Title

Privacy preservation in continuous-time average consensus algorithm via deterministic additive obfuscation signals

Authors

Rezazadeh, Navid
Kia, Solmaz S

Publication Date

2019-04-10

Privacy preservation in continuous-time average consensus algorithm via deterministic additive perturbation signals

Navid Rezazadeh and Solmaz S. Kia

Abstract—**This paper considers the problem of privacy preservation against passive internal and external malicious agents in the continuous-time Laplacian average consensus algorithm over strongly connected and weight-balanced digraphs. For this problem, we evaluate the effectiveness of use of additive perturbation signals as a privacy preservation measure against malicious agents that know the graph topology. Our results include (a) identifying the necessary and sufficient conditions on admissible additive perturbation signals that do not perturb the convergence point of the algorithm from the average of initial values of the agents; (b) obtaining the necessary and sufficient condition on the knowledge set of a malicious agent that enables it to identify the initial value of another agent; (c) designing observers that internal and external malicious agents can use to identify the initial conditions of another agent when their knowledge set on that agent enables them to do so. We demonstrate our results through a numerical example.**

I. Introduction

Decentralized multi-agent cooperative operations have been emerging as effective solutions for some of today's important socio-economical challenges. However, in some areas involving sensitive data, for example in smart grid, banking or healthcare applications, adaption of these solutions are hindered by concerns regarding the privacy preservation guarantees of the participating clients. Motivated by the demand for privacy preservation evaluations and design of privacy preserving augmentations for existing decentralized solutions, in this paper we consider the privacy preservation problem in the distributed static average consensus problem using additive perturbation signals.

Static average consensus problem in a network of agents each endowed with a local static reference value consists of designing a distributed algorithm that enables each agent to asymptotically obtain the average of the static reference values across the network. The solutions to this problem has been used in various distributed computing, synchronization and estimation problems as well as control of multi-agent cyber physical systems. Average consensus problem has been studied extensively in the literature (see e.g., [1]–[3], [4]). The widely adopted distributed solution for the static average consensus problem is the simple first order Laplacian algorithm in which each agent initializes its local dynamics with its local reference value and transmits this local value to its neighboring agents. Therefore, the reference value is readily revealed to outside world, and thus the privacy of the agents implementing this algorithm is trivially breached. This paper studies the multi-agent static average consensus problem under the privacy preservation requirement against internal and external passive malicious agents in the network. By passive, we mean agents that only listen to the communication messages and want to obtain the reference value of the other agents without interrupting the distributed operation. The solution we examine is to induce privacy preservation property by adding perturbation signals to the internal dynamics and the transmitted output of the agents.

Literature review: Privacy preservation solutions for the average consensus problem have been investigated in the literature mainly in the context of discrete-time consensus algorithms over connected undirected graphs. The general idea is to add perturbation signals to the transmitted out signal of the agents. For example, in one of the early privacy preserving schemes, Kefayati, Talebi and Khalaj [5] proposed that each agent adds a random number generated by zero-mean Gaussian processes to its initial condition. This way the reference value of the agents is guaranteed to stay private but the algorithm does not necessarily converge to the anticipated value. Similarly, in recent years, Nozari, Tallapragada and Cortes [6] also relied on adding zero mean noises to protect the privacy of the agents. However, they develop their noises according to a framework defined based on the concept of differential privacy, which is initially developed in the data science literature [7]–[10]. In this framework, [6] characterizes the convergence degradation and proposes an optimal noise in order to keep a level of privacy to the agents while minimizing the rate of convergence deterioration. To eliminate deviation from desired convergence point, Manitara and Hadjicostis [11] proposed to add a zero sum finite sequence of noises to the transmitted signal of each agent, and Mo and Murray [12] proposed to add a zero sum infinite sequences. Because of the zero sum condition on the perturbation signals, however [11] and [12] show that the privacy of an agent can only be preserved when the malicious agent does not have access to at least one of the signals transmitted to that agent. Additive noises have also been used as a privacy preservation mechanism in other distributed algorithms such as distributed optimization [13] and distributed estimation [14], [15]. A thorough review of these results can be found in a recent tutorial paper [16]. For the discrete-time average consensus, on a different approach, [17] uses a cryptographic approach to preserve the privacy of the agents.

Statement of contributions: We consider the problem of privacy preservation of the continuous-time static Laplacian average consensus algorithm over strongly connected and weight-balanced digraphs using additive perturbation signals. The previous work reviewed above considers discrete-time algorithms over connected undirected graphs. Similar to the reviewed literature, in our privacy preservation analysis, we consider the extreme case that the malicious agents know the

The authors are with the Department of Mechanical and Aerospace Engineering, University of California Irvine, Irvine, CA 92697, {nrezazad,solmaz}@uci.edu. This work is supported by NSF award ECCS-1653838.

graph topology. But, instead of random noises, we use the set of continuous-time integrable additive perturbation signals. In addition to the commonly used additive perturbation signal to the transmitted out signal of the agents, we also add another perturbation signal to the agreement dynamics of the agents as another source of obfuscation in the algorithm. Also, instead of using the customary zero-sum vanishing additive signals, we carefully examine the stability and convergence proprieties of the static average consensus algorithm in the presence of the perturbation signals to find the necessary and sufficient conditions on the perturbation signals such that the integrity of the algorithm is preserved, i.e., despite the perturbation signals the agents still converge to the average of their reference values. We refer to such signals as admissible perturbation signals. An interesting finding, which has not been observed in the literature, is that the perturbation signals do not have to be vanishing. Understanding the nature of the admissible perturbation signals is crucial in the privacy preservation evaluations, as it is rational to assume that the malicious agents are aware of the necessary conditions on such signals.

The necessary and sufficient conditions that specify the admissible perturbation signals of the agents are highly coupled. We discuss how the agents can choose their admissible perturbation signals locally with or without coordination among themselves. The conditions we obtain to define the locally chosen admissible perturbation signals are coupled through a set of under-determined linear algebraic constraints with constant scalar free variables. Then, we evaluate the privacy preservation of the Laplacian average consensus algorithm with additive locally chosen admissible perturbation signals against internal and external malicious agents, depending on whether the coupling variables of the necessary conditions defining the locally chosen admissible perturbation signals are known to the malicious agent or not. We show that when the coupling variables are known to the malicious agents, they can use this extra piece of information to enhance their knowledge set to discover the private value of the other agents. In this case, Our main result then states that the necessary and sufficient condition for a malicious agent to be able to identify the initial value of another agent is to have direct access to all the signals transmitted to and out of the agent. Our next contribution is to design asymptotic observers that internal and external malicious agents can use to identify the initial condition of another agent when their knowledge set on that agent enables them to do so. We characterize also the estimation error of these observers at each time. Our results show that external malicious agents need to use an observer with a higher numerical complexity to compensate for the local state information that internal malicious agents can use. As another contribution, we identify examples of graphs topologies in which the privacy of all the agents are preserved using additive admissible perturbation signals. On the other hand, if the coupling variables of the necessary conditions defining the locally chosen admissible perturbation signals are unknown to the malicious agents, we show that the malicious agents cannot reconstruct the private reference value of the other agents even if they have full access to all the transmitted input and output signals of an agent. We use input-to-state stability (ISS) results (see [18], [19]) to perform our analysis. We demonstrate our results through a numerical example with non-vanishing perturbation signals. A preliminary version of our work has appeared in [20]. In this paper the results are extended in the following directions: (a) we derive the necessary and sufficient conditions to characterize the admissible signals; (b) we study privacy preservation also with respect to external malicious agents; (c) we consider a general class of set of measurable essentially bounded perturbation signals; (d) we improve our main result from sufficient condition to necessary and sufficient condition.

II. Preliminaries

We denote the standard Euclidean norm of vector $\mathbf{x} \in \mathbb{R}^n$ by $\|\mathbf{x}\| = \sqrt{\mathbf{x}^\top \mathbf{x}}$, and the (essential) supremum norm of a signal $f : \mathbb{R}_{\geq 0} \to \mathbb{R}^n$ by $\|f\|_{\text{ess}} = (\text{ess})\sup\{\|f(t)\|, t \geq 0\}$. The set of measurable essentially bounded functions $f : \mathbb{R}_{\geq 0} \to \mathbb{R}^n$ is denoted by $\mathcal{L}_n^\infty$. The set of measurable functions $f : \mathbb{R}_{\geq 0} \to \mathbb{R}^n$ that satisfy $\int_0^t \|f(\tau)\| \text{d}\tau < \infty$ is denoted by $\mathcal{L}_n^1$. For a sets $\mathcal{A}$ and $\mathcal{B}$, the relative complement of $\mathcal{B}$ in $\mathcal{A}$ is $\mathcal{A} \backslash \mathcal{B} = \{x \in \mathcal{A} \,|\, x \notin \mathcal{B}\}$. For a vector $\mathbf{x} \in \mathbb{R}^n$, the sum of its elements is $\texttt{sum}(\mathbf{x})$. In a network of N agents, to emphasize that a variable is local to an agent $i \in \{1, \dots, N\}$, we use superscripts. Moreover, if $p^i \in \mathbb{R}$ is a variable of agent $i \in \{1, \dots, N\}$, the aggregated p^i's of the network is the vector $\mathbf{p} = [\{p^i\}_{i=1}^N] = [p^1, \cdots, p^N]^\top \in \mathbb{R}^N$.

Graph theory: in the following, we review some basic concepts from algebraic graph theory following [21]. A weighted directed graph (digraph) is a triplet $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$, where $\mathcal{V} = \{1, \dots, N\}$ is the *node set*, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the *edge set* and $\mathbf{A} = [\mathsf{a}_{ij}] \in \mathbb{R}^{N \times N}$ is a weighted *adjacency* matrix with the property that $\mathsf{a}_{ij} > 0$ if $(i,j) \in \mathcal{E}$ and $\mathsf{a}_{ij} = 0$, otherwise. A weighted digraph is *undirected* if $\mathsf{a}_{ij} = \mathsf{a}_{ji}$ for all $i, j \in \mathcal{V}$. An edge from i to j, denoted by (i,j), means that agent j can send information to agent i. For an edge $(i,j) \in \mathcal{E}$, i is called an *in-neighbor* of j and j is called an *out-neighbor* of i. We denote the set of the out-neighbors of an agent $i \in \mathcal{V}$ by $\mathcal{N}_{\text{out}}^i$. We define $\mathcal{N}_{\text{out}+i}^i = \mathcal{N}_{\text{out}}^i \cup \{i\}$. A *directed path* is a sequence of nodes connected by edges. A digraph is called *strongly connected* if for every pair of vertices there is a directed path connecting them. We refer to a strongly connected and undirected graph as a *connected graph*. The *weighted out-degree* and *weighted in-degree* of a node i, are respectively, $\mathsf{d}_{\text{in}}^i = \sum_{j=1}^N \mathsf{a}_{ji}$ and $\mathsf{d}_{\text{out}}^i = \sum_{j=1}^N \mathsf{a}_{ij}$. A digraph is *weight-balanced* if at each node $i \in \mathcal{V}$, the weighted out-degree and weighted in-degree coincide (although they might be different across different nodes). The *(out-) Laplacian* matrix is $\mathbf{L} = [\ell_{ij}]$ is $\mathbf{L} = \mathbf{D}^{\text{out}} - \mathbf{A}$, where $\mathbf{D}^{\text{out}} = \text{Diag}(\mathsf{d}_{\text{out}}^1, \cdots, \mathsf{d}_{\text{out}}^N) \in \mathbb{R}^{N \times N}$. Note that $\mathbf{L}\mathbf{1}_N = \mathbf{0}$. A digraph is weight-balanced if and only if $\mathbf{1}_N^\top \mathbf{L} = \mathbf{0}$. For a strongly connected and weight-balanced digraph, $\text{rank}(\mathbf{L}) = N-1$, $\text{rank}(\mathbf{L} + \mathbf{L}^\top) = N-1$, and $\mathbf{L}$ has one zero eigenvalue $\lambda_1 = 0$ and the rest of its eigenvalues have positive real parts. We let $\mathbf{R} \in \mathbb{R}^{N \times (N-1)}$ be a matrix whose columns are normalized orthogonal complement of $\mathbf{1}_N$. Then

$$\mathbf{T}^\top \mathbf{L} \mathbf{T} = \begin{bmatrix} 0 & \mathbf{0} \\ 0 & \mathbf{L}^+ \end{bmatrix}, \quad \mathbf{T} = \begin{bmatrix} \frac{1}{\sqrt{N}} \mathbf{1}_N & \mathbf{R} \end{bmatrix}, \quad \mathbf{L}^+ = \mathbf{R}^\top \mathbf{L} \mathbf{R}. \quad (1)$$

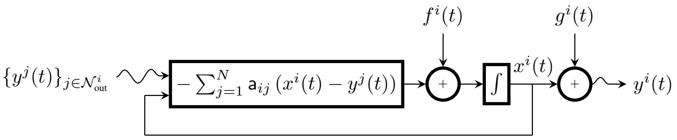

$f^i(t)$
$g^i(t)$
$\{y^j(t)\}_{j\in\mathcal{N}^i_{\text{out}}}$
$-\sum_{j=1}^{N}\mathsf{a}_{ij}\,(x^i(t)-y^j(t))$
$+$
$\int$
$x^i(t)$
$+$
$y^i(t)$

Fig. 1: Graphical representation of algorithm 3.

For a strongly connected and weight-balanced digraph, $-\mathbf{L}^{+}$ is a Hurwitz matrix.

III. Problem formulation

Consider the static average consensus algorithm

$$\dot{x}^i(t) = -\sum\nolimits_{j=1}^{N}\mathsf{a}_{ij}\,(x^i(t)-x^j(t)),\quad x^i(0)=\mathsf{r}^i,\quad (2)$$

over a strongly connected and weight-balanced digraph $\mathcal{G}(\mathcal{V},\mathcal{E},\mathbf{A})$. For such an interaction typology, x^i of each agent $i\in\mathcal{V}$ converges to $\frac{1}{N}\sum_{j=1}^{N}\mathsf{r}^j$ as $t\to\infty$ [4]. In this algorithm, r^i, represents a *reference value* of agent $i\in\mathcal{V}$. Because in (2), the reference value r^i of each agent $i\in\mathcal{V}$ is transmitted to its in-neighbors, this algorithm trivially reveals the reference value r^i of each agent $i\in\mathcal{V}$ to all its in-neighbors and any external agent that is listening to the communication messages. In this paper, we investigate whether in a network of $N\geq 3$ agents, the reference value of the agents can be concealed from the *malicious agents* by adding the perturbation signals $f^i\in\mathcal{L}_1^{\infty}$ and $g^i\in\mathcal{L}_1^{\infty}$ to, respectively, the internal dynamics and the transmitted signal of each agent $i\in\mathcal{V}$ (see Fig. 1), i.e.,

$$\dot{x}^i(t) = -\sum\nolimits_{j=1}^{N}\mathsf{a}_{ij}\,(x^i(t)-y^j(t))+f^i(t),\qquad (3a)$$
$$y^i(t)=x^i(t)+g^i(t),\qquad (3b)$$
$$x^i(0)=\mathsf{r}^i,\qquad (3c)$$

while still guaranteeing that x^i converges to $\frac{1}{N}\sum_{j=1}^{N}\mathsf{r}^j$ as $t\to\infty$. We define the malicious agents formally as follows.

Definition 1 (malicious agent): A malicious agent is an agent inside (internal agent) or outside (external agent) the network that stores and processes the transmitted inter-agent communication messages that it can access so that it can obtain the private reference value of the other agents in the network, without interfering with the execution of algorithm (3). That is, the malicious agents are passive attackers. □

We refer to the set of perturbation signals $\{f^j,g^j\}_{j=1}^{N}$ for which each agent $i\in\mathcal{V}$ still converges to the exact average of the reference values across the network, i.e., $\lim_{t\to\infty}x^i(t)=\frac{1}{N}\sum_{j=1}^{N}x^j(0)=\frac{1}{N}\sum_{j=1}^{N}\mathsf{r}^j$, as the *admissible perturbation signals*.

Theorem 3.1 (The set of necessary and sufficient conditions on the admissible perturbation signals): *Consider algorithm* (3) *over a strongly connected and weight-balanced digraph with perturbation signals $f^i,g^i\in\mathcal{L}_1^{\infty}$, $i\in\mathcal{V}$. Then, the trajectory $t\mapsto x^i(t)$, of all agents $i\in\mathcal{V}$ converges to $\frac{1}{N}\sum_{j=1}^{N}x^j(0)=\frac{1}{N}\sum_{j=1}^{N}\mathsf{r}^j$ as $t\to\infty$ if and only if*

$$\lim_{t\to\infty}\int_0^t\sum\nolimits_{k=1}^{N}(f^k(\tau)+\mathsf{d}^k_{\text{out}}\,g^k(\tau))\,\mathrm{d}\tau=0,\qquad (4a)$$
$$\lim_{t\to\infty}\int_0^t\mathrm{e}^{-\mathbf{L}^{+}(t-\tau)}\mathbf{R}^{\top}(\mathbf{f}(\tau)+\mathbf{A}\,\mathbf{g}(\tau))\,\mathrm{d}\tau=\mathbf{0},\qquad (4b)$$

where $\mathbf{L}^{+}$ and $\mathbf{R}$ are defined in (1). □

The proof of Theorem 3.1 is given in the appendix. The necessary and sufficient conditions in (4) that specify the admissible signals of the agents are highly coupled. The following result gives a representation that the coupling is in the form of a set of linear algebraic constraints.

Theorem 3.2 (Locally chosen admissible signals): *Consider algorithm* (3) *over a strongly connected and weight-balanced digraph. Let each agent $i\in\mathcal{V}$ choose its local perturbation signals $f^i,g^i\in\mathcal{L}_1^{\infty}$ such that*

$$\lim_{t\to\infty}\int_0^t(f^i(\tau)+\mathsf{d}^i_{\text{out}}\,g^i(\tau))\,\mathrm{d}\tau=\beta^i,\qquad (5)$$

where $\beta^i\in\mathbb{R}$. Then, the necessary and sufficient conditions to satisfy (4) *are*

$$\sum\nolimits_{k=1}^{N}\beta^k=0,\qquad (6a)$$
$$\lim_{t\to\infty}\int_0^t\mathrm{e}^{-(t-\tau)}g^i(\tau)\mathrm{d}\tau=\alpha\in\mathbb{R},\quad i\in\mathcal{V}.\qquad (6b)$$

□

The proof of Theorem 3.2 is given in the appendix. We refer to the admissible signals chosen according to (5) and (6) as the *locally chosen admissible signals*. For a given set of $\{\beta^i\}_{i=1}^{N}$ and α, Theorem 3.2 enables the agents to choose their admissible perturbation signals locally with guaranteed convergence to the exact average consensus. Choosing signals that satisfy condition (5) is rather easy. However, condition (6b) appears to be more complex. The result below, whose proof is given in the appendix, identifies three classes of signals that are guaranteed to satisfy condition (6b).

Lemma 3.1 (Signals that satisfy (6b)): *For a given $\alpha\in\mathbb{R}$, let $g=g_1+g_2\in\mathcal{L}_1^{\infty}$ satisfy one of the conditions (a) $\lim_{t\to\infty}g(t)=\alpha$ (b) $\lim_{t\to\infty}g_1(t)=\alpha$ and $\lim_{t\to\infty}\int_0^t g_2(\tau)\mathrm{d}\tau=\bar{g}<\infty$ (c) $\lim_{t\to\infty}g_1(t)=\alpha$ and $\int_0^t\sigma(|g_2(\tau)|)\mathrm{d}\tau<\infty$ for $t\in\mathbb{R}_{\geq 0}$, where σ is any class $\mathcal{K}_{\infty}$ function. Then, $\lim_{t\to\infty}\int_0^t\mathrm{e}^{-(t-\tau)}g(\tau)\mathrm{d}\tau=\alpha$.* □

An interesting fact that Lemma 3.1 reveals is that the admissible perturbation signals $\{f^j,g^j\}_{j=1}^{N}$, unlike the existing results for the discrete-time average consensus algorithm, e.g., in [12], do not necessarily need to be vanishing signals even for $\alpha=0$ and $\beta^i=0$, $i\in\mathcal{V}$. For example, in the numerical example in Section V where $\alpha=0$ and $\beta^i=0$, $i\in\mathcal{V}$, we use $g^i(t)=\sin(i\,t^2)$, $i\in\mathcal{V}$, which is a non-vanishing signal that satisfies condition (b) of Lemma 3.1 ($\lim_{t\to\infty}\int_0^t\sin(i\tau^2)=\sqrt{\frac{\pi}{8i}}$).

We examine the privacy preservation properties of algorithm (3) against non-collaborative malicious agents. The malicious agents are non-collaborative if they do not share their *knowledge sets* with each other. The knowledge set of a malicious agent is the information that it can use to infer the private reference value of the other agents. The extension of our results to collaborative agents is rather straightforward and is omitted for brevity. Without loss of generality, we assume that agent 1 is the malicious internal agent that wants to obtain reference value of other agents in the network. At each time $t \in \mathbb{R}_{\geq 0}$, the signals that are available to agent 1 are

$$\mathcal{Y}^1(t) = \{x^1(\tau), y^1(\tau), \{y^i(\tau)\}_{i \in \mathcal{N}^1_{\text{out}}}\}_{\tau=0}^{t}.$$

For an external malicious agent, the available signals depend on what channels it intercepts. We assume that the external malicious agent can associate the intercepted signals to the corresponding agents. We represent the set of these signals with $\mathcal{Y}^{ext}(t)$. We assume that the malicious agent knows the graph topology. It is also rational to assume that the malicious agents are aware of the form of the necessary conditions on the admissible perturbation signals.

Remark 3.1 (Locally chosen admissible signals): If there exists an ultimately secure and trusted authority that oversees the operation, this authority can assign to each agent its admissible private perturbation signals that collectively satisfy (4). However, in what follows, we consider a scenario where such an authority does not exist, and each agent $i \in \mathcal{V}$, to increase its privacy protection level, wants to choose its own admissible signals (f^i, g^i) privately without revealing them explicitly to the other agents. In this setting, the agents do not know if others are using perturbation signals or not. The only information available to the agents is that their collective choices should satisfy (4). Then, in light of Theorem 3.2, to ensure (4a) each agent $i \in \mathcal{V}$ chooses its local admissible perturbation signals according to (5) with $\beta^i = 0$. Consequently, according to Theorem 3.2 again, each agent $i \in \mathcal{V}$ needs to choose its respective g^i according to (6b) with $\alpha = 0$. Any other choice of $\{\beta^i\}_{i=1}^N$ and α needs an inter-agent coordination/agreement procedure. In case of the locally chosen admissible perturbation signals without inter-agent coordination, since the agents need to satisfy (5) and (6) with $\alpha = \beta^i = 0$, $i \in \mathcal{V}$, these values will be known to the malicious agents. In case that the agents coordinate to choose non-zero values for α and $\{\beta\}_{i=1}^N$ such that (5) and (6) are satisfied, it is likely that these choices to be known to the malicious agents. In our privacy preservation analysis below, we consider both cases when the choices of α and $\{\beta\}_{i=1}^N$ are either known or unknown to the malicious agents. □

Definition 2 (Knowledge set of a malicious agent): The knowledge set of the malicious internal agent 1 and external agent ext is assumed to be one of the cases below,

- Case 1:

$$\mathcal{K}^a = \{\mathcal{Y}^a(\infty), \mathcal{G}(\mathcal{V}, \mathcal{E}, \mathbf{A}), \text{form of conditions (5) and (6)}, \alpha, \{\beta^i\}_{i=1}^N\}, \quad (7)$$

- Case 2:

$$\mathcal{K}^1 = \{\mathcal{Y}^1(\infty), \mathcal{G}(\mathcal{V}, \mathcal{E}, \mathbf{A}), \text{form of conditions (5) and (6)}, \alpha\}, \quad (8)$$

$$\mathcal{K}^{\text{ext}} = \{\mathcal{Y}^{\text{ext}}(\infty), \mathcal{G}(\mathcal{V}, \mathcal{E}, \mathbf{A}), \text{form of conditions (5) and (6)}\}, \quad (9)$$

where $a \in \{1, \text{ext}\}$. □

Our objective in this paper is to determine the effectiveness of use of additive perturbation signals, as introduced in (3), as a privacy preserving measure for the Laplacian average consensus algorithm against internal or external malicious agents with a knowledge set belonging to one of the cases in Definition 2. Our study intends to determine: (a) whether the malicious agents inside or outside the network can obtain the reference value of the other agents by storing and processing the transmitted messages; (b) more specifically, what *knowledge set* enables an agent inside or outside the network to discover the reference value of the other agents in the network; (c) what observers such agents can employ to obtain the reference value of the other agents in the network.

IV. Privacy preservation evaluation

In this section, we evaluate the privacy preservation properties of the modified average consensus algorithm (3) against malicious internal agent 1 and a malicious external agent whose knowledge sets are either of the two cases given in Definition 2. From the perspective of a malicious agent interested in private reference value of another agent $i \in \mathcal{V}$, the dynamical system to observe is (3) with x^i as the internal state, (f^i, g^i, $\{y^j\}_{j \in \mathcal{N}^i_{\text{out}}}$) as the inputs and y^i as the measured output. When inputs and measured outputs over some finite time interval (resp. infinite time) are known, the traditional observability (resp. detectability) tests (see [22], [23]) can determine whether the initial conditions of the system can be identified. However, here the inputs f^i and $g^i : \mathbb{R}_{\geq 0} \to \mathbb{R}$ of agent $i \in \mathcal{V}$ are not available to the malicious agent. All is known is the conditions (5) and (6) that specify the perturbation signals. With regard to inputs $\{y^j\}_{j \in \mathcal{N}^i_{\text{out}}}$ and output y^i, an external agent should intercept these signals while the internal malicious agent 1 has only access to these inputs if it is an in-neighbor of agent i and all the out-neighbors of agent i (e.g., in Fig. 2, agent 1 is an in-neighbor of agent 2 and all the out-neighbors of agent 2).

A. Case 1 knowledge set

The following results show that in a scenario that the malicious agent with the knowledge set (7) has access to all the transmitted input and output signals of another agent i, it can identify the reference value of agent i despite the perturbation signals.

Theorem 4.1 (Observer design for an internal malicious agent with the knowledge set (7)): *Consider the modified static average consensus algorithm* (3) *with a set of locally chosen admissible perturbation signals* $\{f^j, g^j\}_{j=1}^N$ *over a* *strongly connected and weight-balanced digraph $\mathcal{G}$. Let agent 1 be the in-neighbor of agent $i \in \mathcal{V}$ and all the out-neighbors of agent i, i.e., agent 1 knows $\{y^j(t)\}_{j\in\mathcal{N}^i_{\mathrm{out}+i}}$, $t \in \mathbb{R}_{\geq 0}$. Let the knowledge set of agent 1 be (7). Then, agent 1 can employ the observer*

$$\dot{\zeta} = \sum\nolimits_{j=1}^{N} \mathsf{a}_{ij}(y^i - y^j), \qquad \zeta(0) = -\beta^i, \tag{10a}$$

$$\nu(t) = \zeta(t) + x^1(t), \tag{10b}$$

to asymptotically obtain r^i, i.e., $\nu \to \mathsf{r}^i$ as $t \to \infty$. Moreover, at any time $t \in \mathbb{R}_{\geq 0}$, the estimation error of the observer satisfies

$$\nu(t) - \mathsf{r}^i = x^1(t) - x^i(t) + \int_0^t (f^i(\tau) + \mathsf{d}^i_{\mathrm{out}}\, g^i(\tau))\mathrm{d}\tau \;-\beta^i. \tag{11}$$

Proof 1: Given (3) and (10) we can write

$$\dot{\zeta} + \dot{x}^i = f^i + \mathsf{d}^i_{\mathrm{out}}\, g^i$$

which, because of $x^i(0) = \mathsf{r}^i$ and $\zeta(0) = -\beta^i$, gives

$$\zeta(t) = -x^i(t) + \mathsf{r}^i + \int_0^t (f^i(\tau) + \mathsf{d}^i_{\mathrm{out}}\, g^i(\tau))\mathrm{d}\tau - \beta^i, \;\; t \in \mathbb{R}_{\geq 0}.$$

Then, using (10b) and (3b) we obtain (11) as the estimation error. Subsequently, because of (5) and since $\lim_{t\to\infty}(x^1(t) - x^i(t)) = 0$, from (11) we obtain $\lim_{t\to\infty}\nu(t) = \mathsf{r}^i$.

To construct the observer (10), the internal malicious agent used its own local state. The following result shows that an external malicious agent can compensate for the lack of this internal state information by employing a higher order observer and also invoking condition (6b), which the internal malicious agent does not need. This means that an external malicious agent incurs a higher computational cost.

Theorem 4.2 (Observer design for an external malicious agent with the knowledge set (7)): *Consider the modified static average consensus algorithm (3) with a set of locally chosen admissible perturbation signals $\{f^j, g^j\}_{j=1}^N$ over a strongly connected and weight-balanced digraph $\mathcal{G}$. Consider an external malicious agent that has access to the output signals of agent $i \in \mathcal{V}$ and all its out-neighbors, i.e., $\{y^j(t)\}_{j\in\mathcal{N}^i_{\mathrm{out}+i}}$, $t \in \mathbb{R}_{\geq 0}$. Let the knowledge set of this agent be (7). Then, this external malicious agent can employ the observer*

$$\dot{\zeta} = \sum\nolimits_{j=1}^{N} \mathsf{a}_{ij}(y^i - y^j), \qquad \zeta(0) = -\beta^i - \alpha, \tag{12a}$$

$$\dot{\eta} = -\eta + y^i, \qquad \eta(0) \in \mathbb{R}, \tag{12b}$$

$$\nu(t) = \zeta(t) + \eta(t), \tag{12c}$$

to asymptotically obtain r^i, $i \in \mathcal{V}$, i.e., $\nu \to \mathsf{r}^i$ as $t \to \infty$. Moreover, at any time $t \in \mathbb{R}_{\geq 0}$, the estimation error of the observer satisfies

$$\nu(t) - \mathsf{r}^i = \eta(t) - x^i(t) + \int_0^t (f^i(\tau) + \mathsf{d}^i_{\mathrm{out}}\, g^i(\tau))\mathrm{d}\tau - \beta^i - \alpha, \tag{13}$$

where

$$\eta(t) = \mathrm{e}^{-t}\eta_0 + \int_0^t \mathrm{e}^{-(t-\tau)}x^i(\tau)\mathrm{d}\tau + \int_0^t \mathrm{e}^{-(t-\tau)}g^i(\tau)\mathrm{d}\tau. \tag{14}$$

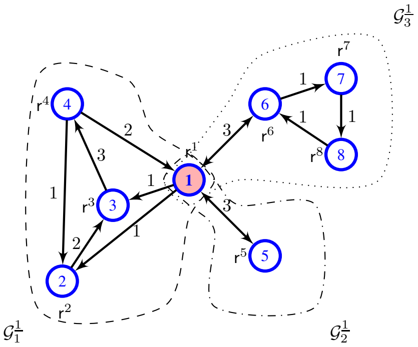


Fig. 2: A strongly connected and weight-balanced digraph $\mathcal{G}$ in which node 1 is an articulation point of the undirected representation of $\mathcal{G}$. $\mathcal{G}^1_1$, $\mathcal{G}^1_2$ and $\mathcal{G}^1_3$ are the islands of agent 1.

Proof 2: Given (3) and (12a), we can write

$$\dot{\zeta} + \dot{x}^i = f^i + \mathsf{d}^i_{\mathrm{out}}\, g^i,$$

which given $x^i(0) = \mathsf{r}^i$ and $\zeta(0) = -\beta^i - \alpha$, for $t \in \mathbb{R}_{\geq 0}$ gives

$$\zeta(t) = -x^i(t) + \mathsf{r}^i + \int_0^t (f^i(\tau) + \mathsf{d}^i_{\mathrm{out}}\, g^i(\tau))\mathrm{d}\tau \;- \beta^i - \alpha. \tag{15}$$

On the other hand, using (3b), $t \mapsto \eta(t)$ is obtained from (14). Then, tracking error (13) is readily deduced from (12c) and (15). Next, given (5) and (6b) and also $\lim_{t\to\infty} \mathrm{e}^{-t}\eta_0 = 0$, we obtain $\lim_{t\to\infty}\nu(t) = \mathsf{r}^i + \lim_{t\to\infty}(-x^i(t) + \int_0^t \mathrm{e}^{-(t-\tau)}x^i(\tau)\mathrm{d}\tau)$. Subsequently, since $\lim_{t\to\infty} x^i(t) = \frac{1}{N}\sum_{j=1}^{N}\mathsf{r}^j$, we can conclude our proof by invoking Lemma 7.2 that guarantees $\lim_{t\to\infty}\int_0^t \mathrm{e}^{-(t-\tau)}x^i(\tau)\mathrm{d}\tau = \lim_{t\to\infty} x^i(t) = \frac{1}{N}\sum_{j=1}^{N}\mathsf{r}^j$.

When a malicious agent does not have direct access to all the signals in $\{y^j(t)\}_{j\in\mathcal{N}^i_{\mathrm{out}+i}}$, a rational strategy appears to be that the malicious agent estimates the signals it does not have access to. If those agents also have out-neighbors that their output signals are not available to the malicious agent, then the malicious agent should estimate the state of those agents as well, until the only inputs to the dynamics that it observes are the additive admissible perturbation signals. For example, in Fig. 2, to obtain the reference value of agent 6, agent 1 compensates for the lack of direct access to $y^7(t)$, which enter the dynamics of agent 6, by estimating the state of all the agents in subgraph $\mathcal{G}^1_3$. Our results below however show that this strategy is not effective. In fact, we show that a malicious agent (internal or external) is able to uniquely identify the reference value of an agent $i \in \mathcal{V}$ if and only if it has direct access to $\{y^j(t)\}_{j\in\mathcal{N}^i_{\mathrm{out}+i}}$ for all $t \in \mathbb{R}_{\geq 0}$.

To present our results, we first introduce some notations. Let $\bar{\mathcal{V}}^1_k$, $k \in \{1,\dots,\bar{n}^1\}$ be the set of the agents in the k^{th} induced disjoint subgraph obtained from removal of agent 1 and its incident edges.

Recall that if 1 is an articulation point[1] of the undirected representation of digraph $\mathcal{G}$, then $\bar{n}^1 > 1$, otherwise $\bar{n}^1 = 1$.

[1] An articulation point of an undirected connected graph is a node whose removal along with its incident edges disconnects the graph [24].

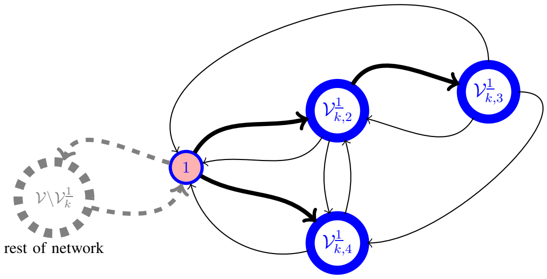


Fig. 3: The k^{th} induced island of malicious agent 1. The super node $\mathcal{V}^1_{k,2}$ in $\mathcal{G}^1_k$ is the set of the out-neighbors of agent 1 that each of them has at least one out-neighbor that is not an out-neighbor of agent 1. The super node $\mathcal{V}^1_{k,4}$ is the set of the out-neighbors of agent 1 whose out-neighbors are all also out-neighbors of agent 1. Finally, the super node $\mathcal{V}^1_{k,3}$ is the set of the agents in $\mathcal{G}^1_k$ that are not an out-neighbor of agent 1. An arrow from each node a (agent 1 or each super node) to another node b (agent 1 or each super node) indicates that at least one agent in a can obtain information from at least one agent in b. The thin connection lines may or may not exist in a network.

We refer to every induced subgraph $\mathcal{G}^1_k = (\mathcal{V}^1_k, \mathcal{E}^1_k) \subset \mathcal{G}(\mathcal{V}, \mathcal{E})$, $k \in \{1, \dots, \bar{n}^1\}$, where $\mathcal{V}^1_k = \bar{\mathcal{V}}^1_k \cup \{1\}$ and $\mathcal{E}^1_k = \{(l, j) \in \mathcal{E} | \, l \in \mathcal{V}^1_k, \; j \in \mathcal{V}^1_k\}$, as the k^{th} island of agent 1. Note that every island of agent 1 is connected to the rest of the digraph $\mathcal{G}$ only through agent 1 (see Fig. 2 for an example). Let $\mathcal{G}^1_1 = (\mathcal{V}^1_1, \mathcal{E}^1_1)$ be the island of agent 1 that includes agent 2, the out-neighbor of agent 1 that agent 1 wants to obtain its reference value r^2. Because every agent in $\mathcal{G}^1_1$ is connected to the rest of the agents in digraph $\mathcal{G}$ only through agent 1, all the out-neighbors and in-neighbors of agent 2 are necessarily in $\mathcal{G}^1_1$. Based on how each agent interacts with agent 1, we divide the agents of island $\mathcal{G}^1_1$ into three groups as described below (see Fig. 3)

- $\mathcal{V}^1_{1,2} = \big\{i \in \mathcal{V}^1_1 \,\big|\, i \in \mathcal{N}^1_{\text{out}}, \; \mathcal{N}^i_{\text{out}} \not\subset \mathcal{N}^1_{\text{out}+1}\big\}$,
- $\mathcal{V}^1_{1,3} = \big\{i \in \mathcal{V}^1_1 \,\big|\, i \notin \mathcal{N}^1_{\text{out}}\big\}$.
- $\mathcal{V}^1_{1,4} = \big\{i \in \mathcal{V}^1_1 \,\big|\, i \in \mathcal{N}^1_{\text{out}}, \; \mathcal{N}^i_{\text{out}} \subseteq \mathcal{N}^1_{\text{out}+1}\big\}$,

Without loss of generality, in what follows we assume that the agents in the network are labeled according to the ordered set $(1, \mathcal{V}^1_{1,2}, \mathcal{V}^1_{1,3}, \mathcal{V}^1_{1,4}, \mathcal{V}\backslash\mathcal{V}^1_1)$. We let the aggregated states and perturbation signals of the agents in $\mathcal{V}^1_{1,l}$, $l \in \{2, 3, 4\}$, be $\mathbf{x}_l = [x^i]_{i \in \mathcal{V}^1_{1,l}}$, $\mathbf{g}_l = [g^i]_{i \in \mathcal{V}^1_{1,l}}$ and $\mathbf{f}_l = [f^i]_{i \in \mathcal{V}^1_{1,l}}$. Similarly, we let the aggregated states and perturbation signals of the agents in $\mathcal{V}\backslash\mathcal{V}^1_1$ be $\mathbf{x}_5 = [x^i]_{i \in \mathcal{V}\backslash\mathcal{V}^1_1}$, $\mathbf{g}_5 = [g^i]_{i \in \mathcal{V}\backslash\mathcal{V}^1_1}$ and $\mathbf{f}_5 = [f^i]_{i \in \mathcal{V}\backslash\mathcal{V}^1_1}$. We partition L, A and D^{out}, respectively, to subblock matrices L_{ij}'s, A_{ij}'s and $\mathsf{D}^{\text{out}}_{ij}$'s in a comparable manner to the partitioned aggregated state $(x^1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5)$ (see (44)). By definition $\mathsf{L}_{ij} = -\mathsf{A}_{ij}$, $i, j \in \{1, \cdots, 5\}$, $i \neq j$.

Lemma 4.1 (A case of indistinguishable admissible initial conditions for an internal malicious agent): *Consider the modified static average consensus algorithm* (3) *with a set of locally chosen admissible perturbation signals* $\{f^j, g^j\}_{j=1}^N$ *over a strongly connected and weight-balanced digraph* $\mathcal{G}$. *Let* $t \mapsto y^i(t)$ *be the transmitted signal from agent* $i \in \mathcal{V}$ *for* $t \in \mathbb{R}_{\geq 0}$. *Let* $\mathcal{G}^1_1 = (\mathcal{V}^1_1, \mathcal{E}^1_1)$ *be an island of agent 1 that satisfies* $\mathcal{V}^1_{1,2} \neq \{\}$. *Now consider an alternative implementation of algorithm* (3a)-(3b) *with initial condition*

$$\begin{aligned} &x^{i\prime}(0) = x^i(0) = \mathsf{r}^i, && i \in \mathcal{V}\backslash(\mathcal{V}^1_{1,2} \cup \mathcal{V}^1_{1,3}),\\ &x^{i\prime}(0) \in \mathbb{R}, && i \in (\mathcal{V}^1_{1,2} \cup \mathcal{V}^1_{1,3}), \;\; s.t. \\ &\mathbf{x}'_2(0) - \mathbf{x}_2(0) = -\mathsf{A}_{23}\mathsf{L}_{33}^{-1}(\mathbf{x}'_3(0) - \mathbf{x}_3(0)), \end{aligned} \tag{16}$$

and perturbation signals

$$\begin{aligned} &g^{i\prime}(t) = g^i(t), \quad f^{i\prime}(t) = f^i(t), && i \in \mathcal{V}\backslash\mathcal{V}^1_{1,2},\\ &g^{i\prime}(t) = g^i(t) + \mathsf{e}^{-\mathsf{d}^i_{\text{out}} t}(x^{i\prime}(0) - x^i(0)), && i \in \mathcal{V}^1_{1,2}, \end{aligned} \tag{17}$$

$$f^{i\prime}(t) = f^i(t) - \Big[\mathsf{A}_{23}\mathsf{e}^{-\mathsf{L}_{33} t}(\mathbf{x}'_3(0) - \mathbf{x}_3(0))\Big]_{i-1}, \quad i \in \mathcal{V}^1_{1,2}.$$

Let $t \mapsto x^{i\prime}(t)$ *and* $t \mapsto y^{i\prime}(t)$, $t \in \mathbb{R}_{\geq 0}$, *respectively, be the state and the transmitted signal of agent* $i \in \mathcal{V}$ *in this case. Then,*

$$\sum\nolimits_{j=1}^N x^{j\prime}(0) = \sum\nolimits_{j=1}^N x^j(0) = \sum\nolimits_{j=1}^N \mathsf{r}^j, \tag{18}$$

$$\lim_{t \to \infty} x^{i\prime}(t) = \frac{1}{N}\sum\nolimits_{j=1}^N \mathsf{r}^j, \qquad i \in \mathcal{V}. \tag{19}$$

Moreover,

$$y^j(t) = y^{j\prime}(t), \quad t \in \mathbb{R}_{\geq 0}, \qquad j \in \mathcal{V}\backslash\mathcal{V}^1_{1,3}. \tag{20}$$

□

The proof of Lemma 4.1 is given in the Appendix.

Lemma 4.1 states that there exists infinite number of admissible initial conditions and admissible perturbation signals for an agent $i \in \mathcal{N}^1_{\text{out}}$ and any agent $j \in \mathcal{N}^i_{\text{out}}\backslash\mathcal{N}^1_{\text{out}+1} \neq \{\}$ that agent 1 cannot distinguish between, because for all of these cases, the signals transmitted from any out-neighbor of agent 1 are identical. We can develop similar results, as stated in the corollary below, for an external malicious agent that does not have direct access to the output signal of some of the out-neighbors of agent $i \in \mathcal{V}$. The proof of this corollary is omitted for brevity.

Corollary 4.1 (A case of indistinguishable admissible initial conditions for an external malicious agent): *Consider the modified static average consensus algorithm* (3) *with a set of locally chosen admissible perturbation signals* $\{f^j, g^j\}_{j=1}^N$ *over a strongly connected and weight-balanced digraph* $\mathcal{G}$. *Let* $t \mapsto y^i(t)$ *be the transmitted signal from agent* $i \in \mathcal{V}$ *for* $t \in \mathbb{R}_{\geq 0}$. *Consider an external malicious agent that has direct access to the output signal of agent* $2 \in \mathcal{V}$ *but not that of the agent* $3 \in \mathcal{N}^2_{\text{out}}$. *Now consider an alternative implementation of algorithm* (3a)-(3b) *with initial condition* $x^{i\prime}(0) = x^i(0) = \mathsf{r}^i$ *for* $i \in \mathcal{V}\backslash\{2, 3\}$, *and* $x^{2\prime}(0), x^{3\prime}(0) \in \mathbb{R}$ *such that* $x^{2\prime}(0) - x^2(0) = -\frac{\mathsf{a}_{23}}{\ell_{33}}(x^{3\prime}(0) - x^3(0))$, *and perturbation signals* $g^{i\prime}(t) = g^i(t)$, $f^{i\prime}(t) = f^i(t)$, *for* $i \in \mathcal{V}\{2\}$, *and* $g^{2\prime}(t) = g^2(t) + \mathsf{e}^{-\mathsf{d}^2_{\text{out}} t}(x^{2\prime}(0) - x^2(0))$ *and* $f^{2\prime}(t) = f^2(t) - \mathsf{a}_{23}\mathsf{e}^{-\ell_{33} t}(x^{3\prime}(0) - x^3(0))$. *Let* $t \mapsto x^{i\prime}(t)$ *and* $t \mapsto y^{i\prime}(t)$, $t \in \mathbb{R}_{\geq 0}$, *respectively, be the state and the transmitted signal of agent* $i \in \mathcal{V}$ *in this case. Then, the equations* (18) *and* (19) *hold. Moreover,*

$$y^j(t) = y^{j\prime}(t), \quad t \in \mathbb{R}_{\geq 0}, \qquad j \in \mathcal{V}\backslash\{3\}.$$

□

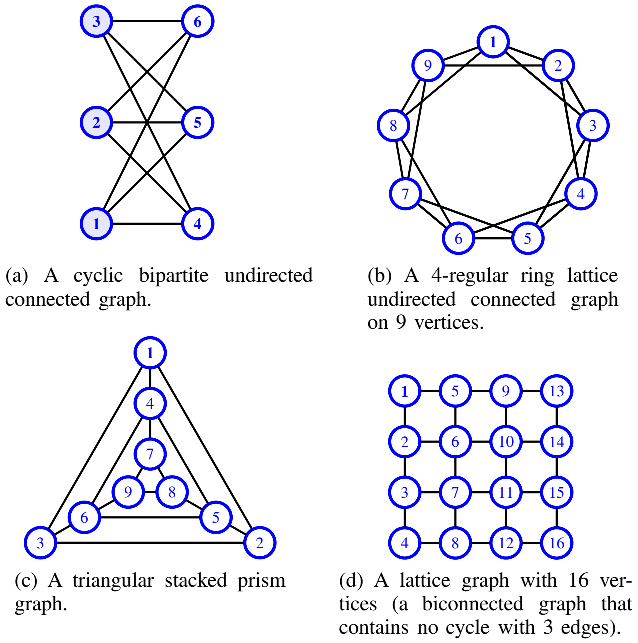
Building on our results so far, we are now ready to state the necessary and sufficient condition under which a malicious agent with knowledge set (7) can discover the reference value of an agent $i \in \mathcal{V}$.

Theorem 4.3 (Privacy preservation using the modified average consensus algorithm (3) when the knowledge set of the malicious agents is given by Case 1 in Definition 2): *Consider the modified static average consensus algorithm* (3) *with a set of locally chosen admissible perturbation signals* $\{f^i, g^i\}_{i=1}^N$ *over a strongly connected and weight-balanced digraph* $\mathcal{G}$. *Let the knowledge set of the internal malicious agent* 1 *and external agent* ext *be* (7). *Then, (a) agent* 1 *can reconstruct the exact initial value of agent* $i \in \mathcal{V}\backslash\{1\}$ *if and only if* $i \in \mathcal{N}^1_{\text{out}}$ *and* $\mathcal{N}^i_{\text{out}} \subseteq \mathcal{N}^1_{\text{out}+1}$*; (b) the external agent* ext *can reconstruct the exact initial value of agent* $i \in \mathcal{V}$ *if and only if* $\{\{y^j(\tau)\}_{j\in\mathcal{N}^i_{\text{out}+i}}\}_{\tau=0}^{\infty} \subseteq \mathcal{Y}^{\text{ext}}(\infty)$.

Proof 3: Proof of statement (a): If $i \in \mathcal{N}^1_{\text{out}}$ and $\mathcal{N}^i_{\text{out}} \subseteq \mathcal{N}^1_{\text{out}+1}$, Theorem (4.1) guarantees that agent 1 can employ an observer to obtain the reference value of agent i. Next, we show that if $i \notin \mathcal{N}^1_{\text{out}}$ or $\mathcal{N}^i_{\text{out}} \not\subset \mathcal{N}^1_{\text{out}+1}$, then agent 1 cannot uniquely identify the reference value r^i of agent i. Suppose agent $i \in \mathcal{V}\backslash\{1\}$ satisfies $i \notin \mathcal{N}^1_{\text{out}}$ (resp. $i \in \mathcal{N}^1_{\text{out}}$ and $\mathcal{N}^i_{\text{out}} \not\subset \mathcal{N}^1_{\text{out}+1}$). Without loss of generality let $\mathcal{V}^1_1$ be the island of agent 1 that contains this agent i. Consequently, $i \in \mathcal{V}^1_{1,3}$ (resp. $i \in \mathcal{V}^1_{1,2}$). Then, by virtue of Lemma 4.1, we know that there exists infinite number of alternative admissible initial conditions and corresponding admissible perturbation signals for any agents in $\mathcal{V}^1_{1,3} \cup \mathcal{V}^1_{1,2}$ for which the time histories of each signal transmitted to agent 1 are identical. Therefore, agent 1 cannot uniquely identify the initial condition of any agents in $\mathcal{V}^1_{1,3} \cup \mathcal{V}^1_{1,2}$. In light of Theorem 4.2 and Corollary 4.1, the proof of statement (b) is similar to that of statement (a) and is omitted for brevity.

Remark 4.1 (Examples of privacy preserving graph topologies): Cyclic bipartite undirected graphs, 4-regular ring lattice undirected graphs with $N > 5$, planar stacked prism graphs, directed ring graphs, and any biconnected undirected graph that does not contain a cycle with 3 edges are examples of graph topologies for which the privacy of every agent is preserved with respect to any internal malicious agent (see [25] for the formal definition of these graph topologies). This is because for every malicious agent i in the network, every $j \in \mathcal{N}^i_{\text{out}}$ has a neighbor $k \in \mathcal{N}^j_{\text{out}}$ such that $k \notin \mathcal{N}^i_{\text{out}}$ (recall Theorem 4.3). Some examples of these privacy preserving topologies is shown in Fig. 4. □

Next, we show that even though agent 1 cannot obtain the initial condition of the individual agents in $\mathcal{V}^1_{k,2} \neq \{\}$ and $\mathcal{V}^1_{k,3}$, $k \in \{1, \cdots \bar{n}^1\}$, it can obtain the average of the initial conditions of those agents. Without loss of generality, we demonstrate our results for $k = 1$.



(a) A cyclic bipartite undirected connected graph.

(b) A 4-regular ring lattice undirected connected graph on 9 vertices.

(c) A triangular stacked prism graph.

(d) A lattice graph with 16 vertices (a biconnected graph that contains no cycle with 3 edges).

Fig. 4: Examples of privacy preserving graph topologies.

Proposition 4.1 (Island anonymity): *Consider the dynamic consensus algorithm* (3) *over a strongly connected and weight-balanced digraph* $\mathcal{G}$ *in which* $\mathcal{V}^1_{1,2} \neq \{\}$. *Let* $n_{2,3} = |\mathcal{V}^1_{1,2} \cup \mathcal{V}^1_{1,3}|$ *and* $\mathsf{d}^{1,1}_{\text{out}} = \sum_{j\in(\mathcal{V}^1_{1,2}\cup\mathcal{V}^1_{1,4})} \mathsf{a}_{1j}$ *be the out-degree of agent* 1 *in subgraph* $\mathcal{G}^1_1$. *Then, the malicious agent* 1 *with the knowledge set* (7) *can employ the observer*

$$\begin{aligned}
\dot{\zeta}_i &= \sum\nolimits_{j=1}^{N} \mathsf{a}_{ij}(y^i - y^j), & \zeta_i(0) &= -\beta^i, \quad i \in \mathcal{V}^1_{1,4},\\
\dot{\eta} &= -\sum_{j\in(\mathcal{V}^1_{1,2}\cup\mathcal{V}^1_{1,4})} a_{1j}(y^1 - y^j), & \eta(0) &= -\sum_{j\in\mathcal{V}^1_1\backslash\{1\}} \beta^i,
\end{aligned}$$

$$\mu(t) = \frac{\eta(t) - \sum_{i\in\mathcal{V}^1_{1,4}} \zeta_i}{n_{2,3}} + x^1(t).$$

to have $\lim_{t\to\infty} \mu(t) = \frac{1}{n_{2,3}} \sum_{j\in(\mathcal{V}^1_{1,2}\cup\mathcal{V}^1_{1,3})} \mathsf{r}^j$.

Proof 4: Consider the aggregate dynamics of η and $\mathbf{x}_i$, $i \in \{2, 3, 4\}$, which reads as

$$\begin{bmatrix} \dot{\eta} \\ \dot{\mathbf{x}}_2 \\ \dot{\mathbf{x}}_3 \\ \dot{\mathbf{x}}_4 \end{bmatrix} = -\underbrace{\begin{bmatrix} \mathsf{d}^{1,1}_{\text{out}} & -\mathbf{A}_{12} & \mathbf{0} & -\mathbf{A}_{14} \\ -\mathbf{A}_{21} & \mathbf{D}^{\text{out}}_{22} & -\mathbf{A}_{23} & -\mathbf{A}_{24} \\ -\mathbf{A}_{31} & -\mathbf{A}_{32} & \mathbf{D}^{\text{out}}_{33} & \mathbf{0} \\ -\mathbf{A}_{41} & -\mathbf{A}_{42} & \mathbf{0} & \mathbf{D}^{\text{out}}_{44} \end{bmatrix}}_{\mathbf{L}^1_1} \begin{bmatrix} y^1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \\ \mathbf{y}_4 \end{bmatrix} + \begin{bmatrix} 0 \\ \mathbf{f}_2 + \mathbf{D}^{\text{out}}_{22}\mathbf{g}_2 \\ \mathbf{f}_3 + \mathbf{D}^{\text{out}}_{33}\mathbf{g}_3 \\ \mathbf{f}_4 + \mathbf{D}^{\text{out}}_{44}\mathbf{g}_4 \end{bmatrix}.$$

Notice that $\mathbf{L}^1_1$ is the Laplacian matrix of graph $\mathcal{G}^1_1$. By Virtue of Lemma 7.3 in the appendix we know that $\mathcal{G}^1_1$ is a strongly connected and weight-balanced digraph. Consequently, left multiplying both sides of equation above with $\mathbf{1}^\top_{|\mathcal{V}^1_1|}$ gives

$$\dot{\eta} + \sum_{j\in\mathcal{V}^1_1\backslash\{1\}} x^i = \sum_{j\in\mathcal{V}^1_1\backslash\{1\}} (f^j(t) + \mathsf{d}^j_{\text{out}}\, g^j(t)).$$

Thereby, given $\eta(0) = -\sum\limits_{j\in\mathcal{V}^{1}_{1}\backslash\{1\}} \beta^i$ and $x^i(0) = \mathsf{r}^i$, we obtain

$$\eta(t) = \sum_{j\in\mathcal{V}^{1}_{1}\backslash\{1\}} \mathsf{r}^j - \sum_{j\in\mathcal{V}^{1}_{1}\backslash\{1\}} x^j(t) + \sum_{j\in\mathcal{V}^{1}_{1}\backslash\{1\}} \int_0^t (f^j(\tau) + \mathsf{d}^j_{\text{out}}\, g^j(\tau))\mathrm{d}\tau \\ - \sum_{j\in\mathcal{V}^{1}_{1}\backslash\{1\}} \beta^i.$$

On the other hand, following the proof of Theorem 4.1, we can conclude that

$$\sum_{i\in\mathcal{V}^{1}_{1,4}} \zeta_i(t) = \sum_{i\in\mathcal{V}^{1}_{1,4}} \mathsf{r}^i - \sum_{i\in\mathcal{V}^{1}_{1,4}} x^i(t) + \sum_{i\in\mathcal{V}^{1}_{1,4}} \int_0^t (f^i(\tau) + \mathsf{d}^i_{\text{out}}\, g^i(\tau))\mathrm{d}\tau \\ - \sum_{i\in\mathcal{V}^{1}_{1,4}} \beta^i.$$

Therefore, we can write

$$n_{2,3}\,\mu(t) = \sum_{j\in(\mathcal{V}^{1}_{1,2}\cup\mathcal{V}^{1}_{1,3})} \mathsf{r}^i - \sum_{j\in(\mathcal{V}^{1}_{1,2}\cup\mathcal{V}^{1}_{1,3})} x^i(t) - \sum_{j\in(\mathcal{V}^{1}_{1,2}\cup\mathcal{V}^{1}_{1,3})} \beta^i \\ + \sum_{j\in(\mathcal{V}^{1}_{1,2}\cup\mathcal{V}^{1}_{1,3})} \int_0^t (f^j(\tau) + \mathsf{d}^j_{\text{out}}\, g^j(\tau))\mathrm{d}\tau + n_{2,3}\, x^1(t).$$

The proof then follows from the necessary condition (5) on the perturbation signals, and the fact that $\lim_{t\to\infty} n_{2,3}\, x^1(t) - \sum_{j\in(\mathcal{V}^{1}_{1,2}\cup\mathcal{V}^{1}_{1,3})} x^i(t) = 0$ (recall that $\lim_{t\to\infty} x^i(t) = \lim_{t\to\infty} x^j(t),\ \forall i, j \in \mathcal{V}$).

B. Case 2 knowledge set

The first result below shows that if β^i corresponding to the locally chosen admissible perturbation signals of an agent $i \in \mathcal{V}$ is not known to the malicious agent, the privacy of the agent i is preserved even if the malicious agent knows all the transmitted input and output signals of agent i and the parameter α. The proof of this lemma is given in the appendix.

Lemma 4.2 (Privacy preservation for $i \in \mathcal{V}$ via a concealed β^i): *Consider the modified static average consensus algorithm* (3) *with a set of locally chosen admissible perturbation signals* $\{f^j, g^j\}_{j=1}^N$ *over a strongly connected and weight-balanced digraph* $\mathcal{G}$*. Let the knowledge set of the malicious agent* 1 *include the form of conditions* (5) *and* (6)*, and also the parameter* α *that the agents agreed to use. Let agent* 1 *be the in-neighbor of agent* $i \in \mathcal{V}$ *and all the out-neighbors of agent* i*, i.e., agent* 1 *knows* $\{y^j(t)\}_{j\in\mathcal{N}^i_{\text{out}+i}}$*,* $t \in \mathbb{R}_{\geq 0}$*. Then, the malicious agent* 1 *can obtain* r^i *of agent* i *if and only if it knows* β^i*.*

A similar statement to that of Lemma 4.2 can be made about an external malicious agent. In case of the external malicious agent, it is very likely that the malicious agent does not know α, as well. Building on the result of Lemma 4.2, we make our final formal privacy preservation statement as follows.

Theorem 4.4 (Privacy preservation using the modified average consensus algorithm (3) when the knowledge set of the malicious agents is given by Case 2 in Definition 2): *Consider the modified static average consensus algorithm* (3) *with a set of locally chosen admissible perturbation signals* $\{f^j, g^j\}_{j=1}^N$ *over a strongly connected and weight-balanced digraph* $\mathcal{G}$*. Let the knowledge set of the internal malicious agent* 1 *and the external malicious agent* ext *be given by Case 2 in Definition 2. Then, the malicious agent 1 (resp. agent* ext*) cannot reconstruct the reference value* r^i *of any agent* $i \in \mathcal{V}\backslash\{1\}$ *(resp.* $i \in \mathcal{V}$*).*

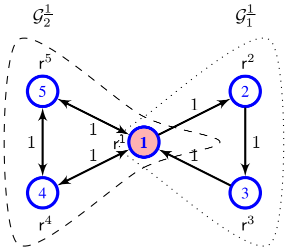


Fig. 5: A strongly connected and weight-balanced digraph $\mathcal{G}$.

Proof 5: Any agent $i \in \mathcal{V}\backslash\{1\}$ satisfies either $\mathcal{N}^i_{\text{out}+i} \subset \mathcal{N}^1_{\text{out}+1}$ or $\mathcal{N}^i_{\text{out}+i} \not\subset \mathcal{N}^1_{\text{out}+1}$. Since the malicious agent 1 does not know $\{\beta^i\}_{j=2}^N$, if $\mathcal{N}^i_{\text{out}+i} \subset \mathcal{N}^1_{\text{out}+1}$, $i \in \mathcal{V}\backslash\{1\}$, (agent 1 has access to all the transmitted input and output signals of agent i), it follows from Lemma 4.2 that it cannot reconstruct r^i. Consequently, if $\mathcal{N}^i_{\text{out}+i} \not\subset \mathcal{N}^1_{\text{out}+1}$, $i \in \mathcal{V}\backslash\{1\}$, since the malicious agent 1 lacks more information (it does not have access to some or all of the transmitted input and output signals of agent i), we conclude that the malicious agent 1 cannot reconstruct r^i. The proof of the statement for the external malicious agent is similar to that of the internal malicious agent 1, and is omitted for brevity (note here that the malicious external agent ext lacks the knowledge of α, as well).

Remark 4.2 (Guaranteed privacy preservation when an ultimately secure authority assigns the admissible perturbation signals): If there exists an ultimately secure and trusted authority that assigns the agents' admissible private perturbation signals in a way that they collectively satisfy (4), the privacy of the agents is not trivially guaranteed. This is because, it is rational to assume that the malicious agents know the necessary condition (4) and may be able to exploit it to their benefit. However, in light of Theorem 4.4, we are now confident to offer the privacy preservation guarantee for such a case. This is because, in this case the malicious agents' knowledge set lacks more information than Case 2 in Definition 2 (note that the locally chosen admissible perturbation signals are a specially structured subset of all the possible classes of the admissible perturbation signals).

V. Numerical example

We demonstrate our results using an execution of the modified static average consensus algorithm (3) over the strongly connected and weight-balanced digraph in Fig. 5 where the parameters specifying the admissible signals are set at $\alpha = 0$ and $\beta^i = 0$, $i \in \mathcal{V}$ and are known to the malicious agents.

The local reference value of the agents as well the admissible perturbations they each use are given by

$$\mathsf{r}^1\!=\!3,\ \ \mathsf{r}^2\!=\!2,\ \ \mathsf{r}^3\!=\!5, \mathsf{r}^4\!=\!-3,\ \ \mathsf{r}^5\!=\!-1,$$

$$f^i(t) = \mathsf{d}^i_{\text{out}}\frac{\sqrt{(2\,i)\pi}}{4i}\mathsf{e}^{-t},\quad g^i(t) = \sin(i\,t^2),\quad i \in \mathcal{V}. \quad (21)$$

The malicious agent here is agent 1. In regards to agents 4 and 5, despite use of non-vanishing perturbation signals g^4 and g^5, as guaranteed in Theorem 4.2, agent 1 can employ observers of the form (10) to obtain $x^4(0) = \mathsf{r}^4 = -3$ and $x^5(0) = \mathsf{r}^5 = -1$ (see Fig. 6(c)). Agent 1 however, cannot uniquely identify r^2 and r^3, since $\mathcal{N}^2_{\text{out}} = \{3\} \not\subseteq \mathcal{N}^1_{\text{out}+1} = \{1, 2, 4, 5\}$. To show this, consider an *alternative* implementation of algorithm (3) with initial conditions and admissible perturbation signals

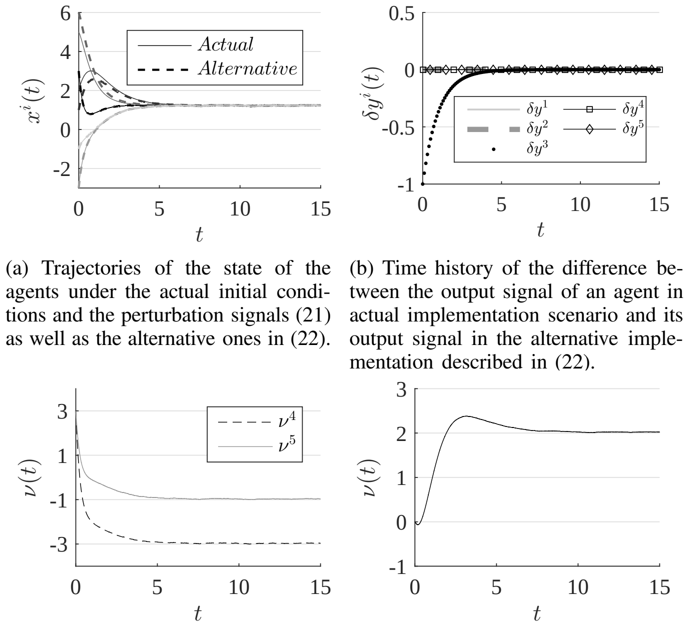
$$\begin{aligned} &x^{1'}(0)\!=\!3,\ x^{2'}(0)\!=\!1,\ x^{3'}(0)\!=\!6,\ x^{4'}(0)\!=\!-3,\ x^{5'}(0)\!=\!-1,\\ &f^{i'}(t)\!=\!f^i(t), \qquad\qquad g^{i'}(t)\!=\!g^i(t), \qquad\quad i \in \{1, 3, 4, 5\},\\ &f^{2'}(t)\!=\!f^2(t) - \mathsf{e}^{-t},\quad g^{2'}(t)\!=\!g^2(t) + \mathsf{e}^{-t}, \qquad\qquad\quad (22)\end{aligned}$$

where $\frac{1}{5}\sum_{i=1}^{5} x^{i'}(0) = \frac{1}{5}\sum_{i=1}^{5} x^i(0) = \frac{1}{5}\sum_{i=1}^{8} \mathsf{r}^i = 1.2$. As Fig. 6(a) shows the execution of algorithm (3) using the initial conditions and perturbation signals (21) (the actual case) and those in (22) (an alternative case) converge to the same final value of 1.2. Let $\delta y^i = y^i - y^{i'}$, $i \in \{1, \ldots, 5\}$ be the error between the output of the agents in the actual and the alternative cases. As Fig. 6(b) shows $\delta y^i \equiv 0$ for all $i \in \mathcal{N}^1_{\text{out}} = \{2, 4, 5\}$. This means that agent 1 cannot distinguish between the actual and the alternative cases and therefore, fails to identify uniquely the initial values of agent 2 and also agent 3. Figure 6(d) shows that an external malicious agent that has access to the output signals of agents 2 and 3 can employ an observer of the form (12) to identify the initial value of agent 2, i.e., $\mathsf{r}^2 = 2$.

VI. Conclusions

In this paper, we considered the problem of preserving the privacy of the reference value of the agents in an average consensus algorithm using additive perturbation signals. We started our study by characterizing the set of the necessary and sufficient conditions on the admissible perturbation signals, which do not perturb the final convergence point of the algorithm.

We assessed the privacy preservation property of the average consensus algorithm with the additive perturbation signals against internal and external malicious agents, depending on how much knowledge the malicious agents have about the necessary conditions that specify the class of the signals agents choose their local admissible perturbation signals from. We showed that if the necessary conditions are fully known to the malicious agents, then a malicious internal or external agent that have access to all the transmitted input and out signals of an agent can employ an asymptotic observer to obtain the reference value of that agent. Next, we showed that indeed having access to all the transmitted input and out signals of an agent at all $t \in \mathbb{R}_{\geq 0}$ is the necessary and sufficient condition for a malicious agent to identify the initial value of that particular agent. On the other hand, we showed that if the necessary conditions defining the locally chosen admissible perturbation signals are not fully known to the malicious agents, then the malicious agents cannot reconstruct the reference value of any other agent in the network.



(a) Trajectories of the state of the agents under the actual initial conditions and the perturbation signals (21) as well as the alternative ones in (22).

(b) Time history of the difference between the output signal of an agent in actual implementation scenario and its output signal in the alternative implementation described in (22).

(c) Time history of the observers of the form (10) that agent 1 with knowledge set (7) uses to obtain r^4 and r^5.

(d) Time history of the observer (12) of an external malicious agent with knowledge set (7) that wants to obtain r^2 and has direct access to y^2 and y^3 for all $t \in \mathbb{R}_{\geq 0}$.

Fig. 6: Simulation results when agents implement the modified average consensus algorithm (3) over the network in Fig. 5

Our problem of interest, identifying the initial condition of the agents in the presence of unknown additive perturbation signals, appears to be related to the concept of strong observability/detectability [26], [27] in control theory. However, our work is different from these classical results because a malicious agent has extra information given in the form of the necessary conditions on the unknown admissible perturbation signals, which it can exploit to reconstruct the initial condition of the other agents. Our future work includes extending our results to other multi-agent distributed algorithms such as dynamic average consensus and distributed optimization algorithms.

References

[1] R. Olfati-Saber and R. M. Murray, “Consensus problems in networks of agents with switching topology and time-delays,” *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1520–1533, 2004.

[2] W. Reb and R. W. Beard, “Consensus seeking in multi-agent systems under dynamically changing interaction topologies,” *IEEE Transactions on Automatic Control*, vol. 50, no. 5, pp. 655–661, 2005.

[3] L. Xiao and S. Boyd, “Fast linear iterations for distributed averaging,” *Systems and Control Letters*, vol. 53, pp. 65–78, 2004.

[4] R. Olfati-Saber, J. A. Fax, and R. M. Murray, “Consensus and cooperation in networked multi-agent systems,” *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, 2007.

[5] M. Kefayati, M. S. Talebi, B. H. Khalaj, and H. R. Rabiee, “Secure consensus averaging in sensor networks using random offsets,” in *IEEE International Conference on Telecommunications*, pp. 556–560, 2007.

[6] E. Nozari, P. Tallapragada, and J. Cortés, “Differentially private average consensus: obstructions, trade-offs, and optimal algorithm design,” *Automatica*, vol. 81, pp. 221–231, 2017.

[7] F. McSherry and K. Talwar, “Mechanism design via differential privacy,” in *IEEE Symposium on Foundations of Computer Science, 48th Annual*, pp. 94–103, 2007.

[8] A. Friedman and A. Schuster, “Data mining with differential privacy,” in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 493–502, 2010.

[9] C. Dwork, “Differential privacy: A survey of results,” in *International Conference on Theory and Applications of Models of Computation*, pp. 1–19, Springer, 2008.

[10] C. Dwork, A. Roth, *et al.*, “The algorithmic foundations of differential privacy,” *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[11] N. E. Manitara and C. N. Hadjicostis, “Privacy-preserving asymptotic average consensus,” in *European Control Conference*, pp. 760–765, 2013.

[12] Y. Mo and R. M. Murray, “Privacy preserving average consensus,” *IEEE Transactions on Automatic Control*, vol. 62, no. 2, pp. 753–765, 2017.

[13] Z. Huang, S. Mitra, and N. Vaidya, “Differentially private distributed optimization,” in *Proceedings of the 2015 International Conference on Distributed Computing and Networking*, p. 4, ACM, 2015.

[14] J. Le Ny and G. J. Pappas, “Differentially private Kalman filtering,” in *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, pp. 1618–1625, IEEE, 2012.

[15] J. L. Ny and G. J. Pappas, “Differential private filtering,” *IEEE Transactions on Automatic Control*, vol. 59, no. 2, pp. 341–354, 2014.

[16] J. Cortes, G. E. Dullerud, S. Han, J. L. Ny, S. Mitra, and G. J. Pappas, “Differential privacy in control and network systems,” in *IEEE Int. Conf. on Decision and Control*, pp. 4252–4272, 2016.

[17] M. Ruan, M. Ahmad, and Y. Wang, “Secure and privacy-preserving average consensus,” in *ACM Proceedings of the 2017 Workshop on Cyber-Physical Systems Security and Privacy*, pp. 123–129, 2017.

[18] E. D. Sontag, “Input to state stability: Basic concepts and results,” in *Nonlinear and Optimal Control Theory*, pp. 163–220, Springer, 2006.

[19] S. N. Dashkovskiy, D. V. Efimov, and E. D. Sontag, “Input to state stability and allied system properties,” *Automation and Remote Control*, vol. 72, no. 8, pp. 1579–1614, 2011.

[20] N. Rezazadeh and S. S. Kia, “Privacy preservation in a continuous-time static average consensus algorithm over directed graphs,” in *American Control Conference*, 2018. to appear.

[21] F. Bullo, J. Cortés, and S. Martínez, *Distributed Control of Robotic Networks*. Applied Mathematics Series, Princeton University Press, 2009.

[22] R. Hermann and A. J. Krener, “Nonlinear controllability and observability,” *IEEE Transactions on Automatic Control*, no. 5, pp. 728–740, 1977.

[23] E. D. Sontag, *Mathematical control theory: deterministic finite dimensional systems*. Springer Science & Business Media, 2013.

[24] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*. MIT Press, 3 ed., 2009.

[25] R. C.R. and R. Wilson, *An atlas of graphs*. Oxford University Press, 2005.

[26] M. Hou and R. J. Patton, “Input observability and input reconstruction,” *Automatica*, vol. 34, no. 6, pp. 789–794, 1998.

[27] M. L. J. Hautus, “Strong detectability and observers,” *Linear Algebra and its Applications*, vol. 50, pp. 353–368, 1983.

VII. Appendix

To provide proofs for our lemmas and theorems we rely on a set of auxiliary results, which we state first.

Lemma 7.1 (Auxiliary result 1): *Let* $\mathbf{L}$ *be the Laplacian matrix of a strongly connected and weight-balanced digraph. Recall* $\mathbf{L}^+=\mathbf{R}^\top\mathbf{L}\mathbf{R}$ *from* (1). *Let* $\mathbf{g}(t)=[g_1(t),...,g_n(t)]^\top\in\mathcal{L}_n^\infty$. *Then,*

$$\lim_{t\to\infty}\int_0^t \mathrm{e}^{-\mathbf{L}^+(t-\tau)}\mathbf{R}^\top\mathbf{L}\,\mathbf{g}(\tau)\mathrm{d}\tau=\mathbf{0},\tag{23}$$

is guaranteed to hold if and only if

$$\lim_{t\to\infty}\int_0^t \mathrm{e}^{-(t-\tau)}g^i(\tau)\,\mathrm{d}\tau=\alpha\in\mathbb{R},\quad i\in\{1,\dots,N\}.\tag{24}$$

Proof 6: Let

$$\dot{\boldsymbol{\zeta}}=-\mathbf{L}^+\boldsymbol{\zeta}+\mathbf{R}^\top\mathbf{L}\mathbf{g}(t),\qquad \boldsymbol{\zeta}(0)\in\mathbb{R}^{N-1},\tag{25}$$

$$\dot{\boldsymbol{\eta}}=-\boldsymbol{\eta}+\mathbf{R}^\top\mathbf{L}\mathbf{g}(t),\qquad \boldsymbol{\eta}(0)\in\mathbb{R}^{N-1}.\tag{26}$$

The trajectories $t\mapsto\boldsymbol{\zeta}$ and $t\mapsto\boldsymbol{\eta}$ of these two dynamics for $t\in\mathbb{R}_{\geq0}$ are given by

$$\boldsymbol{\zeta}(t)=\mathrm{e}^{-\mathbf{L}^+t}\boldsymbol{\zeta}(0)+\int_0^t\mathrm{e}^{-\mathbf{L}^+(t-\tau)}\mathbf{R}^\top\mathbf{L}\mathbf{g}(\tau)\mathrm{d}\tau,\tag{27}$$

$$\boldsymbol{\eta}(t)=\mathrm{e}^{-t}\boldsymbol{\eta}(0)+\mathbf{R}^\top\mathbf{L}\int_0^t\mathrm{e}^{-(t-\tau)}\,\mathbf{g}(\tau)\mathrm{d}\tau.\tag{28}$$

Let $\mathbf{e}=\boldsymbol{\zeta}-\boldsymbol{\eta}$. Then, the error dynamics between (25) and (26) is given by

$$\dot{\mathbf{e}}=-\mathbf{e}+(\mathbf{I}-\mathbf{L}^+)\boldsymbol{\zeta}.\tag{29}$$

or equivalently

$$\dot{\mathbf{e}}=-\mathbf{L}^+\mathbf{e}+(\mathbf{L}^++\mathbf{I})\boldsymbol{\eta}.\tag{30}$$

Let (23) hold. Since $-\mathbf{L}^+$ is a Hurwitz matrix, we have $\lim_{t\to\infty}\boldsymbol{\zeta}(t)=0$. Moreover, since $\mathbf{g}$ is essentially bounded, the trajectories of $\boldsymbol{\zeta}$ are guaranteed to be bounded. Therefore, considering error dynamics (29), by invoking the ISS stability results [19], we have the guarantees that $\lim_{t\to\infty}\mathbf{e}(t)=\mathbf{0}$, and consequently $\lim_{t\to\infty}\boldsymbol{\eta}(t)=\mathbf{0}$. As such, from (28) we obtain

$$\mathbf{R}^\top\mathbf{L}\lim_{t\to\infty}\int_0^t\mathrm{e}^{-(t-\tau)}\mathbf{g}(\tau)\mathrm{d}\tau=\mathbf{0}.\tag{31}$$

The nullspace of $\mathbf{R}^\top\mathbf{L}\in\mathbb{R}^{(N-1)\times N}$ is spanned by $\mathbf{1}_N$, therefore,

$$\lim_{t\to\infty}\int_0^t\mathrm{e}^{-(t-\tau)}\mathbf{g}(\tau)\mathrm{d}\tau=\alpha\mathbf{1}_N,\quad\alpha\in\mathbb{R},$$

which validates (24). Now let (24) hold. Then, using (28), we obtain $\lim_{t\to\infty}\boldsymbol{\eta}(t)=\mathbf{0}$. Since $\mathbf{g}$ is essentially bounded, the trajectories of $\boldsymbol{\zeta}$ are guaranteed to be bounded. Thereby, considering error dynamics (30), by invoking the ISS stability results [19], we have the guarantees that $\lim_{t\to\infty}\mathbf{e}(t)=\mathbf{0}$, and consequently $\lim_{t\to\infty}\boldsymbol{\eta}(t)=\mathbf{0}$. Since $-\mathbf{L}^+$ is a Hurwitz matrix, we obtain (23) from (27).

Lemma 7.2 (Auxiliary result 2): *Let* $\mathbf{u}:\mathbb{R}_{\geq0}\to\mathbb{R}^n$ *be an essentially bounded signal and* $\mathbf{E}\in\mathbb{R}^{n\times n}$ *be a Hurwitz matrix.*

(a) If $\lim_{t\to\infty}\mathbf{u}(t)=\bar{\mathbf{u}}\in\mathbb{R}^n$, *and* $\mathbf{E}\in\mathbb{R}^{n\times n}$, *then*

$$\lim_{t\to\infty}\int_0^t\mathrm{e}^{\mathbf{E}\,(t-\tau)}\mathbf{u}(\tau)\mathrm{d}\tau=-\mathbf{E}^{-1}\,\bar{\mathbf{u}}.\tag{32}$$

(b) If $\lim_{t\to\infty}\int_0^t\mathbf{u}(\tau)\mathrm{d}\tau=\bar{\mathbf{u}}\in\mathbb{R}^n$, *then*

$$\lim_{t\to\infty}\int_0^t\mathrm{e}^{\mathbf{E}\,(t-\tau)}\mathbf{u}(\tau)\mathrm{d}\tau=\mathbf{0}.\tag{33}$$

Proof 7: To prove statement (a) we proceed as follows. Let $\boldsymbol{\mu}(t) = \mathbf{u}(t) - \bar{\mathbf{u}}$. Next, consider $\dot{\boldsymbol{\zeta}} = \mathbf{E}\,\boldsymbol{\zeta} + \boldsymbol{\mu}$, $\boldsymbol{\zeta}(0) \in \mathbb{R}^n$, which gives $\boldsymbol{\zeta}(t) = \mathrm{e}^{\mathbf{E}\,t}\boldsymbol{\zeta}(0) + \int_0^t \mathrm{e}^{\mathbf{E}(t-\tau)}\boldsymbol{\mu}(\tau)\mathrm{d}\tau$, $t \geq 0$. Since $\mathbf{E}$ is Hurwitz and $\boldsymbol{\mu}$ is an essentially bounded and vanishing signal, by virtue of the ISS results for linear systems [19] we have $\lim_{t\to\infty}\boldsymbol{\zeta}(t) = 0$. Consequently, $\lim_{t\to\infty}\int_0^t \mathrm{e}^{\mathbf{E}\,(t-\tau)}\boldsymbol{\mu}(\tau)\mathrm{d}\tau = \mathbf{0}$, which guarantees (32).

To prove statement (b) we proceed as follows. Consider

$$\dot{\boldsymbol{\zeta}} = \mathbf{u},\ \dot{\boldsymbol{\eta}} = \mathbf{E}\boldsymbol{\eta} + \mathbf{u}, \quad \boldsymbol{\zeta}(0) = \mathbf{0},\ \boldsymbol{\eta}(0) \in \mathbb{R}^n,$$

which result in $\boldsymbol{\zeta}(t) = \int_0^t \mathbf{u}(\tau)\mathrm{d}\tau$ and

$$\boldsymbol{\eta}(t) = \mathrm{e}^{\mathbf{E}\,t}\boldsymbol{\eta}(0) + \int_0^t \mathrm{e}^{\mathbf{E}\,(t-\tau)}\mathbf{u}(\tau)\mathrm{d}\tau. \tag{34}$$

Given the conditions on $\mathbf{u}$ both $\boldsymbol{\zeta}$ and $\boldsymbol{\eta}$ are essentially bounded signals (recall that $\mathbf{E}$ is Hurwitz). Let $\mathbf{e} = \boldsymbol{\eta} - \boldsymbol{\zeta}$. Therefore, we can write

$$\dot{\mathbf{e}} = \mathbf{E}\,\mathbf{e} + \mathbf{E}\,\boldsymbol{\zeta}, \quad \mathbf{e}(0) = \boldsymbol{\eta}(0) \in \mathbb{R}^n.$$

Since $\boldsymbol{\zeta}$ is essentially bounded and satisfies $\lim_{t\to\infty}\mathbf{E}\boldsymbol{\zeta}(t) = \mathbf{E}\bar{\mathbf{u}}$, with an argument similar to that of the proof of statement (a), we can conclude that $\lim_{t\to\infty}\mathbf{e}(t) = -\bar{\mathbf{u}}$. As a result $\lim_{t\to\infty}\boldsymbol{\eta}(t) = \mathbf{0}$. Consequently, from (34), we obtain (33).

Lemma 7.3 (Auxiliary result 3): *Let $\mathcal{G}$ be a strongly connected and weight-balanced digraph. Then, every island of any agent i, is strongly connected and weight-balanced.*

Proof 8: Without loss of generality, we prove our argument by showing that the island $\mathcal{G}_1^{\bar{1}}$ of agent 1 is strongly connected and weight-balanced. By construction, we know that there is a directed path from every agent to every other agent in $\mathcal{G}_1^{\bar{1}}$, therefore, $\mathcal{G}_1^{\bar{1}}$ is strongly connected. Next we show that $\mathcal{G}_1^{\bar{1}}$ is weight-balanced. Let $\mathcal{V}_2 = \mathcal{V}_1^{\bar{1}}\backslash\{1\}$ and $\mathcal{V}_3 = \mathcal{V}\backslash\mathcal{V}_2$.Let the nodes of $\mathcal{G}$ be labeled in accordance to $(1, \mathcal{V}_2, \mathcal{V}_3)$, respectively, and partition the graph Laplacian L accordingly as

$$\mathsf{L} = \begin{bmatrix} \mathsf{d}_{\text{out}}^1 & -\mathsf{A}_{12} & -\mathsf{A}_{13} \\ -\mathsf{A}_{21} & \mathsf{L}_{22} & \mathbf{0} \\ -\mathsf{A}_{31} & \mathbf{0} & \mathsf{L}_{33} \end{bmatrix}.$$

Since $\mathcal{G}$ is strongly connected and weight-balanced, we have $\mathsf{L}\mathbf{1}_N = \mathbf{0}$ and $\mathbf{1}_N^\top\mathsf{L} = \mathbf{0}$, which guarantee that

$$\mathbf{1}_{|\mathcal{V}_1^{\bar{1}}|}^\top \begin{bmatrix} -\mathsf{A}_{12} \\ \mathsf{L}_{22} \end{bmatrix} = \mathbf{0}, \qquad \begin{bmatrix} -\mathsf{A}_{21} & \mathsf{L}_{22} \end{bmatrix} \mathbf{1}_{|\mathcal{V}_1^{\bar{1}}|} = \mathbf{0}. \tag{35}$$

Therefore,

$$\mathbf{1}_{|\mathcal{V}_1^{\bar{1}}|}^\top \begin{bmatrix} -\mathsf{A}_{12} \\ \mathsf{L}_{22} \end{bmatrix} \mathbf{1}_{|\mathcal{V}_1^{\bar{1}}|} = 0, \qquad \mathbf{1}_{|\mathcal{V}_1^{\bar{1}}|}^\top \begin{bmatrix} -\mathsf{A}_{21} & \mathsf{L}_{22} \end{bmatrix} \mathbf{1}_{|\mathcal{V}_1^{\bar{1}}|} = 0,$$

which we can use to conclude that $\mathtt{sum}(\mathsf{A}_{12}^\top) = \mathtt{sum}(\mathsf{A}_{21})$.

Let the Laplacian matrix of $\mathcal{G}_1^{\bar{1}}$ be $\mathsf{L}_1^{\bar{1}}$. Partitioning this matrix according to order node set $(1, \mathcal{V}_2)$, we obtain

$$\mathsf{L}_1^{\bar{1}} = \begin{bmatrix} \mathsf{d}_{\text{out}}^{1,1} & -\mathsf{A}_{12} \\ -\mathsf{A}_{21} & \mathsf{L}_{22} \end{bmatrix},$$

where $\mathsf{d}_{\text{out}}^{1,1} = \sum_{j\in\mathcal{V}_2}\mathsf{a}_{1j} = \mathtt{sum}(\mathsf{A}_{12}^\top)$. To establish $\mathcal{G}_1^{\bar{1}}$ is weight-balanced digraph, we show next that $\mathbf{1}_{|\mathcal{V}_1^{\bar{1}}|}^\top\mathsf{L}_1^{\bar{1}} = \mathbf{0}$. From $\mathbf{1}_N^\top\mathsf{L} = \mathbf{0}$, it follows that $\mathbf{1}_{|\mathcal{V}_1^{\bar{1}}|}^\top \begin{bmatrix} -\mathsf{A}_{12} \\ \mathsf{L}_{22} \end{bmatrix} = \mathbf{0}$. Therefore, to prove $\mathcal{G}_1^{\bar{1}}$ is weight-balanced, we need to show that $\mathsf{d}_{\text{out}}^{1,1} + \mathtt{sum}(-\mathsf{A}_{21}) = 0$, which follows immediately from $\mathsf{d}_{\text{out}}^{1,1} = \mathtt{sum}(\mathsf{A}_{12}^\top)$ and $\mathtt{sum}(\mathsf{A}_{12}^\top) = \mathtt{sum}(\mathsf{A}_{21})$.

Next we present the proof of our main results.

Proof 9 (Proof of Theorem 3.1): *To prove necessity, we proceed as follows. We write the algorithm* (3) *in compact form*

$$\dot{\mathbf{x}} = -\mathsf{L}\,\mathbf{x} - \mathsf{L}\,\mathbf{g} + \mathbf{f} + \mathsf{D}^{\text{out}}\,\mathbf{g} = -\mathsf{L}\,\mathbf{x} + \mathbf{f} + \mathbf{A}\,\mathbf{g}. \tag{36}$$

Left multiplying both sides of (36) *by* $\mathbf{1}_N^\top$ *gives*

$$\sum\nolimits_{j=1}^{N}\dot{x}^j(t) = \sum\nolimits_{j=1}^{N}(f^i(t) + \mathsf{d}_{\text{out}}^i\,g^i(t)),$$

which results in

$$\sum\nolimits_{j=1}^{N}x^j(t) = \sum\nolimits_{j=1}^{N}x^j(0) + \int_0^t\sum\nolimits_{j=1}^{N}(f^i(\tau) + \mathsf{d}_{\text{out}}^i\,g^i(\tau))\,\mathrm{d}\tau.$$

Because $x^i(0) = \mathsf{r}^i$, *to ensure* $\lim_{t\to\infty}x^i(t) = \frac{1}{N}\sum_{j=1}^N\mathsf{r}^j$, $i \in \mathcal{V}$, *we necessarily need* (4b).

Next, we apply the change of variable

$$\mathbf{p} = \begin{bmatrix} p_1 \\ \mathbf{p}_{2:N} \end{bmatrix} = \mathbf{T}\,\mathbf{x}, \tag{37}$$

where $\mathbf{T}$ *is defined in* (1), *to write* (36) *in the equivalent form*

$$\dot{p}_1 = \frac{1}{\sqrt{N}}\sum\nolimits_{i=1}^{N}(f^i + \mathsf{d}_{\text{out}}^i\,g^i), \tag{38a}$$

$$\dot{\mathbf{p}}_{2:N} = -\mathsf{L}^+\,\mathbf{p}_{2:N} + \mathbf{R}^\top(\mathbf{f} + \mathbf{A}\,\mathbf{g}). \tag{38b}$$

The solution of (38) *is*

$$p_1(t) = \frac{1}{\sqrt{N}}\sum\nolimits_{i=1}^{N}x^i(0) + \tag{39a}$$

$$\frac{1}{\sqrt{N}}\int_0^t\sum\nolimits_{i=1}^{N}(f^i(\tau) + \mathsf{d}_{\text{out}}^i\,g^i(\tau))\mathrm{d}\tau,$$

$$\mathbf{p}_{2:N}(t) = \mathrm{e}^{-\mathsf{L}^+\,t}\,\mathbf{p}_{2:N}(0) +$$

$$\int_0^t \mathrm{e}^{-\mathsf{L}^+(t-\tau)}\mathbf{R}^\top(\mathbf{f}(\tau) + \mathbf{A}\,\mathbf{g}(\tau))\,\mathrm{d}\tau. \tag{39b}$$

Given (4a), (39a) *results in* $\lim_{t\to\infty}p_1(t) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}x^i(0) = \frac{1}{N}\sum_{j=1}^{N}\mathsf{r}^j$. *Consequently, given* (37), *to ensure* $\lim_{t\to\infty}x^i(t) = \frac{1}{N}\sum_{j=1}^N\mathsf{r}^j$, $i \in \mathcal{V}$, *we need*

$$\lim_{t\to\infty}\mathbf{p}_{2:N}(t) = \mathbf{0}. \tag{40}$$

Because for a strongly connected and weight-balanced digraph, $-\mathsf{L}^+$ *is a Hurwitz matrix,* $\lim_{t\to\infty}\mathrm{e}^{-\mathsf{L}^+\,t}\mathbf{p}_{2:N}(0) = \mathbf{0}$. *Then, the necessary condition for* (40) *is* (4b).

The sufficiency proof follows from noting that under (4), *the trajectories of* (39) *satisfy* $\lim_{t\to\infty}p_1(t) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}x^i(0)$ *and* $\lim_{t\to\infty}\mathbf{p}_{2:N}(t) = \mathbf{0}$. *Then, given* (37) *and* $x^i(0) = \mathsf{r}^i$ *we obtain* $\lim_{t\to\infty}x^i(t) = \frac{1}{N}\sum_{j=1}^{N}\mathsf{r}^j$, $i \in \mathcal{V}$.

Proof 10 (Proof of Theorem 3.2): *Given* (5), *it is straightforward to see that* (6a) *is necessary and sufficient for* (4a). *Next, we observe that using* (5), *we can write* $\lim_{t\to\infty}\int_0^t \mathbf{R}^\top(\mathbf{f}(\tau)+\mathbf{D}^{\text{out}}\mathbf{g}(\tau))\mathrm{d}\tau = \mathbf{R}^\top\begin{bmatrix}\beta^1 & \cdots & \beta^N\end{bmatrix}^\top$. *Then, it follows from the statement (b) of Lemma 7.2 that* $\lim_{t\to\infty}\int_0^t \mathrm{e}^{-\mathbf{L}^+(t-\tau)}\mathbf{R}^\top(\mathbf{f}(\tau)+\mathbf{D}^{\text{out}}\mathbf{g}(\tau))\mathrm{d}\tau = \mathbf{0}$. *As a result, given* $\mathbf{f}+\mathbf{A}\,\mathbf{g}=\mathbf{f}+\mathbf{D}^{\text{out}}\mathbf{g}-\mathbf{L}\,\mathbf{g}$, *we obtain*

$$\lim_{t\to\infty}\int_0^t \mathrm{e}^{-\mathbf{L}^+(t-\tau)}\mathbf{R}^\top(\mathbf{f}(\tau)+\mathbf{A}\,\mathbf{g}(\tau))\,\mathrm{d}\tau = \\ -\lim_{t\to\infty}\int_0^t \mathrm{e}^{-\mathbf{L}^+(t-\tau)}\mathbf{R}^\top\mathbf{L}\,\mathbf{g}(\tau)\mathrm{d}\tau. \tag{41}$$

Given (41), *by virtue of Lemma 7.1,* (4b) *holds if and only if* (6b) *holds.*

Proof 11 (Proof of Lemma 3.1): *When condition (a) holds, the proof of the statement follows from statement (a) of Lemma 7.2. When condition (b) is satisfied, the proof follows from the statements (a) and (b) of Lemma 7.2 which, respectively, give* $\lim_{t\to\infty}\int_0^t \mathrm{e}^{-(t-\tau)}g_1(\tau)\mathrm{d}\tau=\alpha$ *and* $\lim_{t\to\infty}\int_0^t \mathrm{e}^{-(t-\tau)}g_2(\tau)\mathrm{d}\tau=0$. *When condition (c) is satisfied, the proof follows from the statement (a) of Lemma 7.2 which gives* $\lim_{t\to\infty}\int_0^t \mathrm{e}^{-(t-\tau)}g_1(\tau)\mathrm{d}\tau=\alpha$ *and noting that* $\int_0^t \mathrm{e}^{-(t-\tau)}g_2(\tau)\mathrm{d}\tau$ *is the zero state response of system* $\dot{\zeta}=-\zeta+g_2$. *Since* $g_2(t)$ *is essentially bounded, this system is ISS, and as a result it is also integral ISS [19]. Then,* $\int_0^t \mathrm{e}^{-(t-\tau)}g_2(\tau)\mathrm{d}\tau=0$, *follows from [19, Lemma 3.1].*

Proof 12 (Proof of Lemma 4.1): *Let the error variables of the two execution of* (3) *described in the statement be* $\delta x^i(t)=x^{i'}(t)-x^i(t)$, $\delta y^i(t)=y^{i'}(t)-y^i(t)$, $\delta g^i(t)=g^{i'}(t)-g^i(t)$, *and* $\delta f^i(t)=f^{i'}(t)-f^i(t)$, $i\in\mathcal{V}$. *Consequently,*

$$\begin{aligned}
&\delta x^1(0)=0,\quad \delta\mathbf{x}_4=\mathbf{0},\quad \delta\mathbf{x}_5(0)=\mathbf{0}, && \text{(42a)}\\
&\delta x^i(0)\in\mathbb{R}, \qquad\qquad i\in(\mathcal{V}_{\bar{1},2}^{1}\cup\mathcal{V}_{\bar{1},3}^{1}), && \text{(42b)}\\
&\delta\mathbf{x}_2(0)=-\mathbf{A}_{23}\mathbf{L}_{33}^{-1}\delta\mathbf{x}_3(0), && \text{(42c)}
\end{aligned}$$

and

$$\begin{aligned}
&\delta g^1(t)\equiv 0,\quad \delta f^1(t)\equiv 0, && \text{(43a)}\\
&\delta\mathbf{g}_l(t)\equiv\mathbf{0},\quad \delta\mathbf{f}_l(t)\equiv\mathbf{0},\qquad l\in\{3,4,5\}, && \text{(43b)}\\
&\delta\mathbf{g}_2(t)=-\mathrm{e}^{-\mathbf{D}_{22}^{\text{out}}t}\delta\mathbf{x}_2(0),\ \ \delta\mathbf{f}_2(t)=-\mathbf{A}_{23}\mathrm{e}^{-\mathbf{L}_{33}t}\delta\mathbf{x}_3(0). && \text{(43c)}
\end{aligned}$$

Given the inter-agent interactions across the network based on agent grouping in accordance to the definition of the island $\mathcal{G}_{\bar{1}}^{1}$ *(see Fig. 3), the error dynamics pertained to the modified static average consensus algorithm* (3) *reads as*

$$\begin{bmatrix}\delta\dot{x}^1\\ \delta\dot{\mathbf{x}}_2\\ \delta\dot{\mathbf{x}}_3\\ \delta\dot{\mathbf{x}}_4\\ \delta\dot{\mathbf{x}}_5\end{bmatrix} = -\underbrace{\begin{bmatrix}\mathsf{d}_{\text{out}}^1 & -\mathbf{A}_{12} & \mathbf{0} & -\mathbf{A}_{14} & -\mathbf{A}_{15}\\ -\mathbf{A}_{21} & \mathbf{L}_{22} & -\mathbf{A}_{23} & -\mathbf{A}_{24} & \mathbf{0}\\ -\mathbf{A}_{31} & -\mathbf{A}_{32} & \mathbf{L}_{33} & -\mathbf{A}_{34} & \mathbf{0}\\ -\mathbf{A}_{41} & -\mathbf{A}_{42} & \mathbf{0} & \mathbf{L}_{44} & \mathbf{0}\\ -\mathbf{A}_{51} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{L}_{55}\end{bmatrix}}_{\mathbf{L}}\begin{bmatrix}\delta x^1\\ \delta\mathbf{x}_2\\ \delta\mathbf{x}_3\\ \delta\mathbf{x}_4\\ \delta\mathbf{x}_5\end{bmatrix} \\ +\underbrace{\begin{bmatrix}0 & \mathbf{A}_{12} & \mathbf{0} & \mathbf{A}_{14} & \mathbf{A}_{15}\\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{A}_{23} & \mathbf{A}_{24} & \mathbf{0}\\ \mathbf{A}_{31} & \mathbf{A}_{32} & \mathbf{A}_{33} & \mathbf{A}_{34} & \mathbf{0}\\ \mathbf{A}_{41} & \mathbf{A}_{42} & \mathbf{0} & \mathbf{A}_{44} & \mathbf{0}\\ \mathbf{A}_{51} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A}_{55}\end{bmatrix}}_{\mathbf{A}}\begin{bmatrix}\delta g^1\\ \delta\mathbf{g}_2\\ \delta\mathbf{g}_3\\ \delta\mathbf{g}_4\\ \delta\mathbf{g}_5\end{bmatrix}+\begin{bmatrix}\delta f^1\\ \delta\mathbf{f}_2\\ \delta\mathbf{f}_3\\ \delta\mathbf{f}_4\\ \delta\mathbf{f}_5\end{bmatrix}. \tag{44}$$

Since for a strongly connected and weight-balanced digraph we have $\operatorname{rank}(\mathbf{L})=N-1$ *and* $-(\mathbf{L}+\mathbf{L}^\top)\leq 0$, *the sub-block matrices* $-\mathbf{L}_{33}$ *and* $-\mathbf{L}_{44}$ *and* $-\mathbf{L}_{55}$ *satisfy* $-(\mathbf{L}_{ii}+\mathbf{L}_{ii}^\top)<0$, $i\in\{1,\dots,5\}$. *Thereby, they are invertible and Hurwitz matrices.*

To establish (18), *we show* $\mathbf{1}_N^\top\delta\mathbf{x}(0)=\mathbf{0}_N$. *For this, note that taking into account* (42), *we can write*

$$\delta\mathbf{x}(0)=\underbrace{\begin{bmatrix}0 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & -\mathbf{A}_{23} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{L}_{33} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0}\end{bmatrix}}_{\mathbf{B}}\begin{bmatrix}0\\ \mathbf{L}_{33}^{-1}\delta\mathbf{x}_3(0)\\ \mathbf{L}_{33}^{-1}\delta\mathbf{x}_3(0)\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix} \tag{45}$$

Comparing $\mathbf{B}$ *with the block partitioned* $\mathbf{L}$ *in* (44), *it is evident that* $\mathbf{1}^\top\mathbf{B}=\mathbf{0}$ *follows from* $\mathbf{1}^\top\mathbf{L}=\mathbf{0}$. *Consequently, we can deduce from* (45) *that* $\mathbf{1}^\top\delta\mathbf{x}(0)=0$. *Next, given* (18), *we validate* (19) *by invoking Theorem 3.2 and showing that the perturbation signals* $(f^{i'},g^{i'})$, $i\in\mathcal{V}$, *satisfy the sufficient conditions in* (6). *For* $i\in\mathcal{V}\backslash\mathcal{V}_{\bar{1},2}^{1}$, *the sufficient conditions in* (6) *are trivially satisfied. To show* (6a) *for* $i\in\mathcal{V}_{\bar{1},2}^{1}$, *we proceed as follows. First note that since* (f^i,g^i), $i\in\mathcal{V}_{\bar{1},2}^{1}$, *are admissible signals, they necessarily satisfy* (6a). *Next, note that using* (16) *we can write*

$$\int_0^t\big(-\mathbf{A}_{23}\mathrm{e}^{-\mathbf{L}_{33}\tau}\delta\mathbf{x}_3(0)+\mathbf{D}_{22}^{\text{out}}\mathrm{e}^{-\mathbf{D}_{22}^{\text{out}}\tau}\delta\mathbf{x}_2(0)\big)\mathrm{d}\tau = \\ \mathbf{A}_{23}\mathbf{L}_{33}^{-1}\mathrm{e}^{-\mathbf{L}_{33}t}\delta\mathbf{x}_3(0)-\mathrm{e}^{-\mathbf{D}_{22}^{\text{out}}\tau}\delta\mathbf{x}_2(0).$$

Let $\mathfrak{B}_2=[\{\beta^i\}_{i\in\mathcal{V}_{\bar{1},2}^{1}}]$. *Then, in light of the aforementioned observations and the fact that* $-\mathbf{L}_{33}$ *and* $-\mathbf{D}_{22}^{\text{out}}$ *are Hurwitz matrices we can write*

$$\lim_{t\to\infty}\int_0^t\big(\mathbf{f}_2'(\tau)+\mathbf{D}_{22}^{\text{out}}\mathbf{g}_2'(\tau)\big)\mathrm{d}\tau = \\ \mathfrak{B}_2+\lim_{t\to\infty}\big(\mathbf{A}_{23}\mathbf{L}_{33}^{-1}\mathrm{e}^{-\mathbf{L}_{33}t}\delta\mathbf{x}_3(0)-\mathrm{e}^{-\mathbf{D}_{22}^{\text{out}}\tau}\delta\mathbf{x}_2(0)\big)=\mathfrak{B}_2,$$

which shows $(f^{i'},g^{i'})$, $i\in\mathcal{V}_{\bar{1},2}^{1}$ *also satisfy the sufficient condition* (6a). *Establishing that* $g^{i'}$, $i\in\mathcal{V}_{\bar{1},2}^{1}$, *satisfies the sufficient condition* (6b) *follows from admissibility of* g^i,

$i \in \mathcal{V}_{1,2}^{1}$, which ensures it satisfies (6b)*, and direct calculations as show below,*

$$\lim_{t\to\infty}\int_0^t \mathrm{e}^{-(t-\tau)} g^{i'}(\tau)\,\mathrm{d}\tau = \alpha + \lim_{t\to\infty}\int_0^t \mathrm{e}^{-(t-\tau)}\mathrm{e}^{-\mathsf{d}_{\mathrm{out}}^i\tau}\delta x^i(0)\,\mathrm{d}\tau = \alpha.$$

Here we used the fact that for a strongly connected digraph we have $\mathsf{d}_{\mathrm{out}}^i \geq 1$.

To establish (20) *we proceed as follows. We assume that* (20) *or equivalently*

$$\delta y^1(t) = \delta x^1(t) + \delta g^1(t) \equiv \mathbf{0}, \qquad t \in \mathbb{R}_{\geq 0}, \tag{46a}$$

$$\delta \mathbf{y}_2(t) = \delta \mathbf{x}_2(t) + \delta \mathbf{g}_2(t) \equiv \mathbf{0}, \qquad t \in \mathbb{R}_{\geq 0}, \tag{46b}$$

$$\delta \mathbf{y}_4(t) = \delta \mathbf{x}_4(t) + \delta \mathbf{g}_4(t) \equiv \mathbf{0}, \qquad t \in \mathbb{R}_{\geq 0}, \tag{46c}$$

$$\delta \mathbf{y}_5(t) = \delta \mathbf{x}_5(t) + \delta \mathbf{g}_5(t) \equiv \mathbf{0}, \qquad t \in \mathbb{R}_{\geq 0}. \tag{46d}$$

hold. Then, for the given initial conditions (42)*, we identify the perturbation signals that make the error dynamics* (44) *render such an output. As we show below, these perturbation signals are exactly the same as* (43)*. Then, the proof is established by the fact that given a set of initial conditions and integrable external signals, the solution of any linear ordinary differential equation is unique. That is, if we implement the identified inputs, the error dynamics is guaranteed to satisfy* (46)*. If* (46) *holds, then the error dynamics* (44) *reads as*

$$\delta \dot{x}^1 = -\mathsf{d}_{\mathrm{out}}^1 \delta x^1 + \delta f^1, \tag{47a}$$

$$\delta \dot{\mathbf{x}}_2 = -\mathsf{D}_{22}^{\mathrm{out}}\delta \mathbf{x}_2 + \mathsf{A}_{23}\delta \mathbf{x}_3 + \mathsf{A}_{23}\delta \mathbf{g}_3 + \delta \mathbf{f}_2, \tag{47b}$$

$$\delta \dot{\mathbf{x}}_3 = -\mathsf{L}_{33}\delta \mathbf{x}_3 + \mathsf{A}_{33}\delta \mathbf{g}_3 + \delta \mathbf{f}_3, \tag{47c}$$

$$\delta \dot{\mathbf{x}}_4 = -\mathsf{D}_{44}^{\mathrm{out}}\delta \mathbf{x}_4 + \delta \mathbf{f}_4, \tag{47d}$$

$$\delta \dot{\mathbf{x}}_5 = -\mathsf{D}_{55}^{\mathrm{out}}\delta \mathbf{x}_5 + \delta \mathbf{f}_5, \tag{47e}$$

Here, we used $\mathsf{L}_{ii} = \mathsf{D}_{ii}^{\mathrm{out}} - \mathsf{A}_{ii}$, $i \in \{1,2,4,5\}$. Next, we choose the perturbation signals according to (43)*. Then, for the given initial conditions* (42)*, we obtain from* (47)*,*

$$\delta \dot{x}^1 = -\mathsf{d}_{\mathrm{out}}^1 \delta x^1, \;\Rightarrow\; \delta x^1(t) = 0 \;\Rightarrow\; \delta y^1(t) \equiv 0, \tag{48a}$$

$$\delta \dot{\mathbf{x}}_3 = -\mathsf{L}_{33}\,\delta \mathbf{x}_3, \;\Rightarrow\; \delta \mathbf{x}_3(t) = \mathrm{e}^{-\mathsf{L}_{33}t}\delta \mathbf{x}_3(0), \tag{48b}$$

$$\delta \dot{\mathbf{x}}_4 = -\mathsf{D}_{44}^{\mathrm{out}}\delta \mathbf{x}_4, \;\Rightarrow\; \delta \mathbf{x}_4(t) \equiv \mathbf{0}, \Rightarrow \delta \mathbf{y}_4(t) \equiv \mathbf{0}, \tag{48c}$$

$$\delta \dot{\mathbf{x}}_5 = -\mathsf{D}_{55}^{\mathrm{out}}\delta \mathbf{x}_5, \;\Rightarrow\; \delta \mathbf{x}_5(t) \equiv \mathbf{0}, \Rightarrow \delta \mathbf{y}_5(t) \equiv \mathbf{0}, \tag{48d}$$

for $t \in \mathbb{R}_{\geq 0}$. Substituting for $\mathbf{x}_3$ ans $\delta \mathbf{f}_2$ in (47b)*, we obtain*

$$\begin{aligned}\delta \dot{\mathbf{x}}_2 &= -\mathsf{D}_{22}^{\mathrm{out}}\delta \mathbf{x}_2 + \mathsf{A}_{23}\mathrm{e}^{-\mathsf{L}_{33}t}\delta \mathbf{x}_3(0) - \mathsf{A}_{23}\mathrm{e}^{-\mathsf{L}_{33}t}\delta \mathbf{x}_3(0)\\ &= -\mathsf{D}_{22}^{\mathrm{out}}\delta \mathbf{x}_2, \;\Rightarrow \delta \mathbf{x}_2(t) = \mathrm{e}^{-\mathsf{D}_{22}^{\mathrm{out}}t}\delta \mathbf{x}_2(0),\end{aligned} \tag{49}$$

for $t \in \mathbb{R}_{\geq 0}$. Finally using $\delta \mathbf{g}_2$ in (43c)*, we*

$$\begin{aligned}\delta \mathbf{y}_2(t) &= \delta \mathbf{x}_2 + \delta \mathbf{g}_2\\ &= \mathrm{e}^{-\mathsf{D}_{22}^{\mathrm{out}}t}\delta \mathbf{x}_2(0) - \mathrm{e}^{-\mathsf{D}_{22}^{\mathrm{out}}t}\delta \mathbf{x}_2(0) \equiv \mathbf{0},\end{aligned} \tag{50}$$

for $t \in \mathbb{R}_{\geq 0}$.

Proof 13 (Proof of Lemma 4.2): *If agent 1 knows β^i, the proof follows from Theorem 4.1. If agent 1 does not know β^i, since it knows* (6a)*, there exists at least one other agent $k \in \mathcal{V}\backslash\{1, i\}$ whose β^k is not known to agent 1. We note that at the best case, $\beta^i + \beta^k$ can be known to agent 1. Now consider $\beta_{ik} \in \mathbb{R}\backslash\{0\}$ and let $\beta^{i'} = \beta^i + \beta_{ik}$ and $\beta^{k\prime} = \beta^k - \beta_{ik}$, and $\beta^{l\prime} = \beta^l$ for $l \in \mathcal{V}\backslash\{i,k\}$. Now consider an alternative implementation of algorithm* (3a)-(3b) *with initial conditions $x^{l\prime}(0) = x^l(0)$ for $l \in \mathcal{V}\backslash\{i,k\}$, $x^{i\prime}(0) = x^i(0) - \beta_{ik}$ and $x^{k\prime}(0) = x^k(0) + \beta_{ik}$ and perturbation signals $f^{l\prime}(t) = f^l(t)$, $g^{l\prime}(t) = g^l(t)$ for $l \in \mathcal{V}\backslash\{i,k\}$, $f^{i'}(t) = f^i(t) + \mathrm{d}\,\beta_{ik}e^{-(d_{out}^i+d)t}$, $g^{i\prime}(t) = g^i(t) + \beta_{ik}e^{-(d_{out}^i+d)t}$ and $f^{k\prime}(t) = f^k(t) - \mathrm{d}\,\beta_{ik}e^{-(d_{out}^k+d)t}$, $g^{k\prime}(t) = g^k(t) - \beta_{ik}e^{-(d_{out}^k+d)t}$, where $\mathrm{d} \in \mathbb{R}$ is chosen such that $\mathrm{d} > \max\{d_{out}^i, d_{out}^k\}$. Let $t \mapsto x^{l\prime}(t)$ and $t \mapsto y^{l\prime}(t)$, $t \in \mathbb{R}_{\geq 0}$, respectively, be the state and the transmitted signal of agent $l \in \mathcal{V}$ in this alternative case. We note that using $\lim_{t\to\infty}\int_0^t d\beta_{ik}e^{-(d_{out}^i+d)\tau}d\tau = \frac{d\beta_{ik}}{d_{out}^i+d}$ and $\lim_{t\to\infty}\int_0^t d\beta_{ik}e^{-(d_{out}^i+d)\tau}d\tau = \frac{1}{d_{out}^i+d}$ we can show $\lim_{t\to\infty}\int_0^t (f^{l'}(\tau) + \mathsf{d}_{\mathrm{out}}^l\, g^{l'}(\tau))\,\mathrm{d}\tau = \beta^{l'}$, and $\lim_{t\to\infty}\int_0^t \mathrm{e}^{-(t-\tau)}g^{l\prime}(\tau)\mathrm{d}\tau = \alpha$ for $l \in \mathcal{V}$. Therefore, since $\sum_{l=j}^N \beta^{j\prime} = 0$, by virtue of Theorem 3.2 we get*

$$\lim_{t\to\infty} x^{l'}(t) = \frac{1}{N}\sum\nolimits_{j=1}^{N} x^{l'}(0) = \frac{1}{N}\sum\nolimits_{j=1}^{N}\mathsf{r}^l, \quad l \in \mathcal{V}. \tag{51}$$

Next, let $\delta x^l(t) = x^l(t) - x^{l\prime}(t)$ and $\delta y^l(t) = y^l(t) - y^{l\prime}(t)$, $l \in \mathcal{V}$. Then,

$$\begin{cases}\delta \dot{x}^l(t) = -\mathsf{d}_{\mathrm{out}}^l\delta x^l(t) + \sum\limits_{j=1}^N \mathsf{a}_{lj}\delta y^j(t), & l \in \mathcal{V}\backslash\{i,k\},\\ \delta \dot{x}^l(t) = -\mathsf{d}_{\mathrm{out}}^l\delta x^l(t) + \sum\limits_{j=1}^N \mathsf{a}_{lj}\delta y^j(t) + f^l - f^{l\prime}, & l \in \{i,k\},\end{cases} \tag{52a}$$

$$\begin{cases}\delta y^l(t) = \delta x^l, & l \in \mathcal{V}\backslash\{i,k\},\\ \delta y^l(t) = \delta x^l + g^l - g^{l\prime}, & l \in \{i,k\}.\end{cases} \tag{52b}$$

To complete our proof, we want to show that $y^l(t) = y^{l\prime}(t)$ (or equivalently $\delta y^l(t) \equiv 0$), $l \in \mathcal{V}$, for $t \in \mathbb{R}_{\geq 0}$, thus agent 1 cannot distinguish between the initial conditions $x^i(0)$ and $x^{i\prime}(0)$. Since, for a given initial condition and integrable external inputs the solution of an ordinary differential equation is unique, we achieve this goal by showing that if $\delta y^l(t) = 0$, $l \in \mathcal{V}$ applied in the state dynamics (52a)*, the resulted output* (52a) *satisfy $\delta y^l(t) \equiv 0$, $l \in \mathcal{V}$, $t \in \mathbb{R}_{\geq 0}$. For this, first note that since $\delta x^l(0) = 0$ for $l \in \mathcal{V}\backslash\{i,k\}$, then it follows from* (52a) *with $\delta y^l(t) = 0$, $l \in \mathcal{V}$, that $\delta x^l(t) \equiv 0$. Subsequently, from* (52b)*, we get the desired $\delta y^l(t) \equiv 0$, $t \in \mathbb{R}_{\geq 0}$ for $l \in \mathcal{V}\backslash\{i,k\}$. Next, we note that, from* (52a) *with $\delta y^l(t) = 0$, $l \in \mathcal{V}$, given $\delta x^i(0) = \beta_{ik}$ and $\delta x^k(0) = -\beta_{ik}$ we obtain*

$$\begin{aligned}\delta x^i(t) =& \beta_{ik}\mathrm{e}^{-\mathsf{d}_{\mathrm{out}}^i t} - \beta_{ik}\mathrm{e}^{-\mathsf{d}_{\mathrm{out}}^i t} + \beta_{ik}\mathrm{e}^{-(\mathsf{d}_{\mathrm{out}}^i+d)t}\\ =& \beta_{ik}\mathrm{e}^{-(\mathsf{d}_{\mathrm{out}}^i+d)t}\\ \delta x^k(t) =& -\beta_{ik}\mathrm{e}^{-\mathsf{d}_{\mathrm{out}}^k t} + \beta_{ik}\mathrm{e}^{-\mathsf{d}_{\mathrm{out}}^k t} - \beta_{ik}\mathrm{e}^{-(\mathsf{d}_{\mathrm{out}}^k+d)t}\\ =& -\beta_{ik}\mathrm{e}^{-(\mathsf{d}_{\mathrm{out}}^k+d)t}\end{aligned}$$

Subsequently, since $g^i - g^{i\prime} = -\beta_{ik}\mathrm{e}^{-(\mathsf{d}_{\mathrm{out}}^i+d)\tau}$ and $g^k - g^{k\prime} = \beta_{ik}\mathrm{e}^{-(\mathsf{d}_{\mathrm{out}}^k+d)\tau}$, from (52b)*, we get the desired $\delta y^l(t) \equiv 0$, $t \in \mathbb{R}_{\geq 0}$ for $l \in \{i,k\}$, which completes our proof.*