emptyset = -\infty$. Here, if all rows of W sum to less than one, we have $|\hat{J}(W)| = n$. This implies that the supremum over $i \notin \hat{J}(W)$ is $-\infty$, therefore, $\widehat{con}W = 0$. Similarly, $\widehat{con}W$ is infinite if $\hat{K}_i(W)$ is empty, meaning there is no path from some row $i \notin \hat{J}(W)$ to a that sums to less than one.

[1, Corollary 2.6] shows that a square substochastic matrix W is convergent if and only if $\widehat{con}W$ is finite. We show example matrices with different contraction indices in Figure 3. We call a substochastic matrix with finite contraction index *weakly chained substochastic matrix*.

Remark 2 *Matrix W is a weakly chained substochastic matrix if and only if for all rows i that are not in the set $\hat{J}(W)$, set $\hat{K}_i(W)$ is non-empty, i.e there is a path $i \to i_1 \to \cdots \to i_j$ in $G(W)$ such that row i_j sums to less than one. Moreover, a weakly chained substochastic matrix is convergent.*

[1] We use path instead of walk in contrast to [1] in our definition, however these definitions are equivalent.

This remark follows directly from the definition of the index of contraction and [1, Corollary 2.6].

The following sequence of results will show that $\overline{W}_{\mathcal{C}_q}$ is weakly chained substochastic. We will first establish a relation between the graph that describes the network and the digraph of $\overline{W}_{\mathcal{C}_q}$. In particular, we will establish that the links $E(\overline{W}_{\mathcal{C}_q})$ in the digraph of $\overline{W}_{\mathcal{C}_q}$ are the inversion of links in the original graph. Then use assumptions of strong connectivity and existence of a directly observing agent to conclude that $\overline{W}_{\mathcal{C}_q}$ is weakly chained substochastic.

Lemma 3 *Let $\overline{W}_{\mathcal{C}_q} \in \mathbb{R}^{u_q \times u_q}$ be defined as before, and let $G_{\mathcal{C}_q}$ be the subgraph of $G_{\mathcal{L}}$ induced by the set of agents $\mathcal{C}_q$. Then, $(i,j) \in G(\overline{W}_{\mathcal{C}_q})$ if and only if $(j,i) \in G_{\mathcal{C}_q}$. In other words, $G_{\mathcal{C}_q}$ is the digraph of $\overline{W}_{\mathcal{C}_q}^T$.*

Proof: Let $(i,j) \in G(\overline{W}_{\mathcal{C}_q})$. Then, by definition of digraph of a matrix, we have that $[\overline{W}_{\mathcal{C}_q}]_{ij} > 0$. So, it means that agent $i \in \mathcal{C}_q$ is receiving information from agent $j \in \mathcal{C}_q$, by the learning protocol (4). Thus, there must be an edge (j,i) in $G_{\mathcal{C}_q}$. Similarly, if (j,i) is an edge in $G_{\mathcal{C}_q}$, then agent j is an in-neighbor of agent i. So, agent i is receiving informattion from agent j, which means that $[\overline{W}_{\mathcal{C}_q}]_{ij} > 0$. □

Corollary 2 *If there is a path $v_1 \to v_2 \to \cdots \to v_l$ in $G_{\mathcal{C}_q}$, then there is a path $v_l \to v_{l-1} \to \cdots \to v_1$ in $G(\overline{W}_{\mathcal{C}_q})$.*

Proof: If there is an edge $v_i \to v_{i+1}$ in $G_{\mathcal{C}_q}$, then by Lemma 3, there exists and edge $v_{i+1} \to v_i$ in $G(\overline{W}_{\mathcal{C}_q})$. Since this holds for each $i = 1, \ldots, l-1$, $v_l \to v_{l-1} \to \ldots v_1$ is a path in $G(\overline{W}_{\mathcal{C}_q})$. □

Theorem 1 *For all agents q, given that the set $\mathcal{C}_q$ is non-empty, the update matrix $\overline{W}_{\mathcal{C}_q}$ is a weakly chained substochastic matrix. Moreover, $\overline{W}_{\mathcal{C}_q}$ is convergent.*

Proof: Let $i \in \mathcal{C}_q$. If agent i has a neighbor $d \in \mathcal{D}_q$ directly observing agent q, row i must sum up to less than one since agent i receives information from d and $d \notin \mathcal{C}_q$. So, $i \in \hat{J}(\overline{W}_{\mathcal{C}_q})$.

Now, assume agent i doesn't have a directly observing neighbor, i.e. $i \notin \hat{J}(\overline{W}_{\mathcal{C}_q})$. We know that there exists some agent $d \in \mathcal{D}_q$ that directly observes agent q by Assumption 1.1 and Assumption 1.2. By Assumption 1.1, the subgraph induced by legitimate agents are strongly connected, so there exists a path

$$d = i_0 \to i_1 \to i_2 \to \ldots i_l \to i$$

in $G_{\mathcal{L}}$ for each $i_j \in \mathcal{L}$ where each arrow denotes a directed edge.[2] Now, choose the largest j such that $i_j \in \mathcal{D}_q$, and consider the path

$$i_j \to i_{j+1} \to \cdots \to i_l \to i$$

Here, since j is chosen as the largest j s.t. $i_j \in \mathcal{D}_q$, we have that $i_{j+1}, \ldots, i_l, i \in \mathcal{C}_q$. Moreover, we assumed $i \notin \hat{J}(\overline{W}_{\mathcal{C}_q})$, so $j < l$ since i does not have a directly observing neighbor.

Now, we know, i_{j+1} has a neighbor directly observing q, i.e. i_j. Therefore, row i_{j+1} of $\overline{W}_{\mathcal{C}_q}$ sums to less than 1, meaning that $i_{j+1} \in \hat{J}(\overline{W}_{\mathcal{C}_q})$. From Corollary 2, there exists a path

$$i \to i_l \to i_{l-1} \to \cdots \to i_{j+2} \to i_{j+1}$$

in the graph $G(\overline{W}_{\mathcal{C}_q})$. So, in the digraph $G(\overline{W}_{\mathcal{C}_q})$ there exists a path from i to a row summing to less than one, i_{j+1}, as desired. Hence, $\hat{K}_i(\overline{W}_{\mathcal{C}_q})$ is non-empty for $i \notin \hat{J}(\overline{W}_{\mathcal{C}_q})$. Therefore, $\overline{W}_{\mathcal{C}_q}$ is weakly chained substochastic and convergent by Remark 2. □

Corollary 3 *For all agents $q \in V$ where the set $\mathcal{C}_q$ is non-empty, $p_{\mathcal{C}_q}(t)$ almost surely converges to $\mathbf{1}_{\{q \in \mathcal{L}\}}$ where $\mathbf{1}_{\{q \in \mathcal{L}\}}$ is a vector with all values equal to 1 if $q \in \mathcal{L}$ and to 0 if $q \in \mathcal{M}$.*

[2] $l \geq 1$ since agent i does not have a directly observing neighbor

Proof: Remember that the error is defined as $\Delta_{\mathcal{C}_q}(t) = p_{\mathcal{C}_q}(t) - \mathbf{1}_{\{q\in\mathcal{L}\}}$. By Corollary 1 we know that there exists a finite time T_f such that for all $t \geq T_f + 1$ we have

$$\|\Delta_{\mathcal{C}_q}(t)\| \leq \|\overline{W}_{\mathcal{C}_q}^{t-T_f}\|\|\Delta_{\mathcal{C}_q}(T_f)\|. \tag{(11)}$$

Since both $p_{\mathcal{C}_q}(t)$ and $\mathbf{1}_{\{q\in\mathcal{L}\}}$ are in $[0,1]^{u_q}$, we have $\|\Delta_{\mathcal{C}_q}(T_f)\| \leq \sqrt{u_q}$. By Theorem 1, we have that $\|\overline{W}_{\mathcal{C}_q}^{t-T_f}\| \to 0$. Therefore, $\|\Delta_{\mathcal{C}_q}(t)\| \to 0$ almost surely. □

4.4 Main Results

In this part, we present our main results which show that all legitimate agents can learn the trustworthiness of of all agents in the system. Let $p_{\mathcal{L}_q}(t)$ denote the trustworthiness estimation of all legitimate agents about an agent q at time t. We show that this vector converges to $\mathbf{1}_{\{q\in\mathcal{L}\}}$.

Theorem 2 (Convergence to the true trust vector almost surely) *For all agents $q \in V$, $p_{\mathcal{L}_q}(t)$ converges almost surely to the true trust vector $\mathbf{1}_{\{q\in\mathcal{L}\}}$, where $\mathbf{1}_{\{q\in\mathcal{L}\}}$ is an $|L| \times 1$ vector with all of its values equal to 1 if $q \in \mathcal{L}$ and equal to 0 if $q \in \mathcal{M}$.*

Proof: Without loss of generality, we reorder the indices of agents such that $\mathcal{C}_q = \{1, 2, \ldots, u_q\}$, and $\mathcal{D}_q = \{u_q + 1, \ldots, |\mathcal{L}|\}$ where $\mathcal{C}_q$ is the set of legitimate agents not observing q and $\mathcal{D}_q$ is the set of legitimate agents directly observing q. We have two different cases where the set $\mathcal{C}_q$ is empty and non-empty. First assume that $\mathcal{C}_q$ is empty. We know that $p_{\mathcal{L}_q}(t) = p_{\mathcal{D}_q}$. There exists a finite time T_f such that for all $t \geq T_f + 1$ we have $p_{\mathcal{L}_q}(t) = \mathbf{1}_{\{q\in\mathcal{L}\}}$ by Lemma 2. Hence, $p_{\mathcal{L}_q}(t)$ converges to $\mathbf{1}_{\{q\in\mathcal{L}\}}$ almost surely.

Now, assume that $\mathcal{C}_q$ is non-empty. Hence, we can represent $p_{\mathcal{L}_q}(t)$ as $p_{\mathcal{L}_q}(t) = \begin{bmatrix} p_{\mathcal{C}_q}(t) \\ p_{\mathcal{D}_q}(t) \end{bmatrix}$. Define $\Delta_{\mathcal{L}_q}(t) = p_{\mathcal{L}_q}(t) - \mathbf{1}_{\{q\in\mathcal{L}\}}$. Using the triangle inequality we obtain

$$\begin{aligned} \|\Delta_{\mathcal{L}_q}(t)\| &\leq \|p_{\mathcal{C}_q}(t) - \mathbf{1}_{\{q\in\mathcal{L}\}}\| + \|p_{\mathcal{D}_q}(t) - \mathbf{1}_{\{q\in\mathcal{L}\}}\| \\ &= \|\Delta_{\mathcal{C}_q}(t)\| + \|p_{\mathcal{D}_q}(t) - \mathbf{1}_{\{q\in\mathcal{L}\}}\|, \end{aligned}$$

where $\Delta_{\mathcal{C}_q}(t)$ is the same one with (8). Now, assume that $t \geq T_f + 1$. Then we have $\|p_{\mathcal{D}_q} - \mathbf{1}_{\{q\in\mathcal{L}\}}\| = 0$ by Lemma 2. Moreover, by Corollary 3, we have that $\|\Delta_{\mathcal{C}_q}(t)\| \to 0$ almost surely. Hence, we can conclude that $\|\Delta_{\mathcal{L}_q}(t)\| \to 0$ and $p_{\mathcal{L}_q}(t)$ converges to $\mathbf{1}_{\{q\in\mathcal{L}\}}$ almost surely. □

Theorem 3 (Convergence in mean to the true trust vector) *For all agents $q \in V$ and $r \geq 1$, $p_{\mathcal{L}_q}(t)$ converges in mean to the true trust vector $\mathbf{1}_{\{q\in\mathcal{L}\}}$. That is,*

$$\lim_{t\to\infty} E[\|p_{\mathcal{L}_q}(t) - \mathbf{1}_{\{q\in\mathcal{L}\}}\|^r] = 0. \tag{13}$$

Proof: Since $p_{\mathcal{L}_q}(t) \in [0,1]^{|\mathcal{L}|}$, $\|p_{\mathcal{L}_q}(t)\|_2 \leq \sqrt{|\mathcal{L}|}$. Also we have $\mathbf{1}_{\{q\in\mathcal{L}\}}$ in $[0,1]^{\mathcal{L}}$. Then, using the triangle inequality we get $\|p_{\mathcal{L}_q}(t) - \mathbf{1}_{\{q\in\mathcal{L}\}}\|^r \leq (\sqrt{|\mathcal{L}|})^r < \infty$. We can apply the dominated convergence theorem [30] to conclude our proof since $p_{\mathcal{L}_q}(t)$ converges to $\mathbf{1}_{\{q\in\mathcal{L}\}}$ almost surely by Theorem 2. □
Finally, the following Corollary shows that following this protocol, every legitimate agent can learn the trustworthiness of all agents in the network, including their in- and out-neighbors, $\mathcal{N}_i^{\text{in}}$, $\mathcal{N}_i^{\text{out}}$, for all $i \in \mathcal{L}$.

Corollary 4 (Learning the Trustworthiness of All Agents) *All legitimate agents $i \in \mathcal{L}$ can learn the trustworthiness of all agents in the network correctly. That is, there exists a finite time T_{max} such that for all $t \geq T_{max}$ and for all $q \in V$, $p_{iq}(t) \geq 1/2$ if $q \in \mathcal{L}$ and $p_{iq}(t) < 1/2$ if $q \in \mathcal{M}$ almost surely.*

Proof: Let i be a legitimate agent. Let q be an arbitrary agent in the system. Then, by Theorem 2, there exist a time T_q almost surely such that for all $t \geq T_q$, $p_{iq}(t) > 1/2$ if $q \in \mathcal{L}$ and $p_{iq}(t) < 1/2$ otherwise. Then we can choose $T_{max} = \max_{q \in V} T_q$. □

5 Numerical Studies

In this section, we evaluate the performance of the algorithm via numerical studies. We show that all legitimate agents can learn the the trustworthiness of all the other agents in the system using our algorithm in various network realizations, which supports our theoretical results.

Communication graph: We generate the graph of legitimate agents, denoted with $G_{\mathcal{L}}$, in two different ways to show that the protocol works with different graph structures. The first way is to use a cyclic graph. We choose this graph model because it is strongly connected by default and the contraction index for learning the trustworthiness of any legitimate agent in the system is $|\mathcal{L}| - 2$, which grows linearly with the number of legitimate agents. The second way is to generate a random graph using Erdős–Rényi model where each edge in the graph is either included or not with probability p [5]. We choose $p = \frac{2 \log |\mathcal{L}|}{|\mathcal{L}|}$ to have a high probability of generating a strongly connected graph [10], and in the case where the generated graph is not strongly connected, we repeat the process to ensure satisfaction of Assumption 1.1. The graphs generated using this model is likely to have a better connectivity and a lower contraction index for learning any legitimate agent compared to the cyclic graphs. In this way the cyclic graph represents the most difficult case where legitimate information takes longest to circulate throughout the network. After generating the graph of legitimate agents, we randomly add malicious agents to the system.

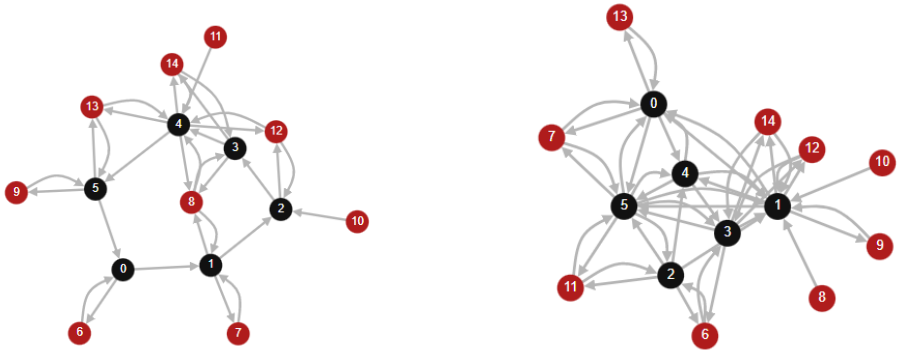
Malicious agents: We assume that all the malicious agents in the system are omnipotent in that they know the trustworthiness of every other agent in the system, this represents a strong attack. The malicious agents do not follow the update rules and they always send the opposite of the true trustworthiness information to other agents, i.e. they assign 1 to all malicious agents and assign 0 to all legitimate agents in the trust vector they share. Since the malicious agents do not follow the learning protocol, we do not explicitly model the communication between the malicious agents.

Trust observations: Following the previous work [29], we model the trust observations $\alpha_{ij}(t)$ as follows: At each time step t we sample $\alpha_{ij}(t)$ uniformly from the interval $[0.35, 0.75]$ if $j \in \mathcal{L}$ and from $[0.25, 0.65]$ if $j \in \mathcal{M}$. This way, $\mathbb{E}[\alpha_{ij}(t)] = 0.55$ if j is a legitimate agent and $\mathbb{E}[\alpha_{ij}(t)] = 0.45$ otherwise. With this setup, Assumption 2 is satisfied.

Metrics: We evaluate the model performance based on two different metrics. The first one is mean squared error (MSE) where we calculate the mean squared error between the true trust vector and trust vector of legitimate agents and take the average across all legitimate agents. Following from Theorem 2, the MSE should converge to 0. The second metric we define is $\hat{T}_{max}$, which is a proxy for T_{max} defined in Corollary 4. If all legitimate agents classify every other agent correctly for N number of time steps after time t, where N is the total number of agents in the system, we assign $\hat{T}_{max} = t$ and stop the experiment, assuming that no further classification error would occur since N is large enough for information to propagate through the whole network.

5.1 Results

Here, we present the results for three different setups with $|\mathcal{L}| \in \{20, 40, 80\}$ and $|\mathcal{M}| = 1.5 \times |\mathcal{L}|$. For each $|\mathcal{L}|$, we generate the network of legitimate agents in two different ways: using a cyclic graph, and an Erdős–Rényi graph over legitimate agents. Then we add the malicious agents randomly, and we track the MSE for both networks. Examples of these graph topologies can be seen in Figure 4 . The results are presented in Figure 5. It can be seen that both MSE and maximum error converges to 0 in all setups.



(a) Cyclic $G_{\mathcal{L}}$ (b) Erdős–Rényi $G_{\mathcal{L}}$

Figure 4: Example graph topologies with $|\mathcal{L}| = 6$, $|\mathcal{M}| = 9$ nodes.

	$\|\mathcal{L}\| = 20, \|\mathcal{M}\| = 30$		$\|\mathcal{L}\| = 40, \|\mathcal{M}\| = 60$		$\|\mathcal{L}\| = 80, \|\mathcal{M}\| = 120$	
	$\hat{T}_{max}$	$\widehat{con}_{max}$	$\hat{T}_{max}$	$\widehat{con}_{max}$	$\hat{T}_{max}$	$\widehat{con}_{max}$
Cyclic	66	19	109	38	192	78
Erdős–Rényi	49	3	64	3	76	3

Table 1: This table shows $\hat{T}_{max}$ and $\widehat{con}_{max}$ for 8 different setups. We can see that a higher $\widehat{con}_{max}$ usually corresponds to a higher $\hat{T}_{max}$. This correlation is intuitive since $\widehat{con}_{max}$ is an indicator of how long it takes for information to propagate from observing agents to non-observing agents. Since the graph topology dictates $\widehat{con}_{max}$, we observe a higher $\hat{T}_{max}$ in Cyclic graphs compared to Erdős–Rényi graphs.

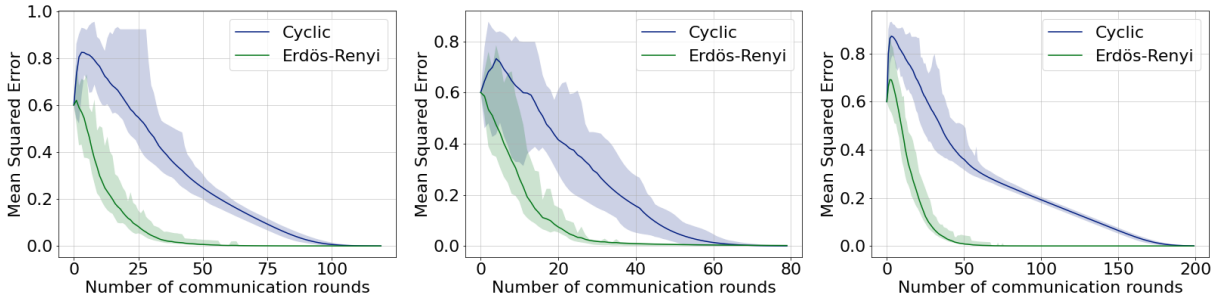
For each setup, we present the maximum contraction index, denoted by $\widehat{con}_{max}$ and $\hat{T}_{max}$ in Table 1. We define the maximum contraction index as $\widehat{con}_{max} = \max_{q \in V} \widehat{con}\overline{W}_{\mathcal{C}_q}$, where $\widehat{con}\overline{W}_{\mathcal{C}_q}$ is defined in (12).

5.2 The Effect of Malicious Agents

In this part, we investigate the effect of malicious agents in the system to the learning protocol. We use the Erdős–Rényi graph setup from the previous part with 40 legitimate agents. We look into two cases: First, we fix the number of malicious agents in the system to 60 and we change the likelihood that the malicious agents make a connection with a legitimate agent. Then, we fix the probability of making a connection to 0.2 and increase the number of malicious agents in the system. The MSE graphs are shown in Fig 6

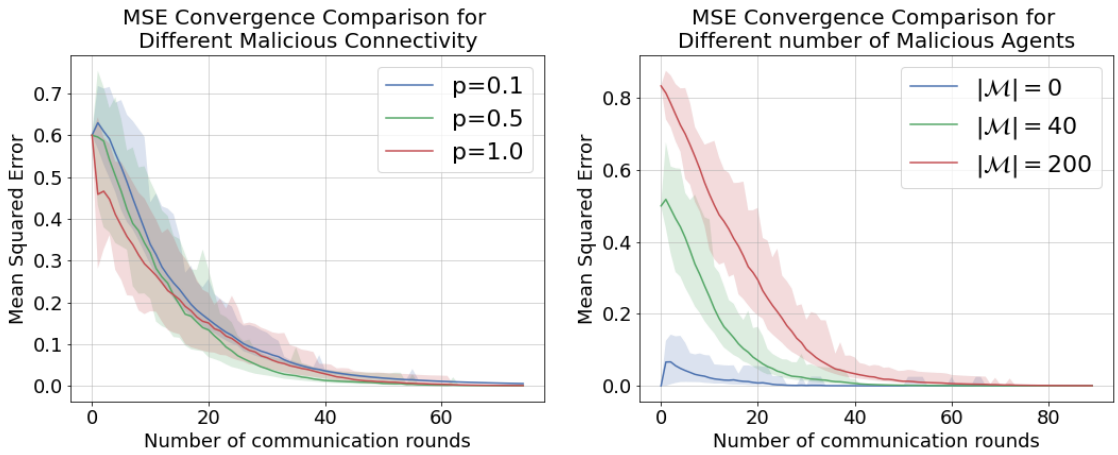
5.3 Necessity of Assumption 1.2

In our experiments, we empirically demonstrate the necessity of Assumption 1.2 for the learning protocol to work. We generated a simple example with two legitimate agents and two malicious agents, which can be seen in Figure 7. In this example, the malicious agent 1 is not an in-neighbor of either of the legitimate nodes while the malicious agent 2 is an in-neighbor of both of the legitimate agents. With this setup, both of the legitimate agents failed to learn the identity of the agent m_1 as expected while learning all of the other agents successfully in ten different trials. This is because before the legitimate agents learn the trustworthiness of the malicious agent 2, the malicious agent 2 changes the opinions of the legitimate agents about the malicious agent 1. After that, since none of the legitimate agents is directly observing the malicious agent 1, their opinion does not change.



(a) $|\mathcal{L}| = 20, |\mathcal{M}| = 30$ (b) $|\mathcal{L}| = 40, |\mathcal{M}| = 60$ (c) $|\mathcal{L}| = 80, |\mathcal{M}| = 120$

Figure 5: Convergence plots for three different cases where the number of malicious agents are chosen as $|\mathcal{M}| = 1.5 \times |\mathcal{L}|$. Solid lines represents the MSE. The shaded areas show the range of error among the legitimate agents. We see that in all cases, the MSE converges to zero eventually as predicted by our theory. Since the malicious agents can have influence on the other nodes in the beginning, we observe an increase in the error at first. This effect is higher in cyclic graphs since the information takes longer to propagate. Moreover, as we increase the size of the graph for cyclic graphs, the convergence time also increases. On the other hand, since Erdős–Rényi graph has good connectivity in all cases, the convergence time is not as sensitive to the graph size compared to the cyclic graphs.



(a) (b)

Figure 6: (a) The effect of increasing the expected number of connections that each malicious agent has. Here, p denotes the probability that the edge (i, m) is present in the system, meaning that $m \in \mathcal{M}$ is an in-neighbor of $i \in \mathcal{L}$. As malicious agents are being observed by more agents, their early effect in the network decreases since all of the directly observing agents learn their trustworthiness using their own observations, without waiting for the information to propagate from the other learning agents. (b) The effect of increasing the number of malicious agents in the system. As we increase the number of malicious agents in the system, their effect on the legitimate agents' opinions also increase. However, the opinions of the legitimate agents still converge to the correct values eventually demonstrating agreement with our main result in Theorems 2 and 3.

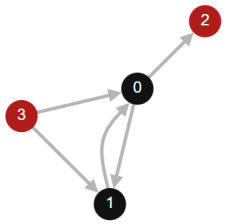


Figure 7: An example network topology where the learning is not guaranteed and the Assumption 1.2 is violated. Since the agent 2 is not an in-neighbor of either of the legitimate agents, they fail to learn the trustworthiness of this agent.

	min	**max**	**mean**	**std**
Cyclic	97	175	112.8	8.85
Erdős–Rényi	36	164	58.2	14.8

Table 2: This table shows the summary statistics of $\hat{T}_{max}$ calculated over 500 random trials. In all trials, we observe a finite time $\hat{T}_{max}$ where no classification errors occur thereafter as predicted by the theory. Since the connections between legitimate agents in Erdős–Rényi graph is random, a higher variation in $\hat{T}_{max}$ is observed.

5.4 Aggregate Results

We present numerical results over multiple trials in this section, each representing a random instantiation of the graph topology and stochastic observations of trust drawn from the distribution described at the beginning of this section. We fix the number of legitimate agents to 40 and the number of malicious agents to 60. Then, we run 500 trials for both cyclic graph and Erdős–Rényi setups. For each trial, we run the protocol for 1000 communication rounds and record the $\hat{T}_{max}$. The resulting statistics of $\hat{T}_{max}$ are shown in Table 2.

6 Conclusion

This paper presents a protocol for learning which agents to trust, and the accompanying analysis, for directed multiagent graphs with stochastic observations of trust. Here, the directed nature of the graph presents an important challenge where the out-neighbors of a node cannot directly observe or receive information from it; this leads to a learning dynamic that makes accurate assessment of malicious agents in the network particularly elusive. The learning protocol developed herein specifically addresses this challenge of learning trust in directed graphs and constitutes the novelty of this paper. Since directed graphs often arise in practical multiagent systems due to heterogeneity in sensing and communication, we believe that the learning protocol and theory presented here can support many optimization, estimation, and learning tasks for general multiagent systems.

Acknowledgments

The authors gratefully acknowledge partial support through NSF awards CNS 2147641, CNS-2147694, and AFOSR grant number FA9550-22-1-0223. The authors would like to thank Ali Taherinassaj for their helpful discussions.

References

[1] Parsiad Azimzadeh. A fast and stable test to check if a weakly diagonally dominant matrix is a nonsingular m-matrix. *Mathematics of Computation*, 88(316):783–800, 2019.

[2] Kai Cai and Hideaki Ishii. Average consensus on general strongly connected digraphs. *Automatica*, 48(11):2750–2761, 2012.

[3] Matthew Cavorsi, Orhan Eren Akgün, Michal Yemini, Andrea Goldsmith, and Stephanie Gil. Exploiting trust for resilient hypothesis testing with malicious robots. *arXiv preprint arXiv:2209.12285*, 2022.

[4] Alejandro D Dominguez-Garcia and Christoforos N Hadjicostis. Distributed matrix scaling and application to average consensus in directed graphs. *IEEE Transactions on Automatic Control*, 58(3):667–681, 2012.

[5] Paul Erdős, Alfréd Rényi, et al. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[6] Michael J Fischer, Nancy A Lynch, and Michael S Paterson. Impossibility of distributed consensus with one faulty process. *Journal of the ACM (JACM)*, 32(2):374–382, 1985.

[7] Stephanie Gil, Cenk Baykal, and Daniela Rus. Resilient multi-agent consensus using wi-fi signals. *IEEE Control Systems Letters*, 3(1):126–131, 2019.

[8] Stephanie Gil, Swarun Kumar, Mark Mazumder, Dina Katabi, and Daniela Rus. Guaranteeing spoof-resilient multi-robot networks. *Autonomous Robots*, 41(6):1383–1400, 2017.

[9] Jairo Giraldo, David Urbina, Alvaro Cardenas, Junia Valente, Mustafa Faisal, Justin Ruths, Nils Ole Tippenhauer, Henrik Sandberg, and Richard Candell. A survey of physics-based attack detection in cyber-physical systems. *ACM Computing Surveys (CSUR)*, 51(4):1–36, 2018.

[10] Alasdair J Graham and David A Pike. A note on thresholds and connectivity in random directed graphs. *Atl. Electron. J. Math*, 3(1):1–5, 2008.

[11] Leslie Lamport, Robert Shostak, and Marshall Pease. The byzantine generals problem. In *Concurrency: the works of leslie lamport*, pages 203–226. 2019.

[12] Rui Liu, Fan Jia, Wenhao Luo, Meghan Chandarana, Changjoo Nam, Michael Lewis, and Katia P Sycara. Trust-aware behavior reflection for robot swarm self-healing. In *Aamas*, pages 122–130, 2019.

[13] Ali Makhdoumi and Asuman Ozdaglar. Graph balancing for distributed subgradient methods over directed graphs. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 1364–1371. IEEE, 2015.

[14] Frederik Mallmann-Trenn, Matthew Cavorsi, and Stephanie Gil. Crowd vetting: Rejecting adversaries via collaboration with application to multirobot flocking. *IEEE Transactions on Robotics*, 38(1):5–24, 2021.

[15] Angelia Nedić and Alex Olshevsky. Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615, 2014.

[16] Angelia Nedić, Alex Olshevsky, and Michael G Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106(5):953–976, 2018.

[17] Reza Olfati-Saber and Richard M Murray. Consensus problems in networks of agents with switching topology and time-delays. *IEEE Transactions on automatic control*, 49(9):1520–1533, 2004.

[18] Alex Olshevsky. Efficient information aggregation strategies for distributed control and signal processing. *arXiv preprint arXiv:1009.6036*, 2010.

[19] Fabio Pasqualetti, Florian Dorfler, and Francesco Bullo. Control-theoretic methods for cyberphysical security: Geometric principles for optimal cross-layer resilient control systems. *IEEE Control Systems Magazine*, 35(1):110–127, 2015.

[20] S. Pu, W. Shi, J. Xu, and A. Nedić. Push-pull gradient methods for distributed optimization in networks. *IEEE Transactions on Automatic Control*, 66(1):1–16, 2021.

[21] Michael Rabbat and Robert Nowak. Distributed optimization in sensor networks. In *Proceedings of the 3rd international symposium on Information processing in sensor networks*, pages 20–27, 2004.

[22] Venkatraman Renganathan and Tyler Summers. Spoof resilient coordination for distributed multi-robot systems. In *2017 International Symposium on Multi-Robot and Multi-Agent Systems (MRS)*, pages 135–141. IEEE, 2017.

[23] Shreyas Sundaram and Bahman Gharesifard. Distributed optimization under adversarial nodes. *IEEE Transactions on Automatic Control*, 64(3):1063–1076, 2018.

[24] Shreyas Sundaram and Christoforos N Hadjicostis. Distributed function calculation via linear iterative strategies in the presence of malicious agents. *IEEE Transactions on Automatic Control*, 56(7):1495–1508, 2010.

[25] Konstantinos I Tsianos, Sean Lawlor, and Michael G Rabbat. Consensus-based distributed optimization: Practical issues and applications in large-scale machine learning. In *2012 50th annual allerton conference on communication, control, and computing (allerton)*, pages 1543–1550. IEEE, 2012.

[26] Konstantinos I Tsianos, Sean Lawlor, and Michael G Rabbat. Push-sum distributed dual averaging for convex optimization. In *2012 ieee 51st ieee conference on decision and control (cdc)*, pages 5453–5458. IEEE, 2012.

[27] Chenguang Xi, Van Sy Mai, Ran Xin, Eyad H Abed, and Usman A Khan. Linear convergence in optimization over directed graphs with row-stochastic matrices. *IEEE Transactions on Automatic Control*, 63(10):3558–3565, 2018.

[28] Jie Xiong and Kyle Jamieson. Securearray: Improving wifi security with fine-grained physical-layer information. In *Proceedings of the 19th annual international conference on Mobile computing & networking*, pages 441–452, 2013.

[29] Michal Yemini, Angelia Nedić, Andrea J. Goldsmith, and Stephanie Gil. Characterizing trust and resilience in distributed consensus for cyberphysical systems. *IEEE Trans. Robot.*, 38(1):71–91, 2022.

[30] Erhan Çınlar. *Probability and Stochastics*. Graduate Texts in Mathematics, 261. Springer New York : Imprint: Springer, New York, NY, 1st ed. 2011. edition, 2011.