Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Towards resilient average consensus in multi-agent systems: a detection and compensation approach*&#

Chongrong FANG[1], Wenzhe ZHENG[1], Zhiyu HE[1], Jianping HE‡[1], Chengcheng ZHAO[2], Jingpei WANG[3]

[1]*Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China*

[2]*State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou 310027, China*

[3]*Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing 100094, China*

E-mail: crfang@sjtu.edu.cn; wzzheng@sjtu.edu.cn; hzy970920@sjtu.edu.cn; jphe@sjtu.edu.cn; chengchengzhao@zju.edu.cn; wjp@csu.ac.cn

Received June 28, 2023; Revision accepted Nov. 6, 2023; Crosschecked Dec. 8, 2023; Published online Dec. 28, 2023

Abstract: Consensus is one of the fundamental distributed control technologies for collaboration in multi-agent systems such as collaborative handling in intelligent manufacturing. In this paper, we study the problem of resilient average consensus for multi-agent systems with misbehaving nodes. To protect consensus value from being influenced by misbehaving nodes, we address this problem by detecting misbehaviors, mitigating the corresponding adverse impact, and achieving the resilient average consensus. General types of misbehaviors are considered, including attacks, accidental faults, and link failures. We characterize the adverse impact of misbehaving nodes in a distributed manner via two-hop communication information and develop a deterministic detection compensation based consensus (D-DCC) algorithm with a decaying fault-tolerant error bound. Considering scenarios wherein information sets are intermittently available due to link failures, a stochastic extension named stochastic detection compensation based consensus (S-DCC) algorithm is proposed. We prove that D-DCC and S-DCC allow nodes to asymptotically achieve resilient accurate average consensus and unbiased resilient average consensus in a statistical sense, respectively. Then, the Wasserstein distance is introduced to analyze the accuracy of S-DCC. Finally, extensive simulations are conducted to verify the effectiveness of the proposed algorithms.

Key words: Resilient consensus; Multi-agent systems; Malicious attacks; Detection; Compensation
https://doi.org/10.1631/FITEE.2300467 **CLC number:** TP13

1 Introduction

Industrial cyber-physical systems (ICPSs) are systems tightly integrating computing, network, and control technologies, and are broadly applied in multiple areas such as intelligent manufacturing. Cooperative tasks in ICPSs, such as collaborative handling in smart factories, could be accomplished by multi-agent systems through assigning well-designed subtasks to multiple nodes in a distributed manner. This is an important cooperation paradigm for ICPSs, especially in dynamic environments. Formation control is one of the key functions for multi-agent

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (Nos. 62103266, 61972345, and U1911401) and the State Key Laboratory of Industrial Control Technology, China (No. ICT2023A03)

& A preliminary version of this paper was presented at Proc 60th IEEE Conf on Decision and Control

Electronic supplementary materials: the online version of this article (https://doi.org/10.1631/FITEE.2300467) contains supplementary materials, which are available to authorized users

ORCID: Chongrong FANG, https://orcid.org/0000-0003-2357-0228; Jianping HE, https://orcid.org/0000-0002-6253-7802

systems to accomplish such tasks, where consensus plays a key role in collaboration (Wang et al., 2014). Consensus aims to achieve global agreements through exchanges of local information among multiple agents by predefined consensus protocols. Recently, the security of multi-agent systems has attracted much attention, because these systems usually run in an open environment and are vulnerable to attack or failure (He et al., 2013; Ma et al., 2022). Even minor failures or malicious attacks on consensus will have an unexpected impact on formation control, thereby affecting the execution of cooperative tasks. Hence, resilient consensus algorithms have been widely investigated to ensure the security of distributed multi-agent systems under faults or malicious attacks.

Against this backdrop, there are a series of studies focusing on resilient consensus (LeBlanc et al., 2013). On one hand, a representative kind of research is studies developed based on mean-subsequence reduced (MSR) algorithms (Kieckhafer and Azadmanesh, 1994). The main idea of these studies is that each agent ignores the extreme states collected from neighbors, while updating its own state with the remaining information. Differently, weighted-mean-subsequence reduced (W-MSR) algorithms (LeBlanc et al., 2013) are developed which do not remove all extreme states as those in the MSR algorithm, but rather adopt a system wherein each agent discards only the extreme states that are strictly larger or smaller than their own states. In addition, Dibaji et al. (2018) proposed a quantized version of the W-MSR algorithm to deal with asynchronous and time-varying time delays. While the abovementioned MSR-based algorithms enable the consensus to be achieved within the range or convex hull of the initial states of normal agents, the accurate average consensus is difficult to guarantee. Additionally, the resilient consensus control strategies are investigated. These involve the development of resilient controllers for ensuring toleration in the event of attacked data or agents, thereby guaranteeing the survivability of the multi-agent systems under misbehaviors. For example, Ge et al. (2023) developed a resilient and safe platooning control strategy for connected automated vehicles under intermittent denial-of-service (DoS) attacks based on a heterogeneous and uncertain vehicle longitudinal dynamic model. Xie et al. (2022) proposed a proportional-integral-observer-based controller to achieve the desired platooning performance under replay attacks. Such resilient studies are also applied to specific practical scenarios, e.g., the dispatch problem in a smart grid (Wen et al., 2021). Yang et al. (2021) tackled the economic dispatch problem of a smart grid under DoS attacks by leveraging a novel distributed event-triggered scheme and an improved multi-agent consensus protocol.

On the other hand, detection-isolation-based methods are also adopted to achieve resilient consensus. The main idea of such studies is to detect anomalous agents while the predefined detection criteria are breached and then to isolate the identified abnormal agents in case they continue to affect the execution of collaboration tasks. Specifically, observer-based detection methods are effective for the detection of abnormal behaviors in the system. Pasqualetti et al. (2012) proposed an observer-based method for synchronous consensus in directed networks, where networks need to be highly connected and agents require global knowledge. Zhao et al. (2018) proposed a mobile-detector-based method, which exploits mobile agents as observers. This approach extends the number of tolerable attacks that are decided by network connectivity. Observer-based techniques are also used in fault detection for interconnected second-order systems (Shames et al., 2011). Gentz et al. (2016) investigated the detection and mitigation in randomized gossiping algorithms based on the observation of temporal or spatial differences in systems of data injection attacks. Multiple communication-information-based approaches, e.g., two-hop information (He et al., 2013; Yuan and Ishii, 2021; Ramos et al., 2022), also contribute to detection. Specifically, Ramos et al. (2022) proposed methods to achieve consensus in multiple distinct subsets of agents, where each normal agent crosschecks its state among different subsets. There will be malicious agents among the subsets with the same values. Based on majority voting, Yuan and Ishii (2021) designed a detection scheme with two-hop information for which the constraint on the graph structure is less stringent than that for MSR algorithms. However, the above-mentioned detection-isolation-based algorithms might mistakenly identify faulty agents as malicious ones when faults occasionally appear, which results in loss of information and system capacity. Additionally,

those detection-isolation-based methods cannot remove adverse impacts introduced before isolation by malicious agents. The research into resilient average consensus (Hadjicostis et al., 2012) considers unreliable heterogeneous communication links, but does not consider misbehaving agents including malicious and faulty ones.

Therefore, in this paper, we are motivated to design resilient average consensus algorithms for multi-agent systems to defend against the adverse impact on the consensus that is brought by misbehaving agents, including malicious and faulty ones. To achieve better performance of resilient average consensus, we adopt the idea of detecting and compensating for the adverse impact of misbehaving agents with two-hop communication information and providing tolerance for faulty agents. In this work, we propose a method to detect misbehaving agents and estimate the corresponding adverse impact, followed by a compensation scheme to mitigate or eliminate the adverse impact. Specifically, the main contributions are summarized as follows:

1. We investigate the problem of resilient average consensus under misbehaving agents. Based on two-hop communication information, we design detection–isolation–mitigation-based methods to detect and isolate misbehaving agents and compensate for the detected errors, thereby achieving resilient accurate average consensus and unbiased resilient average consensus in a statistical sense for deterministic and stochastic scenarios, respectively.

2. We design a deterministic detection compensation based consensus (D-DCC) algorithm for normal agents using two-hop communication information. We prove that the D-DCC algorithm can detect and compensate for the impact of misbehaving agents on consensus, thereby achieving resilient average consensus exactly.

3. We further consider the scenario wherein the communication links could fail due to accidents. In this case, we propose a stochastic detection compensation based consensus (S-DCC) algorithm correspondingly. We also prove that S-DCC can achieve unbiased resilient average consensus in a statistical sense. Moreover, we analyze the accuracy of S-DCC by the Wasserstein distance. We present extensive evaluations to demonstrate the effectiveness of the proposed methods.

Compared with our conference version (Zheng et al., 2021), in this work, we have enriched design details and performance analysis of the resilient consensus algorithm for the deterministic scenario (i.e., D-DCC algorithm) with illustrative examples and discussions. Additionally, for the stochastic scenario, we have relaxed the requirement that the expectation of faults should be zero and provided proof of the resilient consensus performance of the S-DCC algorithm. We have also theoretically analyzed the performance of S-DCC by the Wasserstein distance to show the accuracy of the S-DCC algorithm. Further, we have presented in-depth discussions on assumption relaxation and offered more numerical results in terms of comparisons with related studies.

Notations: $\mathbb{R}$ denotes the set of real numbers, and 1 denotes the vector of all ones with proper dimension. Given a matrix M, M^{T} is its transpose matrix. Given a set $\mathcal{V}$, $|\mathcal{V}|$ is the number of elements in the set and $\mathcal{V}/\{i\}$ is the set removing i from $\mathcal{V}$. We let $\mathbb{E}$ and $\mathbb{D}$ denote expectation and variance operations, respectively.

2 Problem statement

2.1 Network description

In this paper, we consider a multi-agent system, and its network is modeled by an undirected graph $\mathcal{G}=(\mathcal{V},\mathcal{E})$, where $\mathcal{V}=\{1,2,\dots,N\}$ and $\mathcal{E}\subseteq\mathcal{V}\times\mathcal{V}$ are the vertex set and edge set, respectively. The edge $(i,j)\in\mathcal{E}$ means that agents i and j can communicate mutually. The term $\mathcal{N}_i=\{j\mid(i,j)\in\mathcal{E}\}$ represents the set of neighbors for agent i. We denote by $A_{\mathcal{G}}=[a_{ij}]_{N\times N}$ and $L=D_{\mathcal{G}}-A_{\mathcal{G}}$ the adjacency matrix and Laplacian matrix, respectively, where we have that $D_{\mathcal{G}}=\mathrm{diag}(d_1,d_2,\dots,d_N)$ and $d_i=\sum_{j=1}^{N}a_{ij}$. Let $d_{\mathrm{m}}=\max\{d_1,d_2,\dots,d_N\}$. Without loss of generality, we denote by $\mathcal{V}_{\mathrm{s}}=\{1,2,\dots,n\}$ and $\mathcal{V}_{\mathrm{m}}=\{n+1,n+2,\dots,N\}$ the set of normal agents and the set of misbehaving agents, respectively. Note that we have $\mathcal{V}_{\mathrm{s}}\cap\mathcal{V}_{\mathrm{m}}=\varnothing$ and $\mathcal{V}_{\mathrm{s}}\cup\mathcal{V}_{\mathrm{m}}=\mathcal{V}$.

2.2 Consensus protocol

The state vector of the system at time k is given by $x(k)=[x_1(k),x_2(k),\dots,x_N(k)]^{\mathrm{T}}$, where $x_i(k)\in\mathbb{R}$ is the state of agent i. The consensus protocol aims to drive all items in the state vector to the same value. Specifically, we say that the consensus is achieved when $\lim_{k\to\infty}x_i(k)=c$, $\forall i\in\mathcal{V}$, where

the value c is called the consensus value. Particularly, the average consensus can be reached if it holds that $c = \sum_{i=1}^{N} x_i(0)/N$. The basic discrete-time linear average consensus is represented as

$$x(k+1) = Wx(k), \tag{1}$$

where the weight matrix W is doubly stochastic. For each weight w_{ij} in W, we have $w_{ij} \neq 0$ when agents i and j are neighboring. By Eq. (1), the system will achieve the average consensus exponentially if the undirected graph $\mathcal{G}$ is connected. Typical weight matrices, which can guarantee asymptotic convergence, include Metropolis weights (Xiao et al., 2005) and Perron weights, i.e., $W = I - \gamma L$, where $0 < \gamma < 1/d_{\mathrm{m}}$. Under both Metropolis weights and Perron weights, the weights of agent i (namely, w_{ij}, $\forall j \in \mathcal{N}_i$) are available to neighbors if the number of neighbors $|\mathcal{N}_i|$ and N are known by neighbors.

2.3 Information set and misbehavior model

Information set: As indicated in Eq. (1), agents will update their states in a distributed manner based on their own state and those of their adjacent agents. Such two-hop information (i.e., the state of an agent and those of the agent's neighbors) assists in effective detection of misbehaving agents (He et al., 2013; Zhao et al., 2018; Yuan and Ishii, 2021). Specifically, the information set of agent i at time k, $\Psi_i(k)$, is given by

$$\Psi_i(k) = \left\{ i, \pi_i(k), x_i(k), \varepsilon_i(k-1), \left\{ j, x_j^{(i)}(k-1), j \in \mathcal{N}_i \right\} \right\},$$

in which the term $x_j^{(i)}(k-1)$ represents the state of agent j at time $k-1$, which will be sent by agent i to its neighbors at time k. The term $\varepsilon_i(k)$ indicates the compensation value injected by normal agents or the adverse impact brought by misbehaving agents, which will be discussed in detail later. The binary attack detection indicator $\pi_i(k) = 1$ or $\pi_i(k) = 0$ represents "attack" or "no attack," respectively, where the former indicates that the agent i has detected misbehaving agents among its neighbors. Note that for normal agents, $\varepsilon_i(k-1)$ is allowed to be non-zero when $\pi_i(k) = 1$. This implies that the compensation will be added when a misbehavior is detected. Every time an information transmission is made, all agents transmit their own information set $\Psi_i(k)$ $(i \in \mathcal{N})$ to neighbors. Therefore, each agent can easily obtain the two-hop information.

Misbehavior model: As for the misbehavior model, we consider that the misbehaving agents can be either faulty or malicious. Faulty agents may cause adverse impacts on the system because of accidental faults, e.g., miscalculations. Malicious agents aim to disrupt the network functions by manipulating the information set, but such an agent can send only the same information to all of its neighbors at each time slot. This is common, especially when the network is implemented by broadcast communication. In the following, we make four assumptions regarding misbehaviors and misbehaving agents:

Assumption 1 No two misbehaving agents are adjacent.

Assumption 2 Malicious agents can modify the information set by changing their own state values as well as those of neighbors, but will not add any entries.

Assumption 3 A normal agent will no longer communicate with the agent(s) that is (are) to be isolated.

Assumption 4 The subgraph $\mathcal{G}_{\mathrm{s}} = (\mathcal{V}_{\mathrm{s}}, \mathcal{E}_{\mathrm{s}})$ is connected, where $\mathcal{E}_{\mathrm{s}} \subseteq \mathcal{E}$ denotes the edge set between the agents in $\mathcal{V}_{\mathrm{s}}$.

Remark 1 Assumption 1 must hold if there is only one misbehaving agent, and it holds with high probability when misbehaving agents are very sparsely distributed in the network (He et al., 2013). Additionally, it is common to make such an assumption on network connectivity. Note that MSR-based methods assume that the network is $(2F+1)$-robust for the F-local malicious model and $(F+1, F+1)$-robust for the F-total malicious model (Kieckhafer and Azadmanesh, 1994; LeBlanc et al., 2013). These assumptions present the requirement for network connectivity among normal agents. Assumption 1 indicates the requirement for connectivity among misbehaving agents. Nevertheless, we make further discussion on relaxing Assumption 1 in Section 5, with detection and compensation solutions for typical cases.

Remark 2 Assumption 2 specifies the attacker capabilities. Malicious agents are usually reluctant to add any entries since they will be detected easily if normal agents cross-check the data in the two-hop information set. By Assumption 3, misbehaving agents will be isolated effectively by all normal agents, which can be achieved especially when there are mobile agents in the network (Zhao et al., 2018).

As for the update of the weight matrix after isolation, when an agent decides to cut off communication with another agent, it will adjust its weight by transferring the weight of the isolated agent to itself. As for the misbehaving agent, it needs to do a similar adjustment because it will no longer receive the corresponding information set. In this way, the weight matrix W will still be doubly stochastic. Any method of recalculation of the degrees to ensure that $W_{\{i\}}$ remains a doubly stochastic matrix is feasible. Note that the abovementioned strategy requires only local adjustment of the weight matrix and the information set would constitute a source of referral for neighbors for ascertaining the recalculation of the degrees. Assumption 4 is common in resilient-consensus-related literature and can always be satisfied if the number of malicious agents is below the network robustness threshold (Kieckhafer and Azadmanesh, 1994).

Deterministic scenario: Consider that misbehaving agents will cause adverse impacts while normal agents will add compensation values as a confrontation. The state update rule in Eq. (1) can be reformulated as

$$x(k+1) = Wx(k) + \varepsilon(k), \tag{2}$$

where $\varepsilon(k) = [\varepsilon_1(k), \varepsilon_2(k), \ldots, \varepsilon_N(k)]^{\mathrm{T}}$ is the input vector. The term $\varepsilon_i(k)$ is the error input for misbehaving agent $i \in \mathcal{V}_{\mathrm{m}}$, while $\varepsilon_j(k)$ is the compensation input for normal agent $j \in \mathcal{V}_{\mathrm{s}}$.

Stochastic scenario: Given the fact that the possibility for link failure during communication is considered in the present study, we adopt the understanding of link failures as the phenomenon with the potential for preventing the information set from being received at each needed time slot. p denotes the probability of connection between agents; i.e., link failure occurs with probability $1-p$ between each pair of agents independently. Under scenarios where communication link failures may occur, the matrix W will be time-dependent, i.e., $W(k)$. Moreover, due to link failures, some information sets are not available to normal agents, which will lead to undetected errors. Without loss of generality, the errors caused by misbehaving agents could be characterized as obeying an unknown distribution (Marano et al., 2009). Misbehaving agent i affects the system (or the error equals zero) with probability $\theta_i \in [0,1]$. We denote by $X_i(k)$ the random indicator for misbehavior, i.e., $X_i(k) \sim \mathcal{B}(1,\theta_i)$, where $\mathcal{B}$ is the Bernoulli distribution. Additionally, when misbehaving agents affect the system, the error inputs $Y_i(k)$ $(i \in \mathcal{V}_{\mathrm{m}})$ are added. We reasonably assume that $Y_i(k)$ follows a specific distribution with mean μ_i and variance σ_i^2. We consider that the probability of $Y_i(k) = 0$ is 0 since misbehaving agents aim to affect the system, and $X_i(k)$ and $Y_i(k)$ are mutually independent. Hence, we have $\varepsilon_i(k) = X_i(k)Y_i(k)$ for misbehaving agents in $\mathcal{V}_{\mathrm{m}}$. Then, it holds that

$$\begin{aligned}\mathbb{E}[\varepsilon_i(k)] &= \theta_i\mu_i,\\ \mathbb{D}[\varepsilon_i(k)] &= \theta_i\sigma_i^2 + (1-\theta_i)\theta_i\mu_i^2 \triangleq \sigma_{\varepsilon_i}^2.\end{aligned}$$

Nevertheless, $\varepsilon_i(k)$ could be zero and it can obey an arbitrary distribution because we do not pose any restriction on the distributions of $X_i(k)$ and $Y_i(k)$, and do not need to know their expectation and variance, which poses a situation different from the one characterizing the faults. The difference between malicious agents and faulty agents is that malicious agents will attack the system continuously, while faulty agents will cause only accidental disturbances in a limited period.

2.4 Problem formulation

In this paper, we consider a resilient average consensus problem in a multi-agent system with misbehaving nodes, which include faulty and malicious ones. Each agent has an initial state that can be expressed as $x_i(0), i \in \mathcal{V}$, and the system updates the states according to Eq. (2). The goal of this work is to design a detection and compensation method to realize resilient average consensus in a distributed manner. Besides, given that the communication in distributed multi-agent systems could be unreliable due to link failures (especially in the wireless communication scenario), the consensus process will be influenced if the information set is intermittently unavailable. Therefore, we further consider link failures and aim to solve the following two issues:

1. Resilient accurate average consensus: Considering deterministic scenarios where the communication links are reliable, we aim to develop a misbehavior-resilient algorithm to achieve resilient accurate average consensus among the agents in the set after isolation $\mathcal{V}_{\mathrm{r}}$ for the system under

misbehaviors, namely

$$\lim_{k\to\infty} x_i(k) = \frac{1}{|\mathcal{V}_\mathrm{r}|}\sum_{l\in\mathcal{V}_\mathrm{r}} x_l(0), \quad \forall i \in \mathcal{V}_\mathrm{r}, \tag{3}$$

where the term $\mathcal{V}_\mathrm{r}$ is a subset of $\mathcal{V}$ containing the agents that are not isolated.

2. Resilient unbiased average consensus in a statistical sense: Considering stochastic scenarios where the communication links between agents are interrupted randomly, we aim to improve our detection and compensation algorithm to realize resilient unbiased average consensus in a statistical sense, i.e.,

$$\mathbb{E}\left[\lim_{l\to\infty} x_i(l)\right] = \frac{1}{|\mathcal{V}_\mathrm{r}|}\sum_{l\in\mathcal{V}_\mathrm{r}} x_l(0), \quad \forall i \in \mathcal{V}_\mathrm{r}. \tag{4}$$

3 Deterministic detection compensation based consensus

In this section, we propose a detection algorithm to detect misbehaving agents and then compensate for the negative impact caused by these misbehaviors. The underlying design idea is to extract abnormal behaviors from the redundant information in the two-hop information set, so as to detect and compensate for the adverse impact. Using the abovementioned detection and compensation methods, we present the D-DCC algorithm. The design details and performance analysis are provided in the following.

3.1 Detection strategies of D-DCC

The first step for each normal agent is to determine whether there are misbehaving agents in the neighborhood and to estimate the amount of error injected by them. Based on the two-hop information sets, we propose the following two detection strategies. Without loss of generality, we conduct an analysis in a subsystem composed of misbehaving agent i and its neighbors $j \in \mathcal{N}_i$ in the following.

Detection strategy I: The normal agent $j \in \mathcal{N}_i$ will detect whether misbehaving agent i modifies the states of agent j in the information set, which amounts to checking whether $x_j^{(i)}(k) = x_j(k)$ holds. Note that if a malicious agent i removes the ID and state of agent j, it can be considered that the corresponding state is changed to zero.

Detection strategy II: The normal agent j will perform the detection by checking whether the update rule in Eq. (1) is followed for each neighboring agent $i, i \in \mathcal{N}_j$. Specifically, it checks whether $x_i(k+1) = \sum_{h\in\mathcal{N}_i} w_{ih} x_h^{(i)}(k)$ holds.

Concerning the misbehaviors mentioned in Section 2.3, these could be detected with the use of detection strategy I or II or both (see Lemma 2 in Section 3.3). For misbehaving agent i, it holds that

$$\begin{aligned} x_i(k+1) &= \sum_{j\in\mathcal{N}_i} w_{ij}x_j(k) + \varepsilon_i(k) \\ &= \sum_{j\in\mathcal{N}_i} w_{ij}x_j(k) + \sum_{j\in\mathcal{N}_i} \varepsilon_i^{j(1)}(k) + \varepsilon_i^{(2)}(k), \end{aligned} \tag{5}$$

where the terms $\varepsilon_i^{j(1)}(k)$ and $\varepsilon_i^{(2)}(k)$ are the adverse impacts detected by detection strategies I and II, respectively, and the specific representations are given as follows:

$$\varepsilon_i^{j(1)}(k) = w_{ij}(x_j^{(i)}(k) - x_j(k)), \tag{6a}$$

$$\varepsilon_i^{(2)}(k) = x_i(k+1) - \sum_{j\in\mathcal{N}_i} w_{ij}x_j^{(i)}(k). \tag{6b}$$

Note that each neighbor of misbehaving agent i will detect agent i with the same error by detection strategy II, since misbehaving agents will send the same information to all neighbors. Hence, $\varepsilon_i^{(2)}(k)$ is used, instead of $\varepsilon_i^{j(2)}(k)$.

Remark 3 For the above two detection strategies, all information needed for agent j to detect agent i can be obtained from its own state $x_j(k)$ and the available two-hop information sets $\Psi_i(k+1)$ and $\Psi_i(k)$ from agent i. Note that for all normal neighbors of agent i, both detection strategies I and II will be executed. In addition, by detection strategy I, each neighbor agent j can check only its own state $x_j^{(i)}(k)$ in the information set.

3.2 Compensation schemes of D-DCC

To achieve resilient average consensus, the following lemma is given which provides a sufficient condition for consensus on dynamical systems:

Lemma 1 (He et al., 2019) Considering system (2), the average consensus can be exponentially achieved if the input vectors are bounded; namely, we have $||\varepsilon(k)||_\infty \le \alpha\rho^k$ for certain parameters $\alpha > 0$ and $\rho \in [0,1)$ and the sum of the input vectors satisfies

$$\sum_{k=0}^{\infty}\sum_{i=1}^{N} \varepsilon_i(k) = 0. \tag{7}$$

Obviously, the extra data injected by misbehaving agents will make Eq. (7) unsatisfied; i.e., the average consensus is affected. From Lemma 1, to achieve resilient accurate average consensus, we need to design and add compensation inputs $\varepsilon_j(k)$ for all normal agents $j \in \mathcal{V}_{\rm s}$ to compensate for the impact of misbehaviors. In this way, Eq. (7) can still be satisfied under misbehaviors. Since the compensation cannot be completed at once, we introduce an error compensator η_j for each normal agent j to store the amount of error that needs to be compensated for. At each time slot, every normal agent j will determine the compensation input $\varepsilon_j(k)$ from the compensator η_j. Actually, the errors that need to be compensated for come from three aspects: the errors identified by detection strategies I and II, and the error introduced by isolation if the misbehaving agent injects malicious data beyond the tolerance for consensus. To cope with different situations, we propose three compensation schemes as follows:

1. Compensation scheme I: When the misbehavior of agent i in modifying $x_j^{(i)}(k)$ is captured by detection strategy I, compensation scheme I will be employed by the normal agent $j \in \mathcal{N}_i$ to compensate for the impact on consensus. The corresponding error to be compensated for is determined by

$$\eta_j^{i(1)}(k+1) = -w_{ij}(x_j^{(i)}(k) - x_j(k)). \tag{8}$$

2. Compensation scheme II: When the misbehaving agent i does not follow the update rule in Eq. (1), the normal neighboring agent $j \in \mathcal{N}_i$ can detect it by detection strategy II. Correspondingly, the error that needs to be compensated for is $\varepsilon_i^{(2)}(k)$. Since all $|\mathcal{N}_i|$ adjacent agents of agent i will detect the same error, each neighboring agent $j \in \mathcal{N}_i$ will compensate for the detected error averagely. Therefore, the compensation value of each normal agent $j \in \mathcal{N}_i$ is given by

$$\eta_j^{i(2)}(k+1) = -\varepsilon_i^{(2)}(k)/|\mathcal{N}_i|. \tag{9}$$

Isolation scheme: Considering that misbehaving agents cause finite errors such as accidental computation errors and actuator errors, compensation schemes I and II can accurately compensate for them. However, the malicious agent i could be too aggressive such that the adverse impact caused by it is severe enough to breach the conditions such as $||\varepsilon(k)||_\infty \leq \alpha\rho^k$. As a result, the consensus process will be seriously affected and the convergence may not be achieved according to Lemma 1. In this case, the misbehaving agent i should be isolated (i.e., the normal neighboring agents $j \in \mathcal{N}_i$ will cut off the communication with agent i) to avoid a future adverse impact from agent i. From Lemma 1, we design a distributed exponentially decaying error bound (i.e., $\alpha_j\rho_j^k$, $j \in \mathcal{V}_{\rm s}$) as the criterion for isolation. Let $\alpha = N\max_{i\in\mathcal{V}}\alpha_i$ and $\rho = \max_{i\in\mathcal{V}}\rho_i$. Then, we have $||\varepsilon(k)||_\infty \leq \alpha\rho^k$. Specifically, the normal agent $j \in \mathcal{N}_i$ can estimate the error of its neighbor i by Eq. (6). If the obtained error is within the bound, i.e., if $|\varepsilon_i^{j(1)}(k) + \varepsilon_i^{(2)}(k)| \leq \alpha_j\rho_j^k$, then agent j can compensate for the error according to compensation schemes I and II. Otherwise, agent i will be isolated and the following compensation scheme III will be employed to remedy the historical bad impact from the misbehaving agent i. In this way, an accurate average consensus with the remaining agents (after isolation) can be guaranteed.

3. Compensation scheme III: The main idea of compensation scheme III is to ensure that the summation of the remaining agents' states after isolation is the same as that of the initial states of the remaining agents. In other words, the historical adverse impact on consensus that is attributable to misbehaving agent i should be removed. Therefore, each normal agent $j \in \mathcal{N}_i$ will equally compensate for the historical adverse impact. The specific compensation value for isolation is designed as

$$\eta_j^{i(3)}(k+1) = \frac{1}{|\mathcal{N}_i|}\big(x_i(k+1) - x_i(0)\big). \tag{10}$$

Remark 4 The detailed D-DCC algorithm steps are summarized in the supplementary materials. Note that both normal and misbehaving agents have the input term $\varepsilon_i(k)$ (i.e., the error or compensation input) in their information sets. If malicious agents have full knowledge of detection and compensation methods, then they can easily masquerade as normal ones. To avoid this issue, the attack detection indicator $\pi_i(k)$ in the information set $\Psi_i(k)$ could help. A normal agent is allowed to add nonzero compensation input only when it has detected misbehaviors in the neighborhood. Then, we adopt a steady compensation sequence that restricts the changes of compensation, i.e., $|\varepsilon_i(k) - \varepsilon_i(k-1)| \leq \delta$, which is reasonable in practice. We also reasonably assume that malicious agents cannot change the attack detection indicator or have no knowledge of δ.

With the two abovementioned methods, it will be easy to distinguish malicious agents with errors and normal agents with compensation input.

3.3 Performance analysis of D-DCC

The following lemma is given first which demonstrates the effectiveness of D-DCC:

Lemma 2 If Assumptions 1–4 hold, then all misbehaviors mentioned in Section 2.3 will be detected by detection strategies I and II, and some malicious agents will be isolated.

The proof is omitted. Please see Zheng et al. (2021). For ease of understanding, we provide an example in the supplementary materials to illustrate the effectiveness of the proposed detection method.

Next, the consensus performance of D-DCC is analyzed in Theorem 1.

Theorem 1 If Assumptions 1–4 hold, then D-DCC achieves resilient average consensus among the agents in the set after isolation $\mathcal{V}_{\mathrm{r}}$, i.e., Eq. (3) holds, where

$$\mathcal{V}_{\mathrm{r}}=\left\{i\in\mathcal{V}\mid|\varepsilon_i^{j(1)}(k)+\varepsilon_i^{(2)}(k)|<\alpha_j\rho_j^k,\forall j\in\mathcal{N}_i\right\}.$$

The proof is omitted. See Zheng et al. (2021).

Remark 5 The error bound can be used to a certain extent to distinguish between ever-present attacks and accidental faults and to provide tolerance for faults. The parameters α_i and ρ_i are designed to cover the fault, which depends mainly on the prior information of faults in the practical system. Note that a larger $\alpha_i\rho_i$ can increase fault tolerance but slow down convergence. Though the fault tolerance bound may mistakenly classify malicious agents as faulty agents if malicious agents inject false data within the bound, the attack is also restricted to zero as time goes to infinity, which hardly harms the consensus process.

4 Stochastic detection compensation based consensus

4.1 Algorithm design

In this subsection, we further investigate the resilient average consensus problem against misbehaviors while considering the possible link failures among agents. We assume that the link failure occurs between any two agents with the same probability p at each time slot (see Section 2.3). This stochasticity property brings extra challenges compared with the problems in the deterministic scenario (see Section 3); i.e., the corresponding information set of neighbors is not available when a link failure occurs. As a result, not only may state updating be affected, but also some misbehaviors among relevant agents will not be detected. First, to handle the effect on state updating, agent i will adjust its update weight by transferring the weight of agent j to itself at time k if link failures between agents i and j occur, i.e., $w(k)_{ij}=0, w(k)_{ii}=w_{ii}+w_{ij}$. Similarly, agent j will perform the corresponding weight adjustment. In this way, $W(k)$ will be doubly stochastic and state updating would still function normally. Second, if link failures occur between misbehaving agent i and normal agent j, agent j is not able to detect the errors of agent i. Such a situation would cause undetected errors, necessitating further compensation. To ensure the resilience performance against misbehaving agents when the information set is randomly unavailable, we propose an S-DCC algorithm to achieve unbiased resilient average consensus in a statistical sense. The detailed S-DCC algorithm is given in the supplementary materials. Specifically, we further propose compensation scheme IV based on the estimation of the average detected errors:

Compensation scheme IV: Considering the compensation for the adverse impact of undiscovered misbehaviors due to unreliable communication links, the compensation value is designed as

$$\eta_j^{i(4)}(k+1)=-\overline{\varepsilon}_i^j(k)\big(k-k_i^{j0}-m_j(k)\big),\tag{11}$$

where k_i^{j0} is the last detection time before agent i is first detected by agent j as a misbehaving agent, which is treated as the last time before a misbehavior occurs. Additionally, we have

$$\varepsilon_i^j(k)=\varepsilon_i^{j(1)}(k)+\varepsilon_i^{(2)}(k)/|\mathcal{N}_i|,\tag{12a}$$

$$\overline{\varepsilon}_i^j(k)=\sum_{\varepsilon_i^j(k)\in\Omega_j^{(i)}(k)}\varepsilon_i^j(k)/m_j(k),\tag{12b}$$

where $m_j(k)$ is the number of times that agent j detects agent i after time k_i^{j0}. The insight behind compensation scheme IV is that the average value of detected errors could characterize the average impact of misbehaving agents over a period of time. Specifically, the information set with probability p is

available at each occasion of information transmission, with which the normal agent j can perform the detection and compensation for potential errors. To estimate the average impact of misbehaving agents, agent j will store the detected error $\varepsilon_i^j(k)$ in the set $\Omega_j^{(i)}(k)$ for agent i, and then use the average detected error to compensate for undetected errors.

4.2 Performance analysis

Before analysis of the S-DCC algorithm, we define several notations. Summing up the compensation input of all neighbors of misbehaving agent i, the compensation of $\varepsilon_i(k)$ is given by $\overline{\varepsilon}_i(k) \triangleq \sum_{j\in\mathcal{N}_i}\overline{\varepsilon}_i^j(k)$, where $\overline{\varepsilon}_i^j(k)$ is the average of the errors of misbehaving agent i detected by agent $j\in\mathcal{N}_i$. Note that there could be faulty agents non-isolated when they appear accidentally and their errors are within the error bounds. Therefore, we assume that the misbehaviors of the faulty agent i occur in a period from k_i^0 to k_i^1, which is reasonable. We denote by k_i^{iso} the isolation time of agent i, and subsets $\mathcal{V}_{\mathrm{f}}$ and $\mathcal{V}_{\mathrm{M}}$ are the sets of faulty and malicious agents, respectively. The following theorem is provided to demonstrate the performance of S-DCC:

Theorem 2 If Assumptions 1–4 are satisfied, then S-DCC can achieve unbiased resilient average consensus in a statistical sense among the agents in the set after isolation $\mathcal{V}_{\mathrm{r}}$; namely, $\forall j\in\mathcal{V}_{\mathrm{r}}$, we have

$$\mathbb{E}\big[\lim_{l\to\infty}x_j(l)\big]=\frac{1}{|\mathcal{V}_{\mathrm{r}}|}\sum_{u\in\mathcal{V}_{\mathrm{r}}}x_u(0). \tag{13}$$

In addition, the consensus value is bounded, i.e.,

$$\left|\lim_{l\to\infty}x_j(l)-\frac{1}{|\mathcal{V}_{\mathrm{r}}|}\sum_{u\in\mathcal{V}_{\mathrm{r}}}x_u(0)\right|\le\frac{\alpha\rho|\mathcal{V}_{\mathrm{m}}|}{(1-\rho)|\mathcal{V}_{\mathrm{r}}|}. \tag{14}$$

Proof First, we illustrate that all misbehaving agents will be detected. For each malicious agent i, the system is affected with a probability of θ_i, where $0<\theta_i\le 1$. For each normal agent $j\in\mathcal{N}_i$, it detects the misbehavior of agent i with probability $p, 0<p\le 1$. The probability of the event that agent i is detected by agent j in no later than time k is

$$P(k)=1-(1-p\theta_i)^k. \tag{15}$$

By taking the limit on both sides of Eq. (15), we have $\lim_{k\to\infty}P(k)=\lim_{k\to\infty}(1-(1-p\theta_i)^k)=1$. Hence, all misbehaving agents will be detected as time goes to infinity. Similarly, all malicious agents will be isolated with a probability of 1 because the bound $\alpha\rho^k=0$ as $k\to\infty$.

Second, we prove that average consensus is achieved. Since $\mathcal{G}_{\mathrm{s}}$ is connected and the remaining misbehaving agents must have normal neighbors, $\mathcal{G}_{\mathrm{r}}$ will be connected, where $\mathcal{G}_{\mathrm{r}}=(\mathcal{V}_{\mathrm{s}},\mathcal{E}_{\mathrm{s}})$ and $\mathcal{E}_{\mathrm{r}}$ ($\mathcal{E}_{\mathrm{r}}\subseteq\mathcal{E}$) denotes the edge set between the agents in $\mathcal{V}_{\mathrm{r}}$. Due to the detection and isolation, we have $||\varepsilon(k)||_\infty\le\alpha\rho^k$. When $W(k)$ is time-dependent, it holds that $\mathbb{E}[w(k)_{ij}]=pw_{ij}$ if $i\ne j$, and $\mathbb{E}[w(k)_{ii}]=w_{ii}+(1-p)\sum_{j\in\mathcal{N}_i}w_{ij}$. Hence, $\mathbb{E}[W(k)]$ is doubly stochastic (since it is the expected value of matrices, each matrix satisfies this property). Then, we have

$$\mathbb{E}\Big[\prod_{k=1}^{\infty}W(k)\Big]=\prod_{k=1}^{\infty}\mathbb{E}\big[W(k)\big]=11^{\mathrm{T}}/N. \tag{16}$$

Next, we will prove that

$$\mathbb{E}\big[\lim_{l\to\infty}\sum_{j\in\mathcal{V}_{\mathrm{r}}}x_j(l)\big]=\sum_{j\in\mathcal{V}_{\mathrm{r}}}x_j(0). \tag{17}$$

First, we have

$$\begin{aligned}&\mathbb{E}\big[\sum_{j\in\mathcal{V}}x_j(k)\big]\\&=1^{\mathrm{T}}\prod_{l=0}^{k-1}\mathbb{E}\big[W(l)\big]x(0)+1^{\mathrm{T}}\sum_{l=0}^{k-1}\prod_{t=l+1}^{k-1}\mathbb{E}\big[W(t)\big]\varepsilon(l)\\&=\mathbb{E}\big[\sum_{j\in\mathcal{V}}\sum_{l=0}^{k-1}\varepsilon_j(l)\big]+\sum_{j\in\mathcal{V}}x_j(0).\end{aligned}$$

For simplicity, we perform analysis on a subsystem composed of the misbehaving agent i and its neighbors in $\mathcal{N}_i$ here. Recall that we set $\varepsilon_i(l)=X_i(l)Y_i(l)$ in Section 2.3, where $X_i(l)$ and $Y_i(l)$ are independent.

Case 1: Considering a malicious agent i, the expectation of the sum of its error within time k is

$$\mathbb{E}\left[\sum_{l=1}^{k}\varepsilon_i(l)\right]=\mathbb{E}\left[\sum_{l=1}^{k}X_i(l)Y_i(l)\right]=k\theta_i\mu_i. \tag{18}$$

Without loss of generality, we consider that all neighbor agents in $\mathcal{N}_i$ detect the misbehavior of agent i at the same time. The expectation of $\sum_{j\in\mathcal{N}_i}\overline{\varepsilon}_i^j(k)$ satisfies

$$\begin{aligned}\mathbb{E}\Big[\sum_{j\in\mathcal{N}_i}\overline{\varepsilon}_i^j(k)\Big]&=\mathbb{E}\Big[\frac{1}{m_j}\sum_{j\in\mathcal{N}_i}\sum_{\varepsilon_i^j(k)\in\Omega_j^{(i)}}\varepsilon_i^j(k)\Big]\\&=\mathbb{E}\big[\varepsilon_i(l)\big]=\theta_i\mu_i.\end{aligned}$$

The expectation of the sum of compensation values from compensation schemes I, II, and IV is

$$\begin{aligned}&\mathbb{E}\Big[\sum_{j\in\mathcal{N}_i}\sum_{l=1}^{k}(\eta_j^{i(1)}(l)+\eta_j^{i(2)}(l)+\eta_j^{i(4)}(l))\Big]\\=&\mathbb{E}\Big[\sum_{j\in\mathcal{N}_i}\sum_{\varepsilon_i^j(k)\in\Omega_j^{(i)}}\varepsilon_i^j(k)-(k-m_j(k))\sum_{j\in\mathcal{N}_i}\overline{\varepsilon}_i^j(k)\Big]\\=&\mathbb{E}\Big[-k\sum_{j\in\mathcal{N}_i}\overline{\varepsilon}_i^j(k)\Big]=-k\theta_i\mu_i.\end{aligned}\tag{19}$$

Let $k+1=k_i^{\mathrm{iso}}$ for Eq. (10). Therefore, combining Eqs. (10), (18), and (19), we have

$$\mathbb{E}\left[\sum_{l=1}^{k}(\varepsilon_i(l)+\sum_{j\in\mathcal{N}_i}\varepsilon_j(l))\right]=x_i(k+1)-x_i(0).$$

Hence, we have

$$\mathbb{E}\left[\lim_{l\to\infty}\sum_{j\in\mathcal{V}/\{i\}}x_j(l)\right]=\sum_{j\in\mathcal{V}/\{i\}}x_j(0).$$

Case 2: Consider that the errors of faulty agent i occur in a period from k_i^0 to k_i^1. The compensation for agent i is $-\sum_{j\in\mathcal{N}_i}(k_i^{j0}-k_i^{j1})\overline{\varepsilon}_i^j(k)$, where k_i^{j0} is the last time of detection before k_i^0 and k_i^{j1} is the last time of detection before k_i^1. Consequently, it holds that

$$\mathbb{E}[k_i^0-k_i^{j0}]=\mathbb{E}[k_i^1-k_i^{j1}]=\frac{1}{p}-1.$$

Hence, we have $\mathbb{E}[k_i^{j0}-k_i^{j1}]=k_i^0-k_i^1$. Since detection and errors are independent, it holds that

$$\mathbb{E}\left[-\sum_{j\in\mathcal{N}_i}(k_i^{j0}-k_i^{j1})\overline{\varepsilon}_i^j(k)\right]=(k_i^1-k_i^0)\theta_i\mu_i.$$

Then, we have

$$\mathbb{E}\left[\lim_{l\to\infty}\sum_{j\in\mathcal{V}}x_j(l)\right]=\sum_{j\in\mathcal{V}}x_j(0).$$

With the two abovementioned cases, we have Eq. (17) for the general set $\mathcal{V}_{\mathrm{r}}$. Hence, Eq. (4) holds and the unbiased resilient average consensus in a statistical sense among the agents in $\mathcal{V}_{\mathrm{r}}$ is achieved.

According to the proof of Theorem 1 (see Zheng et al. (2021)), $\varepsilon(k)$ satisfies the condition $||\varepsilon(k)||_\infty\leq\alpha\rho^k$. Then, for misbehaving agents, we have

$$\sum_{i\in\mathcal{V}_{\mathrm{m}}}\sum_{l=1}^{\infty}\varepsilon_i(l)\leq\frac{|\mathcal{V}_{\mathrm{m}}|\alpha\rho}{(1-\rho)}.$$

Hence, inequality (14) is proved.

Remark 6 Theorem 2 ensures the unbiased resilient average consensus in a statistical sense. Meanwhile, the expectation of undetected errors is the same as the mean of detected errors. For faulty agents, the compensation period $k_i^{j1}-k_i^{j0}$ has the same expectation as that of errors, i.e., $k_i^1-k_i^0$. Hence, misbehaviors can be unbiasedly compensated for in a statistical sense. For normal agents, the larger attack probability θ_i of neighboring misbehaving agents and detection probability p will reduce the number of expected steps for detection. On one hand, a larger attack probability will improve the attack capability of malicious agents. On the other hand, the detection probability based on the reliable link will be close to 1. Hence, the detection performance will be improved. Furthermore, it is not actually necessary for malicious agents to attack with a constant probability and a certain distribution. The detection method will be effective as long as the attack probability is larger than 0, and the errors may follow a certain attack method. Hence, to simplify the statement, we assume a constant attack probability and present the attack errors by a time-invariant probabilistic model.

Next, we analyze the accuracy of the mean-based compensation scheme IV, i.e., the distance between the mean-based compensation and actual errors. The actual errors may consist of multiple uncertainties. Let $F_{Y_i(k)}(x)$ be the cumulative distribution function (CDF) of $Y_i(k)$. We adopt the Gaussian mixture model (GMM) to represent the error variable $Y_i(k)$, because any distribution can be generally modeled by GMM with arbitrary precision. GMM is defined as a convex combination of N_l Gaussian distributions with expectations μ_l and variances σ_l for the integer $1\leq l\leq N_l$, namely,

$$F_{Y_i(k)}(x)=\sum_{l=1}^{N_l}a_l\Phi(\frac{x-\mu_l}{\sigma_l}),\sum_{l=1}^{N_l}a_l=1,\sum_{l=1}^{N_l}a_l\mu_l=\mu_i,$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution. Since $\varepsilon_i(k)=X_i(k)Y_i(k)$, the CDF of $\varepsilon_i(k)$ is

$$F_{\varepsilon_i(k)}(x)=\begin{cases}\theta_iF_{Y_i(k)}(x), & x<0,\\1-\theta_i+\theta_iF_{Y_i(k)}(x), & x\geq0.\end{cases}\tag{20}$$

Without loss of generality, we consider that all the detection numbers $m_j(k)$ $(j\in\mathcal{N}_i)$ are the same.

When the detection number is large enough, the following holds according to the central limit theorem (Grimmett and Stirzaker, 2020):

$$\overline{\varepsilon}_i(k) \sim \mathcal{N}(\theta_i\mu_i, \sigma^2_{\varepsilon_i}/M_i),$$

where $\mathcal{N}(\cdot,\cdot)$ is the Gaussian distribution and $M_i = \min_{j\in\mathcal{N}_i} m_j(k_i^{\text{iso}})$. The CDF of $\overline{\varepsilon}_i$ is given by

$$F_{\overline{\varepsilon}_i}(x) = \Phi\left(\frac{\sqrt{M_i}(x-\theta_i\mu_i)}{\sigma_{\varepsilon_i}}\right).$$

The close proximity between the statistical distribution of the error and that of the compensation value will guarantee not only the close final consensus value but also the stationarity of the consensus process. The characteristic of proximity of two probability distributions can be described by the Wasserstein distance (Vallender, 1974). The Wasserstein distance $R(\mathcal{P},\mathcal{Q})$ between the two distributions $\mathcal{P}$ and $\mathcal{Q}$ is defined as follows:

$$R(\mathcal{P},\mathcal{Q}) = \inf \mathbb{E}\big[d(\xi,\eta)\big],$$

where $d(\cdot,\cdot)$ is the function of the metric space and the mathematical operation inf is taken over all possible pairs of random variables ξ and η with distributions $\mathcal{P}$ and $\mathcal{Q}$, respectively. In the case of one-dimensional space with the Euclidean metric, the Wasserstein distance is calculated by

$$R(\mathcal{P},\mathcal{Q}) = \int_{-\infty}^{\infty} |F(x) - G(x)| \mathrm{d}x,$$

where $F(x)$ and $Q(x)$ are the CDFs of $\mathcal{P}$ and $\mathcal{Q}$, respectively (Vallender, 1974). Hence, we have the Wasserstein distance between $\varepsilon_i(k)$ and $\overline{\varepsilon}_i(k)$ as

$$R(\varepsilon_i(k),\overline{\varepsilon}_i(k)) = \int_{-\infty}^{\infty} |F_{\varepsilon_i(k)}(x) - F_{\overline{\varepsilon}_i}(x)| \mathrm{d}x. \quad (21)$$

We can use the Wasserstein distance to show the expectation of the absolute error between the mean-based compensation and actual errors. The following theorem is provided to illustrate the bound of $R(\varepsilon_i(k),\overline{\varepsilon}_i(k))$:

Theorem 3 When $Y_i(k)$ is modeled by GMM, we have

$$R(\varepsilon_i(k),\overline{\varepsilon}_i(k)) \le (1-\theta_i)\mathbb{E}[|Y_i|] + \sum_{l=1}^{N_l} a_l \left(|\theta_i\mu_i - \mu_l| + |\frac{\sigma_{\varepsilon_i}}{\sqrt{M_i}-\sigma_l}|\right), \quad (22)$$

where $\mathbb{E}[|Y_i|] \le \sum_{l=1}^{N_l}\left\{\sqrt{\frac{2}{\pi}}\sigma_l \exp\left(\frac{-\mu_l^2}{2\sigma_l^2}\right) + \mu_l[1 - 2\Phi(\frac{-\mu_l}{\sigma_l})]\right\}$.

Proof See the supplementary materials.

5 Discussion on assumption relaxation

In this paper, Assumption 1 makes a strong assumption about the connection relationship between agents; i.e., any two misbehaving agents are not adjacent to each other. This limits the interconnection between misbehaving agents. Although Assumption 1 is more likely to hold when misbehaving agents are sparsely distributed compared to normal agents, it still cannot be ruled out that some misbehaving agents could be adjacent to each other. In this regard, we discuss the relaxation of Assumption 1 in this section by analyzing a typical structure wherein two misbehaving agents are adjacent, followed by the corresponding detection and compensation strategies. Note that we do not fully solve the problem of misbehaving agent adjacency, which is worthy of further research.

A common assumption in related detection methods is that each pair of neighboring misbehaving agents must have at least one common normal neighbor (Zhao et al., 2018). Consider the situation in which two misbehaving agents have common normal neighbor(s). A representative topology is shown in Fig. 1. When two malicious agents (agents 2 and 3 in Fig. 1) are neighbors, they do not need to use detection strategies for each other. However, normal agent 1 may not be able to detect whether agent 2 or 3 is abnormal according to the previously proposed detection strategy. This is because as long as agent 2 forges the information of agent 3 according to the information set so that it satisfies the update rule in the eyes of agent 1, it can prevent agent 1

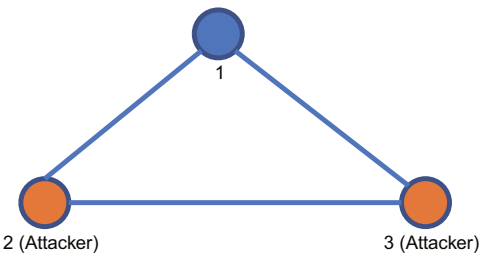

Fig. 1 Typical topology of neighboring misbehaving agents with common normal neighbors

from identifying agent 2 as a misbehaving agent according to detection strategy II. Agent 3 can perform similar operations to prevent agent 1 from detecting it through detection strategy II. In this case, previous detection strategies may not work well. In other words, we need a new detection strategy to deal with it, which is given as follows.

Detection strategy III: Consider a normal agent j and two misbehaving agents i and h, where the three agents are mutually adjacent to each other. To prevent agent j from detecting malicious agents i and h, agents i and h will collaborate to bypass detection strategy II by carefully modifying the information set. Owing to the fact that both malicious agents i and h want to bypass detection strategy II in agent j, the information set of agent i received by agent j may be different from the part about agent i in the information set that is sent by agent h to agent j. This renders the possibility of detection of agents i and h by agent j. In essence, agent j detects agent i using the common neighboring agent h. Specifically, at time k, the normal agent j saves the state values of all neighbor agents $(x_h(k), h \in \mathcal{N}_j)$. At time $k+1$, the normal agent j checks whether the states of each neighboring agent i's neighbors in the information set are correct, i.e., whether $x_h(k) = x_h^{(i)}(k)$ holds, where $i \in \mathcal{N}_j, h \in \mathcal{N}_j, h \in \mathcal{N}_i$. In this way, the detection can be achieved.

Compensation scheme V: After the detection, it is necessary to consider the means to compensate for the error. If the common neighbor h of agents i and j is normal, then agent h will detect misbehaving agent i at the same time, and will compensate for misbehaviors according to compensation scheme I. Hence, agent i does not need to add compensation. If agent h has been marked as a misbehaving agent, it is necessary to add compensation $-w_{ih}(x_h^{(i)}(k)-x_h(k))$ for the error of agent i. If agents i and h have multiple common normal neighbors (i.e., $|\mathcal{N}_i \cap \mathcal{N}_h \cap \mathcal{V}_\mathrm{s}|$) and all normal neighbors can detect and compensate for misbehaviors, then the compensation value of each neighbor will be given as follows:

$$\eta_j^{i(5)}(k+1) = -w_{ih}\big(x_h^{(i)}(k) - x_h(k)\big)/|\mathcal{N}_i \cap \mathcal{N}_h \cap \mathcal{V}_\mathrm{s}|. \tag{23}$$

Since the main idea of the compensation is the same as that of compensation scheme I, the theoretical proof is omitted.

6 Simulation results

In this section, a series of simulations are conducted to illustrate the effectiveness of the proposed D-DCC and S-DCC algorithms. We consider a multi-agent system with $N = 10$ nodes, where its network is described by an Erdös–Rényi random graph. Each edge is generated with a probability of 0.7. The system updates states by Eq. (2) with the weight matrix W designed by Perron weights. All agents' initial states are randomly selected from the interval $[0, 2]$. We set two misbehaving agents in the system, which are not adjacent. Specifically, agent 1 is malicious aiming to affect the consensus process spitefully and agent 5 is a faulty node. Here, the parameters are set as $\rho_i = 0.9$, $\alpha_i = 5$, $\forall i \in \mathcal{V}$.

6.1 Resilient consensus under D-DCC

In this subsection, for malicious agent 1 and faulty agent 5, the adverse impacts injected are given by $\varepsilon_1(k) = 0.5\cos k$ and $\varepsilon_5(k) = 0.5 \times 0.6^k$, respectively. Note that these misbehaviors start at time 0. Then, D-DCC is deployed and the state evolution and extra input of the system are shown in Fig. 2. From Fig. 2a, it can be seen that accurate average consensus is achieved with all agents except agent 1. This is because the error of agent 1 exceeds the predefined decaying error bound and agent 1 is isolated at time 24. The final convergence value (i.e., the consensus value) coincides with the average of the initial states of the agents that are not isolated (see the blue dotted line in Fig. 2a). For comparison, we deploy the MSR algorithm adopted in Kieckhafer and Azadmanesh (1994) and the resilient consensus algorithm used in Ramos et al. (2022), as shown in Fig. 2b. Obviously, both existing algorithms fail to achieve accurate average consensus. The MSR algorithm does not remove the injected adverse impact. The resilient consensus algorithm in Ramos et al. (2022) cannot detect agent 5 as misbehaving, because agent 5 injects only false data but does not aim at deviating the consensus to a specific value. Hence, its consensus process will be affected by agent 5, leading to a deviation from the average consensus. In Fig. 2c, we show the error inputs of agents 1 and 5. Note that agent 5 is not isolated since its error decays exponentially and is always within the error bound.

Note that the parameters α_j and ρ_j play a key

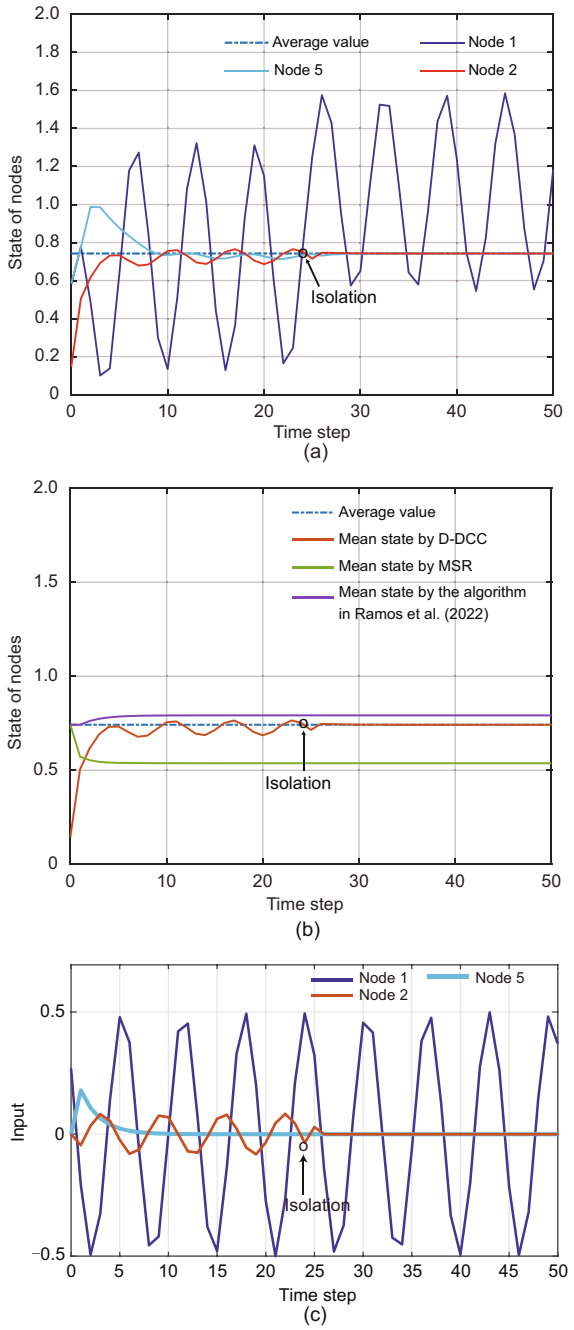

Fig. 2 Resilient consensus performance of D-DCC: (a) state evolution of the system; (b) system state under different methods; (c) error inputs of agents 1 and 5 and compensation of agent 2 (References to color refer to the online version of this figure)

role in trading off the fault tolerance and algorithm convergence. Here, we show the impact of different parameter selections on system performance. Obviously, if the values of parameters α_j and ρ_j are too small, the fault tolerance will be low. Hence, we only show the impact on the algorithm convergence with different parameters. Without loss of generality, let α_j be constant and ρ_j change. The results are given in Fig. 3. The mean state in the figure represents the average state. It can be seen that the mean state converges faster as ρ decreases.

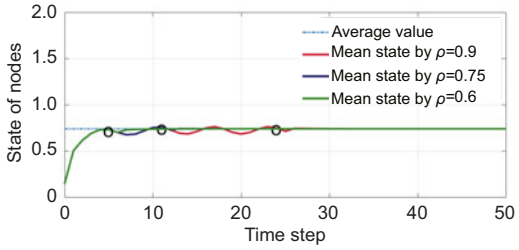

Fig. 3 Convergence of system states with varying ρ (References to color refer to the online version of this figure)

6.2 Resilient consensus under S-DCC

In this subsection, the error input of malicious agent 1 follows a GMM model when an attack is adopted, i.e., $F_{Y_1(k)}(x) = 0.5\Phi(\frac{x-0.05}{\sqrt{0.05}}) + 0.5\Phi(\frac{x-0.15}{\sqrt{0.2}})$. In addition, the error input of faulty agent 5 follows a normal distribution. Here, the attack probability and connection probability are set to be $\theta_1 = 0.8$ and $p = 0.8$, respectively. The performance of resilient consensus under S-DCC is provided in Fig. 4. Specifically, Fig. 4a demonstrates that the consensus is achieved with all agents except agent 1, since agent 1 causes errors continuously and exceeds the fault-tolerant bound, being isolated at time 28. Faulty agent 5 misbehaves only during the first 10 time slots and its error inputs are always within the fault-tolerant bound. Hence, agent 5 is not isolated and its errors are compensated for by neighbors. Here, the true average of the initial states of non-isolated agents is 0.672, while the similarly obtained number corresponding to the use of our methods is 0.693. We accordingly see that the latter is close to the true one. This is because the final consensus value varies randomly in practice, though the unbiased average consensus can be achieved theoretically. We also compare the S-DCC algorithm with the MSR algorithm in Kieckhafer and Azadmanesh (1994) and the consensus algorithm in Ramos et al. (2022). From Fig. 4b, we see that the consensus values of the MSR algorithm and the algorithm in Ramos et al. (2022) are 0.401 and 0.522, respectively. The consensus achieved by S-DCC is more

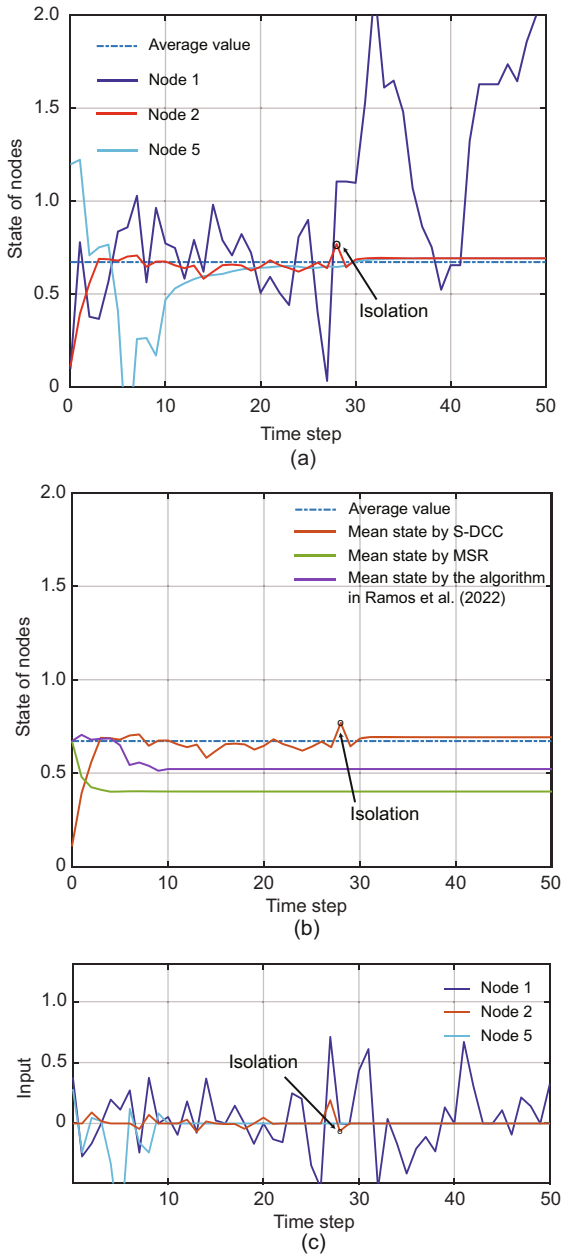

Fig. 4 Resilient consensus performance of S-DCC: (a) system state evolution; (b) system state under different methods; (c) error inputs of agents 1 and 5 and compensation of agent 2 (References to color refer to the online version of this figure)

accurate. In Fig. 4c, we show the error inputs of agents 1 and 5 and the compensation input of agent 2 under S-DCC.

7 Conclusions

In this work, we have studied the issue of resilient average consensus under misbehaving agents. First, considering scenarios with reliable communication, we have designed the D-DCC algorithm to eliminate the adverse impacts introduced by misbehaviors. We have proved that the resilient average consensus can be achieved by D-DCC. Furthermore, the S-DCC algorithm, as proposed in the present research, is imbued with the capability to adapt to scenarios wherein communication link failures may occur. It has been proved that the unbiased resilient average consensus in a statistical sense is achieved by S-DCC, and the absolute error between mean-based compensation and actual adverse impact has been analyzed using the Wasserstein distance. Finally, simulations have been conducted to illustrate the effectiveness of the proposed algorithms. In the future, it would be worthy to study the resilient consensus over time-varying and directed networks or high-dimensional systems. Resilient distributed optimization is also an interesting issue that can be investigated.

Contributors

Chongrong FANG, Wenzhe ZHENG, and Zhiyu HE designed the research. Chongrong FANG and Wenzhe ZHENG processed the data. Chongrong FANG, Wenzhe ZHENG, and Jianping HE drafted the paper. Chengcheng ZHAO and Jingpei WANG helped organize the paper. Jianping HE and Chengcheng ZHAO revised and finalized the paper.

Compliance with ethics guidelines

Chongrong FANG, Wenzhe ZHENG, Zhiyu HE, Jianping HE, Chengcheng ZHAO, and Jingpei WANG declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Dibaji SM, Ishii H, Tempo R, 2018. Resilient randomized quantized consensus. *IEEE Trans Autom Contr*, 63(8):2508-2522. https://doi.org/10.1109/TAC.2017.2771363

Ge XH, Han QL, Wu Q, et al., 2023. Resilient and safe platooning control of connected automated vehicles against intermittent denial-of-service attacks. *IEEE/CAA J Autom Sin*, 10(5):1234-1251. https://doi.org/10.1109/JAS.2022.105845

Gentz R, Wu SX, Wai HT, et al., 2016. Data injection attacks in randomized gossiping. *IEEE Trans Signal Inform Process Netw*, 2(4):523-538. https://doi.org/10.1109/TSIPN.2016.2614898

Grimmett G, Stirzaker D, 2020. Probability and Random Processes. Oxford University Press, Oxford, USA.

Hadjicostis CN, Domínguez-García AD, Vaidya NH, 2012. Resilient average consensus in the presence of heterogeneous packet dropping links. Proc 51st IEEE Conf on Decision Control, p.106-111. https://doi.org/10.1109/CDC.2012.6426666

He JP, Cheng P, Shi L, et al., 2013. SATS: secure average-consensus-based time synchronization in wireless sensor networks. *IEEE Trans Signal Process*, 61(24):6387-6400. https://doi.org/10.1109/TSP.2013.2286102

He JP, Cai L, Cheng P, et al., 2019. Distributed privacy-preserving data aggregation against dishonest nodes in network systems. *IEEE Int Things J*, 6(2):1462-1470. https://doi.org/10.1109/JIOT.2018.2834544

Kieckhafer RM, Azadmanesh MH, 1994. Reaching approximate agreement with mixed-mode faults. *IEEE Trans Parall Distrib Syst*, 5(1):53-63. https://doi.org/10.1109/71.262588

LeBlanc HJ, Zhang HT, Koutsoukos X, et al., 2013. Resilient asymptotic consensus in robust networks. *IEEE J Sel Areas Commun*, 31(4):766-781. https://doi.org/10.1109/JSAC.2013.130413

Ma RK, Zheng H, Wang JY, et al., 2022. Automatic protocol reverse engineering for industrial control systems with dynamic taint analysis. *Front Inform Technol Electron Eng*, 23(3):351-360. https://doi.org/10.1631/FITEE.2000709

Marano S, Matta V, Tong L, 2009. Distributed detection in the presence of Byzantine attacks. *IEEE Trans Signal Process*, 57(1):16-29. https://doi.org/10.1109/TSP.2008.2007335

Pasqualetti F, Bicchi A, Bullo F, 2012. Consensus computation in unreliable networks: a system theoretic approach. *IEEE Trans Autom Contr*, 57(1):90-104. https://doi.org/10.1109/TAC.2011.2158130

Ramos G, Silvestre D, Silvestre C, 2022. General resilient consensus algorithms. *Int J Contr*, 95(6):1482-1496. https://doi.org/10.1080/00207179.2020.1861331

Shames I, Teixeira AMH, Sandberg H, et al., 2011. Distributed fault detection for interconnected second-order systems. *Automatica*, 47(12):2757-2764. https://doi.org/10.1016/j.automatica.2011.09.011

Vallender SS, 1974. Calculation of the Wasserstein distance between probability distributions on the line. *Theory Probab Appl*, 18(4):784-786. https://doi.org/10.1137/1118101

Wang W, Huang JS, Wen CY, et al., 2014. Distributed adaptive control for consensus tracking with application to formation control of nonholonomic mobile robots. *Automatica*, 50(4):1254-1263. https://doi.org/10.1016/j.automatica.2014.02.028

Wen GH, Yu XU, Liu ZW, 2021. Recent progress on the study of distributed economic dispatch in smart grid: an overview. *Front Inform Technol Electron Eng*, 22(1):25-39. https://doi.org/10.1631/FITEE.2000205

Xiao L, Boyd S, Lall S, 2005. A scheme for robust distributed sensor fusion based on average consensus. Proc 4th Int Symp on Information Processing in Sensor Networks, p.63-70. https://doi.org/10.1109/IPSN.2005.1440896

Xie ML, Ding DR, Ge XH, et al., 2022. Distributed platooning control of automated vehicles subject to replay attacks based on proportional integral observers. *IEEE/CAA J Autom Sin*, early assess. https://doi.org/10.1109/JAS.2022.105941

Yang FS, Liang XH, Guan XH, 2021. Resilient distributed economic dispatch of a cyber-power system under DoS attack. *Front Inform Technol Electron Eng*, 22(1):40-50. https://doi.org/10.1631/FITEE.2000201

Yuan LW, Ishii H, 2021. Secure consensus with distributed detection via two-hop communication. *Automatica*, 131:109775. https://doi.org/10.1016/j.automatica.2021.109775

Zhao CC, He JP, Chen JM, 2018. Resilient consensus with mobile detectors against malicious attacks. *IEEE Trans Signal Inform Process Netw*, 4(1):60-69. https://doi.org/10.1109/TSIPN.2017.2742859

Zheng WZ, He ZY, He JP, et al., 2021. Accurate resilient average consensus via detection and compensation. Proc 60th IEEE Conf on Decision and Control, p.5502-5507. https://doi.org/10.1109/CDC45484.2021.9682843

List of supplementary materials

1 Algorithm 1 (D-DCC algorithm)
2 Illustrating example of the D-DCC algorithm
3 Algorithm 2 (S-DCC algorithm)
4 Proof of Theorem 3
Fig. S1 Example: a ring network with a single misbehaving agent 2