Security Protection in Cooperative Control of Multi-agent Systems

Shuai Sun¹, Yilin Mo³
1. Department of Automation and BNRist, Tsinghua University, Beijing 100084, P. R. China
E-mail: suns19@mails.tsinghua.edu.cn, ylmo@tsinghua.edu.cn.

Abstract: Due to the wide application of average consensus algorithm, its security and privacy problems have attracted great attention. In this paper, we consider the system threatened by a set of unknown agents that are both “malicious” and “curious”, who add additional input signals to the system in order to perturb the final consensus value or prevent consensus, and try to infer the initial state of other agents. At the same time, we design a privacy-preserving average consensus algorithm equipped with an attack detector with a time-varying exponentially decreasing threshold for every benign agent, which can guarantee the initial state privacy of every benign agent, under mild conditions. The attack detector will trigger an alarm if it detects the presence of malicious attackers. An upper bound of false alarm rate in the absence of malicious attackers and the necessary and sufficient condition for there is no undetectable input by the attack detector in the system are given. Specifically, we show that under this condition, the system can achieve asymptotic consensus almost surely when no alarm is triggered throughout the execution, and an upper bound of convergence rate and some quantitative estimates about the error of final consensus value are given. Finally, numerical case is used to illustrate the effectiveness of some theoretical results.

Key Words: multi-agent systems; average consensus; security protection; intrusion detection

1 Introduction

Multi-agent systems have attracted widespread attention in recent years because of their better flexibility, good scalability, and excellent computing performance[1][2]. Consensus is one of the most common tasks in multi-agent systems [3] with applications in distributed estimation and optimization [4], sensor fusion [5], distributed energy management [6], sensing scheduling [7], time synchronization [8], and so on. Under the traditional discrete-time average consensus algorithm, at each step, any agent updates its state as a weighted average of the previous state of its own and those of its neighbors. Since there is no fusion center that can monitor the behavior of all agents at any step, systems are very vulnerable to internal and external attacks [9]. Attackers can cause a series of serious problems, such as system security and internal privacy issues.

“Malicious” and “curious” attackers are two common attackers. “Malicious” attackers do not follow the average consensus algorithm but add additional input signals to the system in order to perturb the final consensus or prevent consensus. “Curious” attackers try to infer the initial states of other agents based on the update rule of the average consensus algorithm. This is extremely unfavorable in a privacy-sensitive situation.

In order to address the issues above, a number of security and privacy protection methods related to the average consensus algorithm have been proposed. In order to ensure the security of consensus, in [10][11], Sundaram and Hadjicostis used the method of parity space for fault detection to show the resilience of linear consensus network from the perspective of network topology. Pasqualetti et al discussed the relationship between consensus computation in unreliable networks and fault detection and isolation problem for linear systems and gave some attack detection and identification algorithms based on known input observer method [1]. On the other hand, to protect privacy, Huang et al. proposed an average consensus algorithm that added Laplacian noise with exponential decay characteristics to the calculation, the resulting convergence value would be a random value [12]. In [13], Manitara and Hadjicostis proposed a privacy-preserving average consensus protocol and showed that the privacy of the initial state can be guaranteed when the network topology satisfied certain conditions. Mo and Murray proposed a privacy-preserving average consensus algorithm and proved that the initial state privacy of every benign agent can be effectively protected, under mild situations [14]. In [15], Wang proposed a privacy-preserving protocol in which the state of every agent was randomly decomposed into two substates, such that the mean remained the same with only one of them revealed to other neighboring agents.

However, the security and privacy problems will become more difficult when attackers are both “malicious” and “curious”. Ruan et al. proposed a homomorphic cryptography-based approach with high computational complexity, which can guarantee privacy and security in decentralized average consensus in [16]. Nevertheless, the security problem considered in [16] was the security of communication rather than that of malicious attackers. In [3], Liu et al proposed a privacy-preserving average consensus algorithm equipped with a malicious attack detector by using the method of state estimation and used the reachable set to characterize the maximum disturbance that the attackers can introduce to the system.

In this paper, we consider the case where the system is threatened by a set of unknown attackers that are both “malicious” and “curious”. The main differences between this paper and the reference [3] are as follows.

1) We use an orthogonal projection matrix of the observation matrix of the system to construct the residual vectors, and design an attack detector, while [3] used the method of state estimation.
2) We give the necessary and sufficient condition for there is no undetectable input in the system. Further, un-
der this condition, we show that the system can achieve asymptotic consensus, and give an upper bound of the convergence rate, while the corresponding content is missing in [3].

3) We give quantitative results on the estimate error of final consensus value from the perspective of theoretical analysis when asymptotic consensus is reached. However, [3] characterized the maximum disturbance that the attackers can introduce to the system by using the method of ellipsoid approximation of reachable set, and the estimate error region may be unbounded.

The main contributions of this paper are as follows. Based on the privacy-preserving consensus algorithm proposed in [14], we design a privacy-preserving average consensus algorithm equipped with an attack detector with a time-varying exponentially decreasing threshold for every benign agent, which can guarantee the initial state privacy of every benign agent, under mild conditions. The detector will trigger an alarm if it detects the presence of malicious attackers. An upper bound of false alarm rate in the absence of malicious attacker and the necessary and sufficient condition for there is no undetectable input by the attack detector in the system are given. Under this condition, we show that the system can achieve asymptotic consensus almost surely when no alarm is triggered throughout the execution and give an upper bound of the convergence rate. Quantitative results about the estimate error of final consensus value from the perspective of theoretical analysis are further established.

The rest of this paper is organized as follows: Section 2 briefly reviews the average consensus algorithm and introduces two kinds of attack models. Sections 3 and 4 give the relevant results of privacy protection against curious attackers and security protection against malicious attackers, respectively. Section 5 gives numerical case to illustrate the effectiveness of some theoretical results and Section 6 concludes this paper.

Notations: $\mathbb{N}$ is the set of all non-negative integers. $\mathbb{R}^n$ is the set of $n \times 1$ real vectors. $\mathbb{R}^{n \times m}$ is the set of $n \times m$ real matrices. $\text{tr} M$ is the trace of square matrix $M$. $\mathbf{1}$ is an all one vector of proper dimension. $\mathbf{0}$ is an all zero vector of proper dimension. $\|v\|$ indicates the 2-norm of the vector $v$, while $\|M\|$ is the induced 2-norm of the matrix $M$. $X^+$ is the Moore–Penrose pseudoinverse of the matrix $X$. $\{a(k)\}_{k=0}^n$ stands for the finite set $\{a(0), a(1), \cdots, a(n)\}$ and $\{a(k)\}_{k=0}^\infty$ stands for the infinite set $\{a(0), a(1), \cdots\}$.

2 Problem Formulation

2.1 Average Consensus

In this subsection, we briefly introduce the average consensus algorithm.

Consider a network composed by $n$ agents as an undirected connected graph $G = (V, E)$, where $V = \{1, 2, \cdots, n\}$ is the set of agents, and $E \subseteq V \times V$ represents the communication relationship among the agents. An edge between $i$ and $j$, denoted by $(i, j) \in E$, implies that $i$ and $j$ can communicate with each other. The set of neighbors of $i$ is denoted by $\mathcal{N}_i = \{j \in V : (i, j) \in E, j \neq i\}$.

Suppose that each agent $i \in V$ has an initial state $x_i(0)$. At any time $k$, agent $i$ first broadcasts its state to all of its neighbors and then updates its own state in the following linear combination manner:

$$x_i(k + 1) = a_{ii}x_i(k) + \sum_{j \in \mathcal{N}_i} a_{ij}x_j(k). \quad (1)$$

where $a_{ij} \neq 0$ if and only if $i$ and $j$ are neighbors. Define $x(k) = [x_1(k), x_2(k), \cdots, x_n(k)]^\top \in \mathbb{R}^n$ and weight matrix $A = [a_{ij}] \in \mathbb{R}^{n \times n}$. The state updating rule can be written in the following matrix form:

$$x(k + 1) = Ax(k). \quad (2)$$

We say the agents reach a consensus if $\lim_{k \to \infty} x(k) = \gamma \mathbf{1}_{n \times 1}$, where $\gamma$ is an arbitrary scalar constant. If $\gamma = \frac{1}{n} \sum_{i=1}^n x_i(0)$, then we say the average consensus is reached.

Assume that the eigenvalues of $A$ are arranged as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. It is well known that the necessary and sufficient conditions for average consensus are as follows:

(A1) $\lambda_1 = 1$, and $|\lambda_i| < 1$, $i = 2, \cdots, n$;

(A2) $A \mathbf{1}_{n \times 1} = \mathbf{1}_{n \times 1}$.

In the rest of this paper, assume that $A$ is symmetric and satisfies Assumptions (A1) and (A2) above.

2.2 Attack Models

In this subsection, we introduce two kinds of attack models.

Malicious Attack: Some agents intend to disrupt the average consensus or prevent the consensus from being achieved by adding arbitrary input signals instead of following the updating rule (1), i.e.,

$$x_i(k + 1) = a_{ii}x_i(k) + \sum_{j \in \mathcal{N}_i} a_{ij}x_j(k) + u_i(k), \quad (3)$$

where $u_i(k) \neq 0$ is the attack signal added by agent $i$ at step $k$. Agent $i$ is said to be a malicious attacker if $u_i(k)$ is nonzero for at least one step $k, i \in \mathbb{N}$. The model for malicious attackers considered here is quite general, and the attack signal at every step can be an arbitrary deterministic value.

This kind of malicious attackers can potentially either prevent benign agents, who follow the standard update rule (1), from reaching a consensus or manipulate the final consensus value to be arbitrary.

Curious Attack: Some agents intend to infer the initial states of other agents, which may not be desirable when the initial state privacy is of concern. Such agents are called curious attackers.

In this paper, we deal with a set of unknown agents that are both “malicious” and “curious”. We denote that the set of these unknown agents that are both “malicious” and “curious” as $\{i_1, \cdots, i_p\}$. Meanwhile, assume that each agent knows the weight matrix $A$ defined in (2).

3 Privacy Protection Against Curious Attackers

In this section, we address the issues caused by curious attackers inferring other benign agents’ initial states.

3.1 Privacy Preserving Consensus Algorithm

In order to protect each benign agent’s privacy, we adopt the privacy-preserving algorithm proposed in [14]. For the sake of completeness, we briefly describe the algorithm as follows.
Algorithm 3.1. Let $v_i(k)(i = 1, 2, \ldots, n; k = 0, 1, \cdots)$ be standard normal distributed random variables, which are independent across $i$ and $k$. Denote $v(k) \triangleq [v_1(k), v_2(k), \cdots, v_n(k)]^\top$. Based on $v(k)$ we can construct the following noisy signals

$$w(k) = \begin{cases} v(0), & \text{if } k = 0; \\ \varphi^k v(k) - \varphi^{k-1} v(k-1), & \text{otherwise}; \end{cases}$$

where $0 < \varphi < 1$ is a constant.

To protect their true states, the agents add noisy signals $w(k)$ into their states $x(k)$ and form a new state vector $x^+(k)$, before sharing with their neighbors, i.e., $x^+(k) = x(k) + w(k)$.

Remark 3.1. According to [14], the privacy-preserving average consensus algorithm guarantees the initial state privacy of every benign agent, under mild conditions, and that random noises introduced to the consensus process do not affect the consensus result.

Under this privacy-preserving algorithm, since the set of these unknown agents that are both “malicious” and “curious” are $\{i_1, \cdots, i_p\}$, the state updating rule is as follows:

$$x(k + 1) = Ax^+(k) + Bu(k) = A(x(k) + w(k)) + Bu(k),$$

where $B = [e_{i_1}, e_{i_2}, \cdots, e_{i_p}]$ with $e_i$ being the $i$th canonical basis vector in $\mathbb{R}^n$, and $u(k) = [u_{i_1}(k), u_{i_2}(k), \cdots, u_{i_p}(k)]^\top$ is the attack input signal at step $k$.

Theorem 3.1 ([14]). The initial state $x_j(0)$ of agent $j$ is kept private from these curious agents $\{i_1, \cdots, i_p\}$ if and only if $\mathcal{N}_j \cup \{j\} \not\subseteq \mathcal{N}_{i_1} \cup \cdots \cup \mathcal{N}_{i_p} \cup \{i_1, \cdots, i_p\}$.

4 Security Protection Against Malicious Attackers

In this section, we address the concerns of malicious attackers disturbing the average consensus or preventing consensus.

4.1 Attack Detector

In order to deal with malicious attacks, we will design an attack detector for each benign agent. Without loss of generality, assume that agent 1 is benign, and we focus on designing an attack detector for agent 1. Suppose that the neighbors of agent 1 are $\{j_1, j_2, \cdots, j_{n-1}\}$. The values that are available to agent 1 at $k$-th step will be denoted by

$$y(k) = C(x(k) + w(k)),$$

where $C = [e_1, e_{j_1}, e_{j_2}, \cdots, e_{j_{n-1}}]^\top$.

We first propose a residual generator as follow, which uses the measurement sequence $\{y(k)\}_{k=0}^\infty$ to generate a residual vector sequence $\{r(k)\}_{k=0}^\infty$ that will be a zero vector sequence when there is no noise protecting the agent’s privacy and malicious attackers in the system. The response of linear consensus system of the form (5) and (6) over $n + 1$ steps at each step $k$ is given by

$$
\begin{bmatrix}
    y(k) \\
    y(k + 1) \\
    \vdots \\
    y(k + n)
\end{bmatrix}
= \begin{bmatrix}
    C \\
    CA \\
    \vdots \\
    CA^n
\end{bmatrix}
\begin{bmatrix}
    x(k) \\
    w(k) \\
    \vdots \\
    w(k + n)
\end{bmatrix}
+ \begin{bmatrix}
    0 & 0 & \cdots & 0 \\
    CA & C & \cdots & 0 \\
    \vdots & \vdots & \ddots & \vdots \\
    CA^{n-1}B & CA^{n-2}B & \cdots & 0
\end{bmatrix}
\begin{bmatrix}
    u(k) \\
    u(k + 1) \\
    \vdots \\
    u(k + n)
\end{bmatrix}.
$$

Now, we are ready to proceed with the construction of residual generator which will be used to design an attack detector. In order to make the residual generator not affected by the initial state $x(0)$, we multiply the orthogonal projection matrix $P = I_{m(n+1)} - C_nC_n^\top$ on both the left and right sides of (7), and (7) can be simplified to the following form:

$$P Y_{[k,k+n]} = P \mathcal{H}_n W_{[k,k+n]} + P \mathcal{J}_n U_{[k,k+n]}.$$  

Based on the above results, we can construct the following residual generator, then use it to design an attack detector.

Definition 4.1 (Residual Vector). Define the residue vector $r(k)$ at each step $k$ as below:

$$r(k) \triangleq PY_{[k,k+n]}.$$  

Then a malicious attack detector is obtained, which compares $\|r(k)\|$ with a threshold $c \rho^k$ decreasing exponentially over time and triggers an alarm if and only if $\|r(k)\|$ is greater than the given threshold $c \rho^k$ at some step $k$, where $c > 0$, $\varphi < \rho < 1$ are two fixed constants selected by agent 1.

If there is an alarm triggered at some step $k$, we will regard that there are malicious attackers in the system.

Remark 4.1. It can be seen that there is a delay of $n$ steps in detection at each step $k$. In general, the delay is inevitable, because at any step $k$, we cannot get enough information about the attack input using the observations up to the current moment.

4.2 False Alarm Rate

In this subsection, we will focus on the situation where there is no malicious attacker in the system. According to the privacy-preserving consensus algorithm, the noisy signal $w(k)$ will be added into agents’ states $x(k)$ at each step $k$. Therefore, even if there is no malicious attacker, an alarm may be triggered at some step $k$. False alarm rate of the attack detector will be characterized here.
According to the linearity of linear consensus system of the form (5) and (6) and the definition of residual vector \( r(k), r'(k) \) can be decomposed into the following two parts:
\[
r(k) = r^a(k) + r^n(k), \tag{10}
\]
where \( r^a(k) \) and \( r^n(k) \) are respectively generated by malicious agents’ input and the noisy signals. By using (8), we can directly get the following equivalent relationship:
\[
r^n(k) = \mathcal{P} \mathcal{J}_n U_{[k,k+n]}, \quad r^n(k) = \mathcal{P} \mathcal{H}_n W_{[k,k+n]}.
\tag{11}
\]
Now we define false alarm rate as follows.

**Definition 4.2** (False Alarm Rate). Define false alarm rate \( \alpha \) as the probability of triggering false alarm at least once from the initial step to infinity when there is no malicious attacker, in other words,
\[
\alpha \triangleq \mathbb{P} \left\{ \| r^n(k) \| > c \rho^k \right\}. \tag{12}
\]

**Remark 4.2.** Under the same conditions, a smaller \( \alpha \) often means the better performance of the corresponding attack detector.

Before characterizing false alarm rate \( \alpha \), we first focus on \( r^n(k) \). For the convenience of notation, the matrix \( \mathcal{P} \mathcal{H}_n \) is uniformly partitioned according to the columns as \([P_0 \ P_1 \ \cdots \ P_n]\), where each \( P_i \) is of dimension \( m(n+1) \times n \) and \( P_0 = \mathcal{P} \mathcal{O}_n = 0_{p \times n} \). Combining it with the definition of \( w(k) \), for any \( k \in \mathbb{N} \), \( r^n(k) \) can then be expressed as
\[
r^n(k) = \sum_{i=0}^{n-1} \varphi^{k+i} (P_i - P_{i+1}) w(k+i) + \varphi^{k+n} P_n w(k+n). \tag{13}
\]
The next theorem gives an upper bound of false alarm rate \( \alpha \).

**Theorem 4.1** (An Estimation Of False Alarm Rate). For a linear consensus system of the form (5) and (6), false alarm rate \( \alpha \) of the attack detector above satisfies
\[
\alpha \leq \frac{1}{e} \rho^2 - \frac{\rho^2}{e} \left( \sum_{i=0}^{n-1} \varphi^{2i} \text{tr} \left[ (P_i - P_{i+1})^\top (P_i - P_{i+1}) \right] + \varphi^{2n} \text{tr} \left[ P_n^\top P_n \right] \right). \tag{14}
\]
For specific proof, please refer to [17].

### 4.3 Detectability

In this subsection, we will focus on the detectability of the attack detector.

**Definition 4.3** (Undetectable Input). For a linear consensus system of the form (5) and (6), the attack input \( u \) introduced by these unknown malicious agents \( \{i_1, \cdots, i_p\} \) is undetectable if
\[
\exists x_1, x_2 \in \mathbb{R}^n, \ s.t. \forall k \in \mathbb{N}, y^{0+u}(x_1, u, k) = y^{0+u}(x_2, 0, k),
\]
where \( y^{0+\alpha}(x_1, u, k) \) is the part of \( y(x_1, u, k) \) generated by the initial state \( x_1 \) and the attack input \( u \) at step \( k \).

**Definition 4.4** (Undetectable Input By The Attack Detector). For a linear consensus system of the form (5) and (6), the attack input \( u \) introduced by these unknown malicious agents \( \{i_1, \cdots, i_p\} \) is undetectable by the attack detector if
\[
\exists x_1, x_2 \in \mathbb{R}^n, \ s.t. \forall k \in \mathbb{N}, r^n(x_1, u, k) = r^n(x_2, 0, k).
\]
Now, we give the necessary and sufficient conditions for there exists no undetectable input by the attack detector.

**Theorem 4.2.** For a linear consensus system of the form (5) and (6), the following statements are equivalent almost surely:
1) there is no undetectable input;
2) there is no undetectable input by the attack detector;
3) \( \text{rank} [O_{n-1} \ J_{n-1}] - \text{rank} [J_{n-1}] = n. \)

For specific proof, please refer to [17]. **Remark 4.3.** Theorem 4.2 states that there is no attack input that can be detected but cannot be detected by the attack detector.

**Theorem 4.3.** For a linear consensus system of the form (5) and (6), there exists a \( p \times m(n+1) \) matrix \( Q_B \) \(^3\) that satisfies
\[
Q_B r^n(u) = u(k), \forall k \in \mathbb{N}, \tag{15}
\]
if \( \text{rank} [O_{n-1} \ J_{n-1}] - \text{rank} [J_{n-1}] = n. \)

For specific proof, please refer to [17]. **Theorem 4.3 will be used later.**

### 4.4 Asymptotic Consensus and Error

In order to protect the security and privacy of the system, we have designed a privacy-preserving average consensus algorithm equipped with an attack detector with a time-varying exponentially decreasing threshold for every benign agent. At this point, there are naturally three problems:

- When there exists no undetectable inputs and no alarm is triggered throughout the execution, will the system eventually achieve consensus?
- If the system can achieve a final consensus, what is the rate of convergence?
- How much error of the final consensus value will be?

We will answer these three problems in turn in this subsection.

First we define convergence rate here.

**Definition 4.5** (Convergence Rate). Define the convergence rate \( \varrho \) of consensus algorithm as
\[
\varrho \triangleq \limsup_{k \to \infty} \| x(k) - \pi(k) \|^2,
\tag{16}
\]
whenever the limit on the right-hand side exists, where \( \pi(k) = \frac{1}{n} x(k) \) denotes the average state vector at step \( k \).

**Theorem 4.4.** For a linear consensus system of the form (5) and (6), an asymptotic consensus will be reached almost surely, i.e. \( \lim_{k \to \infty} x(k) = \pi(k) \overset{\text{a.s.}}{=} 0_{n \times 1} \), and the convergence rate \( \varrho \) satisfies \( \varrho \leq \max\{\rho, |\lambda_2|, |\lambda_n|\} \), if
1) no alarm is triggered;

\(^3\)The subscript \( B \) here means that the matrix \( Q_B \) is related to \( B \).
2) rank $[O_{n-1} \ J_{n-1}] - \text{rank} \ [J_{n-1}] = n$.

For specific proof, please refer to [17].

Now we come to answer the third problem raised at the beginning of this subsection, namely how to estimate the error of the final consensus value.

**Definition 4.6. (The Error Of The Final Consensus Value)**
The error $e$ of the final consensus value of the system of the form (5) and (6) is defined as follow

$$
eq \frac{1}{n} \sum_{i=0}^{\infty} u(i). \tag{17}$$

**Remark 4.4.** The error above is well-defined, for the noisy signals have no effect on the final consensus value, which has been proved in [17].

It is worth noting that the estimation error needs to be carried out under the premise that no alarm is triggered throughout the execution in the system. According to the definition of “not triggering an alarm”, we have $\|r(k)\| \leq c \rho^k$ for all step $k$. By using (11), (15), Algorithm 3.1 and summing the step $k$ from 0 to infinity, we have

$$\sum_{k=0}^{\infty} u(k) = Q_B^2 \left( \sum_{k=0}^{\infty} r(k) + \sum_{i=1}^{n} \varphi_k^i P_i v(i-1) \right). \tag{18}$$

Substituting the above equality into the definition of error $e$, we can get

$$e = \frac{1}{n} \sum_{i=0}^{\infty} u(k) \left( \sum_{k=0}^{\infty} r(k) + \sum_{i=1}^{n} \varphi_k^i P_i v(i-1) \right). \tag{19}$$

For convenience and simplicity of notation, let $s_B$ and $T_B$ denote $1_{x \times p} Q_B \sum_{k=0}^{\infty} r(k)$ and $\sum_{k=0}^{\infty} \varphi_k^i P_i v(i-1)$ respectively. Since $v_i(k)(i = 1, 2, \ldots, n; k = 0, 1, \ldots)$ are standard normal distributed random variables, which are independent across $i$ and $k$, $T_B$ is also a normal distributed random variable with $\mathbb{E}[T_B] = 0$ and $\text{Var}[T_B] = \sum_{k=0}^{\infty} \varphi_k^i P_i v(i-1)$. For any $0 < \beta < 1$, let the point $z_{B,\beta/2}$ satisfy $\mathbb{P} [T_B > z_{B,\beta/2}] = \beta/2$. Therefore, we have

$$\mathbb{P} [T_B < z_{B,\beta/2}] = 1 - \beta.$$

Since $e = s_B + T_B$, it follows that $\mathbb{P} [e < s_B] = z_{B,\beta/2} = 1 - \beta$ for any $e \in \mathbb{R}$.

Now let $\mu_B = \frac{1}{n} \sum_{i=0}^{\infty} u(i)$, and $T_B = \frac{1}{n} \sum_{i=0}^{\infty} u(i)$. Under the detectable condition rank $[O_{n-1} \ J_{n-1}] - \text{rank} \ [J_{n-1}] = n$, “no alarm is triggered” implies that $|s_B| \leq \mu_B$ holds because of Cauchy-Schwarz inequality. Therefore, we can get that

$$\mathbb{P} [\mu_B - z_{B,\beta/2} \leq e \leq \mu_B + z_{B,\beta/2}] \geq 1 - \beta \text{ for any } e \in \mathbb{R}, \text{ i.e., } [-\mu_B - z_{B,\beta/2}, \mu_B + z_{B,\beta/2}] \text{ is a confidence interval for } e \text{ with confidence coefficient of not less than } 1 - \beta.$$

Based on the results above and Theorem 4.2, we can get the following theorem.

**Theorem 4.5.** For a linear consensus system of the form (5) and (6), when an asymptotic consensus is reached, for any $0 < \beta < 1$, $\bigcup_B \left[ -\mu_B - z_{B,\beta/2}, \mu_B + z_{B,\beta/2} \right]$ is a confidence interval for $e$ with confidence coefficient of not less than $1 - \beta$, if the following statements hold:

1) no alarm is triggered;
2) rank $[O_{n-1} \ J_{n-1}] - \text{rank} \ [J_{n-1}] = n$.

**5 Numerical Examples**

Consider the following network composed of 4 agents:

![Fig. 1: Network Topology](image)

Suppose that the weight matrix is

$$A = \begin{bmatrix} 0.136 & 0.461 & 0 & 0.403 \\ 0.461 & 0.153 & 0.386 & 0 \\ 0 & 0.386 & 0.278 & 0.336 \\ 0.403 & 0 & 0.336 & 0.261 \end{bmatrix},$$

which is generated randomly. Suppose that the initial state of agents are $x(0) = [100 \ -50 \ 50 \ -100]^T$. Without loss of generality, assume that agent 1 is benign and it is running an attack detector. Then the matrix $C$ is $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

Suppose that agent 3 is both malicious and curious and other agents are all benign. Since for any agent $j, j = 1, 2, 4, \mathcal{N}_i \cup \{i\} \in \mathcal{N}_i \cup \{j\}$, according to Theorem 4.1, the initial state privacy of every benign agent is guaranteed. Since rank $[O_{3} \ J_{3}] - \text{rank} \ [J_{3}] = 4$, according to Theorem 4.2, there is no undetectable input by the attack detector. In order to avoid being detected, agent 3 injects the attack signals $u_3(k) = -24 \times 0.2^k$ into the system at every step $k$.

![Fig. 2: One snapshot of the comparison between $||r(k)||$ and $c \rho^k$](image)

Suppose that agent 1 selects these parameters as follow: $c = 16.2, \rho = 0.7, \varphi = 0.2$. According to Theorem 4.1, false alarm rate $\alpha$ is no more than 0.01. Since that no
alarm is triggered after 2000 steps, and the state values of the neighbors of agent 1 and its own state value have always been $-7.5000$ since the 30-th step, it can be considered that the system has achieved an asymptotic consensus. According to Theorem 4.4, when an asymptotic consensus is achieved, the convergence rate $\rho \leq \max\{\rho, |\lambda_2|, |\lambda_n|\} = \max\{0.7, 0.2229, -0.6057\} = 0.7$. One snapshot of the comparison between $\|r(k)\|$ and $\epsilon \rho^k$ and the trajectories of agents’ state are shown in Fig. 2, Fig 3, respectively. From Fig 3, it can be seen that although an asymptotic consensus is achieved, the final convergence value $-7.5000$ is not the average value 0 of the initial state $x(0)$.

Now, agent 1 begins to estimate the error of the final convergence value. Since agent 1 does not know which agents are malicious attackers, according to Theorem 4.5, it needs to consider all cases that meet the detectable condition $\lambda_1 \lambda_2 \lambda_n = 0$. According to Theorem 4.5, if $\beta = 0.001$, $[-57.9926, 57.9926]$ is a confidence interval for $e$ with confidence coefficient of not less than 0.999. If agent 1 has known that there is at most one malicious attacker in the system, $[-29.5478, 29.5478]$ is a confidence interval for $e$ with confidence coefficient of not less than 0.999.

6 Conclusion

In this paper, we deal with the case that the consensus system is threatened by a set of unknown agents that are both “malicious” and “curious”. We propose a privacy-preserving average consensus algorithm equipped with an attack detector with a time-varying exponentially decreasing threshold, for every benign agent, which can guarantee the initial state privacy of every benign agent, under mild conditions. An upper bound of false alarm rate and the necessary and sufficient condition for there is no undetectable input by the attack detector in the system are given. We prove that the system can achieve asymptotic consensus almost surely and give an upper bound of convergence rate and some estimates about the error.

References

[1] Pasqualetti F., Bicchi A., Bullo F. Consensus Computation in Unreliable Networks: A System Theoretic Approach[J]. IEEE Transactions on Automatic Control, 2012, 57(1):90-104.
[2] Wang, Jian, Liangren Shi, and Xinpeng Guan. “Semi-global leaderless consensus of linear multi-agent systems with actuator and communication constraints.” Journal of Systems Science and Complexity 33 (2020): 882-902.
[3] Liu, Qipeng, Xiaoying Ren, and Yihui Mo. “Secure and privacy preserving average consensus.” 2017 11th Asian Control Conference (ASC). IEEE, 2017.
[4] Pasqualetti, F., Carli, R., Bicchi, A., & Bullo, F. (2010). Distributed estimation and detection under local information. IFAC Proceedings Volumes, 43(19), 263-268.
[5] Xiao, L., Boyd, S., & Lall, S. (2005, April). A scheme for robust distributed sensor fusion based on average consensus. In IPSN 2005. Fourth International Symposium on Information Processing in Sensor Networks, 2005. (pp. 63-70). IEEE.
[6] Zhao, C., He, J., Cheng, P., & Chen, J. (2016). Consensus-based energy management in smart grid with transmission losses and directed communication. IEEE Transactions on Smart Grid, 8(5), 2049-2061.
[7] He, J., Duan, L., Hou, F., Cheng, P., & Chen, J. (2015). Multi-period scheduling for wireless sensor networks: A distributed consensus approach. IEEE Transactions on Signal Processing, 63(7), 1651-1663.
[8] Schenato, L., & Fiorentini, F. (2011). Average timesynch: A consensus-based protocol for clock synchronization in wireless sensor networks. Automatica, 47(9), 1878-1886.
[9] Michiardi, P., & Molva, R. (2002). Core: a collaborative reputation mechanism to enforce node cooperation in mobile ad hoc networks. In Advanced communications and multimedia security (pp. 107-121). Springer, Boston, MA.
[10] Sundaram, Shreyas, and Christoforos N. Hadjicostis. “Distributed function calculation via linear iterations in the presence of malicious agents? Part I: Attacking the network.” 2008 American Control Conference. IEEE, 2008.
[11] Sundaram, Shreyas, and Christoforos N. Hadjicostis. “Distributed function calculation via linear iterations in the presence of malicious agents? Part II: Overcoming malicious behavior.” 2008 American Control Conference. IEEE, 2008.
[12] Huang, Z., Mitra, S., & Dullerud, G. (2012, October). Differentially private iterative synchronous consensus. In Proceedings of the 2012 ACM workshop on Privacy in the electronic society (pp. 81-90). ACM.
[13] Manitara, Nicolaos E., and Christoforos N. Hadjicostis. “Privacy-preserving asymptotic average consensus.” 2013 European Control Conference (ECC). IEEE, 2013.
[14] Mo Y., Murray R M . Privacy Preserving Average Consensus[J]. IEEE Transactions on Automatic Control, 2017, 62(2):753-765.
[15] Wang, Yongqiang. “Privacy-preserving average consensus via state decomposition.” IEEE Transactions on Automatic Control 64.11 (2019): 4711-4716.
[16] Ruan, Minghao, Muaz Ahmad, and Yongqiang Wang. “Secure and privacy-preserving average consensus.” Proceedings of the 2017 workshop on cyber-physical systems security and privacy. 2017.
[17] Shuai Sun and Yilin Mo. Security Protection in Cooperative Control of Multi-agent Systems, 2021; arXiv:2105.02618.