2018 Annual American Control Conference (ACC)
June 27–29, 2018. Wisconsin Center, Milwaukee, USA

Privacy preservation in a continuous-time static average consensus algorithm over directed graphs

Navid Rezazadeh and Solmaz S. Kia .

Abstract— **In this paper, we study the problem of privacy preservation of the continuous-time Laplacian static average consensus algorithm using additive perturbation signals. We consider this problem over a strongly connected and weight-balanced digraph. Starting from a local reference value, in static average consensus algorithm each agent constantly communicates with its neighboring agents to update its local state to compute the average of the reference values across the network. Since every agent transmits its local reference value to its in-neighbors, the reference value of the agents are trivially disclosed. In this paper, we investigate the possibility of preserving the privacy of the reference value of the agents by adding admissible perturbation signals to the local dynamics and the transmitted out signals of the agents. Admissible additive perturbation signals are those signals that do not perturb the final convergence point of the algorithm from the average of the reference values of the agents. Our results show that if an adversarial agent has access to the output of another agent and all the input signals transmitted to that agent, the adversary can discover the private reference value of that agent, regardless of the perturbation signals. Otherwise, the privacy of the agent can be preserved. We demonstrate our results through a numerical example.**

I. Introduction

In recent years, decentralized multi-agent cooperative operations have been proposed as effective solutions for some of today's important socio-economical challenges. However, privacy preservation concerns sometimes play a discouraging role in client participation in networked solutions in areas such as smart grid, banking or healthcare applications, where even though agents are willing to cooperate towards an effective operating point for the whole group, they do not want to release their local information. Motivated by the demand for privacy preserving network solutions to promote wider adoption of distributed operations in privacy sensitive domains, in this paper, we consider the privacy preservation problem in the distributed static average consensus problem.

In a network of agents each endowed with a local static reference value, static average consensus problem 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 are of interest in distributed computing, synchronization, estimation problems and control of multi-agent cyber physical systems. Static average consensus problem has been studied extensively in the literature (see e.g., [1], [2], [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 transmit this local value to its neighboring agents. Therefore, the privacy of the agents implementing this algorithm is trivially breached by sharing their local reference value at the first step of the algorithm. This paper studies the multi-agent static average consensus problem under the requirement of the privacy preservation of the agents' local reference value against internal non-cooperative and passive adversarial agents in the network.

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], [8], [9] and [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 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 adversarial agent does not have access to at least one of the signals transmitted to that agent.

Statement of contributions: In this paper we consider the problem of privacy preservation of the continuous-time static Laplacian average consensus algorithm over strongly connected and weight-balanced digraphs. The previous work reviewed above considers discrete-time algorithms over connected undirected graphs. Instead of random noises, we use continuous-time integrable additive perturbation signals to

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 CAREER award ECCS-1653838.

disguise the local reference value of the agents. We carefully examine the stability and convergence of the static average consensus algorithm in the presence of the perturbation signals to find necessary and also 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. In our privacy preservation evaluation, we assume that adversarial agents know the necessary conditions on the admissible perturbation signals. They can use this extra piece of information to enhance their knowledge set to discover the private value of the other agents. We show that if an adversarial agent has access to all the signals transmitted into and out of an agent, it can discover the local private value of that agent despite the existence of the perturbation signals. We also construct an observer that such an adversary can employ to obtain the reference value. Our next contribution is to present a class of admissible perturbation signals for which we can formally guarantee that if the adversarial agent does not have access to all the transmitted signals to an agent, it cannot obtain uniquely the local value of that agent. Our final contribution is identifying examples of graphs topologies in which the privacy of all the agents are preserved when they implement our proposed admissible additive perturbation signals. We demonstrate our results through a numerical example. Due to the space limitations, some of the proofs of our results are omitted, and will appear elsewhere.

Notations and definitions: Let $\mathbb{R}$, $\mathbb{R}_{\geq 0}$ and $\mathbb{R}_{>0}$, respectively, be the set of real, nonnegative real and positive real numbers. For a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, we denote its transpose matrix by $\mathbf{A}^\top$. We let $\mathbf{1}_n$ (resp. $\mathbf{0}_n$) denote the vector of n ones (resp. n zeros), and denote by $\mathbf{I}_n$ the $n \times n$ identity matrix. When clear from the context, we do not specify the matrix dimensions. We denote the standard Euclidean norm of vector $\mathbf{x} \in \mathbb{R}^n$ by $\|\mathbf{x}\| = \sqrt{\mathbf{x}^\top \mathbf{x}}$. For sets $\mathcal{A}$ and $\mathcal{B}$, the relative complement of $\mathcal{B}$ in $\mathcal{A}$ is $\mathcal{A}\backslash\mathcal{B} = \{x \in \mathcal{A} \,|\, x \notin \mathcal{B}\}$. In a network of N agents, to distinguish and emphasis that a variable is local to an agent $i \in \mathcal{V}$, we use superscripts, e.g., $f^i(t)$ is the local function of agent i. Moreover, if $p^i \in \mathbb{R}$ is a variable of agent $i \in \mathcal{V}$, the aggregated p^i's of the network is the vector $\mathbf{p} = [\{p^i\}_{i=1}^N] = [p^1, \cdots, p^N]^\top \in \mathbb{R}^N$. A measurable function h is called integrable if $\int |h| \mathrm{d}\mu \leq \infty$.

II. Problem definition

Consider a set of N agents each with a *reference value* $\mathsf{r}^i \in \mathbb{R}$, $i \in \mathcal{V}$ interacting over a strongly connected directed graph (digraph) $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathbf{A})$. Here, $\mathcal{V} = \{1, \cdots, N\}$ is the node set, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the edge set and $\mathbf{A} = [a_{ij}]$ is the weighted adjacency matrix of the digraph which satisfies $a_{ij} > 0$ if $(i, j) \in \mathcal{E}$ and $a_{ij} = 0$ otherwise. For graph theoretic definitions, terminologies and properties we follow [13]. Accordingly, in our developments below, 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. A digraph is called *strongly connected* if there is a directed path from every node to every other node in the digraph.

The objective in the static average consensus problem is for the agents $i \in \mathcal{V}$ to asymptotically compute $\frac{1}{N}\sum_{i=1}^N \mathsf{r}^i$ by only interacting with their out-neighbors. A well-known algorithm to arrive at the average consensus is based on driving a simple integrator dynamics using the weighted sum of the feedback of the difference between the local state of an agent and its out-neighbors (c.f. [1]) i.e.,

$$\dot{x}^i(t) = -\sum_{j=1}^N a_{ij}\,(x^i(t) - x^j(t)), \quad x^i(0) = \mathsf{r}^i, \qquad (1)$$

The asymptotic convergence is guaranteed when the weights a_{ij} of the algorithm are chosen such that the strongly connected digraph is also weight-balanced. Recall that a digraph is weight-balanced iff at each node $i \in \mathcal{V}$, the weighted out-degree $\mathsf{d}^i_{\text{out}} = \sum_{j=1}^N \mathsf{a}_{ij}$ and weighted in-degree $\mathsf{d}^i_{\text{in}} = \sum_{j=1}^N \mathsf{a}_{ji}$ coincide (although they might be different across different nodes). We consider a setting in which agents do not fully trust each other. In this setting, some of the agents in the network act as a passive adversarial eavesdropper (see Fig. 1), which without interrupting the execution of the algorithm (1), aim at obtaining the local reference value r^i of other agents $i \in \mathcal{V}$ by storing and processing the time history of the communication messages they receive. We assume that each adversary acts alone. Because in (1) the private value r^i of agent i 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. To preserve privacy of the agents, one can propose to add locally constructed perturbation signals $f^i(t) : \mathbb{R}_{\geq 0} \to \mathbb{R}$, $g^i(t) : \mathbb{R}_{\geq 0} \to \mathbb{R}$ to the local process and communication message of an agent to disguise this private value, i. e., modify (1) as follows

$$\dot{x}^i(t) = -\sum_{j=1}^N a_{ij}\,(x^i(t) - y^j(t)) + f^i(t), \;\; x^i(0) = \mathsf{r}^i, \quad \text{(2a)}$$

$$y^i(t) = x^i(t) + g^i(t). \quad \text{(2b)}$$

where y^i is the signal transmitted by agent $i \in \mathcal{V}$. Here f^i and g^i are assumed to be locally integrable, to guarantee existence and well-posedness of solutions of (2) (c.f. [14, page 30]). We refer to the set of perturbation signals $\{f^i, g^i\}_{i=1}^N$ for which the integrity of the static average consensus algorithm is preserved (i.e., $x^i(t) \to \frac{1}{N}\sum_{j=1}^N x^j(0) = \frac{1}{N}\sum_{j=1}^N \mathsf{r}^j$, $i \in \mathcal{V}$, as $t \to \infty$) as the *admissible perturbation signals*. Our objective in this paper is (a) to identify such admissible signals and (b) to analyze the privacy preservation properties of the modified algorithm (2) employing such signals.

III. Privacy preservation through additive perturbation signals

We start our study by obtaining a set of necessary and also sufficient conditions on the class of admissible perturbation signals f^i and g^i, $i \in \mathcal{V}$. We write the modified static average consensus algorithm in its compact form as

$$\dot{\mathbf{x}} = -\mathbf{L}\,\mathbf{x} - \mathbf{L}\,\mathbf{g} + \mathbf{f} + \mathbf{D}^{\text{out}}\,\mathbf{g} = -\mathbf{L}\,\mathbf{x} + \mathbf{f} + \mathbf{A}\,\mathbf{g}, \qquad (3)$$

where $\mathbf{L}$ is the graph *(out-) Laplacian* defined according to $\mathbf{L} = \mathbf{D}^{\text{out}} - \mathbf{A}$, in which $\mathbf{D}^{\text{out}} = \text{Diag}(\mathsf{d}^1_{\text{out}}, \cdots, \mathsf{d}^N_{\text{out}}) \in$

$\mathbb{R}^{N\times N}$, and $\mathbf{A}=[a_{ij}]$ is the adjacency matrix of the interaction topology. Here, recall that for a strongly connected and weight-balanced digraphs, L has a simple zero eigenvalue and the rest of eigenvalues have positive real parts. Moreover, $\mathsf{L}\mathbf{1}_N=\mathbf{0}$, $\mathbf{1}_N^\top\mathsf{L}=\mathbf{0}$ and $\mathrm{rank}(\mathsf{L})=N-1$. We denote eigenvalues of L by $\{\lambda_i\}_{i=1}^N$ and sort them such that $\lambda_1=0$, and $\mathrm{Re}(\lambda_i)\le\mathrm{Re}(\lambda_j)$ for any $i,j\in\mathcal{V}$ and $i<j$.

Theorem 3.1 (integrity of (3) in the presence of perturbation signals): *Consider algorithm* (2) *over a strongly connected and weight-balanced digraph. Let* $f^i:\mathbb{R}_{\ge0}\to\mathbb{R}$ *and* $g^i:\mathbb{R}_{\ge0}\to\mathbb{R}$, $i\in\mathcal{V}$, *be locally integrable.*

(a) *Let* f^i *and* g^i, $i\in\mathcal{V}$, *be such that* $x^i(t)\to\frac{1}{N}\sum_{j=1}^N \mathsf{r}^j$ *as* $t\to\infty$. *Then, we should have*

$$\lim_{t\to\infty}\int_0^t\sum\nolimits_{i=1}^N(f^i(\tau)+\mathsf{d}^i_{\mathrm{out}}\,g^i(\tau))\,d\tau=0. \tag{4}$$

(b) *Let* f^i *and* g^i, $i\in\mathcal{V}$ *be essentially bounded and vanishing signals that satisfy* (4). *Then, for any* $i\in\mathcal{V}$ *we have* $x^i(t)\to\frac{1}{N}\sum_{j=1}^N x^j(0)=\frac{1}{N}\sum_{j=1}^N\mathsf{r}^j$ *as* $t\to\infty$.

Since in our privacy preservation framework, each agent chooses its perturbation signal locally decide locally then to ensure that the necessary condition (4) holds, each agent $i\in\mathcal{V}$ should choose its admissible signals such that

$$\lim_{t\to\infty}\int_0^t(f^i(\tau)+\mathsf{d}^i_{\mathrm{out}}\,g^i(\tau))\,\mathrm{d}\tau=0. \tag{5}$$

Evidently, any adversarial agent is aware of the necessary condition (5) and can use this knowledge to identify the private reference value of the other agents. In our study, we also assume that the adversary knows the network topology.

Assumption 1 (Knowledge set of the adversary): *The Knowledge set of the adversarial agent includes signals that it receives from its out-neighbors, the adjacency matrix of the network (network topology) and the necessary condition* (5) *on the admissible perturbation signals* f^i *and* g^i, $i\in\mathcal{V}$.

From the perspective of an adversarial eavesdropper on an agent $i\in\mathcal{V}$ the dynamical system to observe is (2), with x^i as the internal state of agent i, f^i, g^i, and $\{y^j\}_{j\in\mathcal{N}^i_{\mathrm{out}}}$, as its inputs and y^i as its output that can be measured from tapping into the communication messages. Here, $\mathcal{N}^i_{\mathrm{out}}$ is the set of out-neighbors of agent $i\in\mathcal{V}$. Given a known input and measured outputs over some finite time interval (resp. infinite time), the traditional observability (resp. detectability) tests (c.f. [15] and [16]). evaluate whether we can uniquely identify the initial conditions of the system. But here, the inputs f^i and $g^i:\mathbb{R}_{\ge0}\to\mathbb{R}$ of agent $i\in\mathcal{V}$ are not available to the adversary. However, it is reasonable to presume that the adversary knows the necessary conditions stated in Theorem 3.1 for admissible perturbation signals. With regards to inputs $\{y^j\}_{j\in(\mathcal{N}^i_{\mathrm{out}}\cup\{i\})}$ an advertorial agent has only access to these signals if it is an in-neighbor of agent i and all the out-neighbors of agent i–see Fig. 1 for an example. The following result shows that in such a scenario the adversarial agent is able to identify the reference value

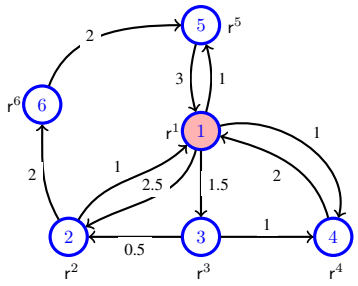


Fig. 1: A strongly connected and weight-balanced digraph in which agent 1 is the in-neighbor of all the out-neighbors of its out-neighbors $\{3,4,5\}$. As a result, agent 1 has direct access to the information transmitted to and from these agents. However, agent 1 is not the in-neighbor of all the out-neighbors of its out-neighbor 2, therefore it does not have direct access to all the information transmitted to this agent, specifically here the information from agent 6 to agent 2.

of the agent it is eavesdropping on despite the perturbation signals. Hereafter, without loss of generality, we assume that the adversarial agent is agent 1.

Theorem 3.2 (Observer design for an internal adversary): *Consider the modified static average consensus algorithm* (2) *with a set of admissible signals* $\{f^i,g^i\}_{i=1}^N$ *over a strongly connected and weight-balanced digraph* $\mathcal{G}$. *Let Assumption 1 hold and agent* 1 *be the internal adversary. Let agent* 1 *be the in-neighbor of agent* 2 *and all the out-neighbors of agent* 2. *Then, agent* 1 *can employ the observer*

$$\dot{\hat{x}}=\sum\nolimits_{j=1}^N a_{2j}(y^2-y^j),\qquad \hat{x}(0)=0, \tag{6a}$$
$$\hat{z}(t)=\hat{x}(t)+x^1(t), \tag{6b}$$

to asymptotically obtain r^2, *i.e.,* $\hat{z}\to\mathsf{r}^2$ *as* $t\to\infty$. *Moreover, at any* $t\in\mathbb{R}_{\ge0}$, *the tracking error of the observer satisfies*

$$\hat{z}(t)-\mathsf{r}^2=x^1(t)-x^2(t)+\int_0^t(f^2(\tau)+\mathsf{d}^2_{\mathrm{out}}\,g^2(\tau))d\tau.\ \square$$

In what follows, we investigate whether an adversarial agent can recover the local reference value of an out-neighbor of its when the adversary does not have access to all the transmitted signals to that out-neighbor. In our study hereafter, we assume that the admissible signals of every agent $i\in\mathcal{V}$ are $f^i(t)=0$ for $t\in\mathbb{R}_{\ge0}$ and g^i as follows

$$\dot{q}^i(t)=h^i(t),\quad 0\neq q^i(0)=-\int_0^\infty h^i(t)\,\mathrm{d}t=c^i<\infty, \tag{7a}$$
$$\dot{p}^i(t)=-\mathsf{d}^i_{\mathrm{out}}\,(p^i(t)+q^j(t)),\quad p^i(0)=0, \tag{7b}$$
$$g^i(t)=p^i(t)+q^i(t), \tag{7c}$$
$$\lim_{t\to\infty}h^i(t)=0. \tag{7d}$$

where $h^i(t):\mathbb{R}_{\ge0}\to\mathbb{R}$ is a bounded and continuous function chosen locally by agent $i\in\mathcal{V}$. The next result shows that $\{f^i,g^i\}_{i=1}^N$ as described above are admissible perturbation signals for modified static average consensus algorithm (2).

Lemma 3.1 ($f^i = 0$ and g^i given in (7) are admissible signals for (2)): *Consider the modified average consensus algorithm* (2) *over a strongly connected and weight-balanced digraph. For every $i \in \mathcal{V}$, let $f^i = 0$ for $t \in \mathbb{R}_{\geq 0}$ and g^i be given by* (7)*, where h^i is a bounded continuous function. Then, we have $x^i(t) \to \frac{1}{N}\sum_{j=1}^N \mathsf{r}^j$, as $t \to \infty$.*

Proof: Our proof is based on showing that f^i and g^i, $i \in \mathcal{V}$ satisfy the set of sufficient conditions that is given in statement (b) of Theorem (3.1). Note that $q^i(t) = q^i(0) + \int_0^t h^i(t)\,\mathrm{d}t, \quad t \in \mathbb{R}_{\geq 0}$, which under the given $q^i(0)$, indicates $\lim_{t\to\infty} q^i(t) = 0$. Then, (7b) is an internally exponentially stable LTI system with a bounded and vanishing external input signal $q^i(t)$. Therefore, the ISS analysis results (c.f. [17, page 175]) guarantees that the trajectories of p^i are bounded and also satisfy $\lim_{t\to\infty} p^i(t) = 0$. Thereby, $g^i(t)$ is an essentially bounded and vanishing signal ($\lim_{t\to\infty} g^i(t) = \lim_{t\to\infty}(q^i(t) + p^i(t) = 0$). To complete the proof, we need to show that $\int_0^\infty (f^i(\tau) + \mathsf{d}^i_{\text{out}} g^i(\tau))\mathrm{d}\tau = 0$, which given $f^i(t) = 0$, simplifies to $\int_0^\infty g^i(\tau)\mathrm{d}\tau = 0$. In this regard note that the differential equation describing $g^i(t) = p^i(t) + q^i(t)$ is given by $\dot{p}^i(t) + \dot{q}^i(t) = -\mathsf{d}^i_{\text{out}}\,(p^i(t) + q^j(t)) + h^i(t)$. The solution of this differential equation is given by

$$p^i(t)+q^i(t) = \mathrm{e}^{-\mathsf{d}^i_{\text{out}}\,t}(p^i(0)+q^i(0)) + \int_0^t \mathrm{e}^{-\mathsf{d}^i_{\text{out}}\,(t-\tau)} h^i(\tau)\mathrm{d}\tau.$$

Therefore, we can write (recall the initial conditions of (7))

$$\begin{aligned}\int_0^t g^i(\tau)\mathrm{d}\tau &= q^i(0)\int_0^t \mathrm{e}^{-\mathsf{d}^i_{\text{out}}\,\nu}\mathrm{d}\nu + \int_0^t\int_0^\nu \mathrm{e}^{-\mathsf{d}^i_{\text{out}}\,(\nu-\tau)} h^i(\tau)\mathrm{d}\tau\,\mathrm{d}\nu \\ &= \frac{-1}{\mathsf{d}^i_{\text{out}}}\Big(\int_0^\infty h^i(\tau)\mathrm{d}\tau\Big)\,(1 - \mathrm{e}^{-\mathsf{d}^i_{\text{out}}\,t}) + \\ &\qquad \int_0^t \mathrm{e}^{-\mathsf{d}^i_{\text{out}}\,\nu}\int_0^\nu \mathrm{e}^{\mathsf{d}^i_{\text{out}}\,\tau} h^i(\tau)\mathrm{d}\tau\,\mathrm{d}\nu. \end{aligned} \tag{8}$$

Using integration by parts, the second summand in the right hand side of the equation above can be written as

$$\begin{aligned}&\int_0^t \mathrm{e}^{-\mathsf{d}^i_{\text{out}}\,\nu}\int_0^\nu \mathrm{e}^{\mathsf{d}^i_{\text{out}}\,\tau} h^i(\tau)\mathrm{d}\tau\,\mathrm{d}\nu = \\ &\frac{-1}{\mathsf{d}^i_{\text{out}}}\mathrm{e}^{-\mathsf{d}^i_{\text{out}}\,t}\int_0^t \mathrm{e}^{\mathsf{d}^i_{\text{out}}\tau} h^i(\tau)\mathrm{d}\tau - \int_0^t \frac{-1}{\mathsf{d}^i_{\text{out}}}\mathrm{e}^{-\mathsf{d}^i_{\text{out}}\,\nu}\mathrm{e}^{\mathsf{d}^i_{\text{out}}\,\nu} h^i(\nu)\,\mathrm{d}\nu \\ &= \frac{-1}{\mathsf{d}^i_{\text{out}}}\psi(t) + \frac{1}{\mathsf{d}^i_{\text{out}}}\int_0^t h^i(\nu)\,\mathrm{d}\nu. \end{aligned} \tag{9}$$

where $\psi(t) = \mathrm{e}^{-\mathsf{d}^i_{\text{out}}\,t}\int_0^t \mathrm{e}^{\mathsf{d}^i_{\text{out}}\,\tau} h^i(\tau)\mathrm{d}\tau$. Next, we show $\lim_{t\to\infty}\psi(t) = 0$, by showing that

$$\text{for } \forall \epsilon > 0,\ \exists T > 0 \text{ s.t if } t > T \text{ then } |\psi(t)| < \epsilon.$$

Recall that h^i is a continuous and bounded signal that satisfies $\lim_{t\to\infty} h^i(t) = 0$. Therefore, for every given $\epsilon \in \mathbb{R}_{>0}$, there exists a $t_1 \in \mathbb{R}_{>0}$ such that $|h^i(t)| < \frac{\mathsf{d}^i_{\text{out}}\epsilon}{2}$. For $t > t_1$, we write $\psi(t)$ as below

$$\psi(t) = \mathrm{e}^{-\mathsf{d}^i_{\text{out}}\,t}\int_0^{t_1} \mathrm{e}^{\mathsf{d}^i_{\text{out}}\,\tau} h^i(\tau)\,\mathrm{d}\tau + \int_{t_1}^t \mathrm{e}^{-\mathsf{d}^i_{\text{out}}\,(t-\tau)} h^i(\tau)\,\mathrm{d}\tau.$$

Because $h^i(t)$ is a bounded signal, we can write $\big|\int_0^{t_1} \mathrm{e}^{\mathsf{d}^i_{\text{out}}\,\tau} h^i(\tau)\,\mathrm{d}\tau\big| = c^i < \infty$. Thus, we can conclude that

$$\begin{aligned}|\psi(t)| &\leq \mathrm{e}^{-\mathsf{d}^i_{\text{out}}\,t} c^i + \int_{t_1}^t \mathrm{e}^{-\mathsf{d}^i_{\text{out}}\,(t-\tau)}|h(\tau)|\,\mathrm{d}\tau \\ &\leq \mathrm{e}^{-\mathsf{d}^i_{\text{out}}\,t} c^i + \frac{\mathsf{d}^i_{\text{out}}\epsilon}{2}\int_{t_1}^t \mathrm{e}^{-\mathsf{d}^i_{\text{out}}\,(t-\tau)}\,\mathrm{d}\tau, \\ &= \mathrm{e}^{-\mathsf{d}^i_{\text{out}}\,t} c^i + \frac{\mathsf{d}^i_{\text{out}}\epsilon}{2}\frac{1}{\mathsf{d}^i_{\text{out}}}(1 - \mathrm{e}^{-\mathsf{d}^i_{\text{out}}(t-t_1)}), \\ &< \mathrm{e}^{-\mathsf{d}^i_{\text{out}}\,t} c^i + \frac{\epsilon}{2}, \quad t > t_1 > 0. \end{aligned}$$

Because $\lim_{t\to\infty} \mathrm{e}^{-\mathsf{d}^i_{\text{out}}\,t} = 0$, there exists a $t_2 \in \mathbb{R}_{>0}$ such that $\mathrm{e}^{-\mathsf{d}^i_{\text{out}}\,t} < \frac{\epsilon}{2c^i}$. Therefore, by taking $T > \max\{t_1, t_2\}$ we conclude that $|\psi(t)| < \epsilon$. Because $\lim_{t\to\infty}\psi(t) = 0$, from (9) and (8) we can conclude that $\lim_{t\to\infty}\int_0^t g^i(\tau)\,\mathrm{d}\tau = \lim_{t\to\infty}(-\frac{1}{\mathsf{d}^i_{\text{out}}}\int_0^t h(\tau)\,\mathrm{d}\tau + \frac{1}{\mathsf{d}^i_{\text{out}}}\int_0^t h(\tau)\,\mathrm{d}\tau) = 0$,, which concludes the proof. ■

Our next result considers an implementation of the modified static average consensus algorithm (2) in which agents choose their admissible perturbation signals according to $f^i = 0$ and g^i in (7). We show that in this implementation if an adversarial agent does not have direct access to all the signals that are transmitted to any of its out-neighbors, it cannot uniquely identify the initial condition of that out-neighbor, i.e., the local reference value of that agent stays private. In the developments below we denote the set of the in-neighbors of an agent $i \in \mathcal{V}$ by $\mathcal{N}^i_{\text{in}}$.

Theorem 3.3 (Privacy preservation): *Consider the modified static average consensus algorithm* (2) *with a set of admissible signals $\{f^i, g^i\}_{i=1}^N$ over a strongly connected and weight-balanced digraph $\mathcal{G}$. Let the admissible signals for $i \in \mathcal{V}$ be $f^i = 0$ and g^i given in* (7)*. Let Assumption 1 hold and agent* 1 *be the adversary. Let $\mathcal{N}^{2,-1}_{out} = (\mathcal{N}^2_{out} \backslash (\mathcal{N}^1_{out} \cup \{1\})$ be the set of the out-neighbors of agent* 2 *that are not out-neighbors of agent* 1*. Let agent* 2 *be an out-neighbor of agent* 1 *for which $\mathcal{N}^{2,-1}_{out} \neq \{\}$. Then, agent* 1 *cannot uniquely identify the reference value of agents* 2 *and $\mathcal{N}^{2,-1}_{out}$, i.e., agent* 1 *cannot uniquely identify r^i of $i \in (\mathcal{N}^{2,-1}_{out} \cup \{2\})$.*

Proof: Given a set of reference inputs $\{\mathsf{r}^i\}_{i=1}^N$, consider '*the actual scenario*' in which algorithm (2) is driven by $x^i(0) = \mathsf{r}^i$, and admissible perturbation signals from the set described in the statement. We represent the perturbation signal g^i by $g^i \sim (q^i(0), h^i(t))$, $i \in \mathcal{V}$. To show that agent 1 cannot uniquely identify r^i of $i \in (\mathcal{N}^{2,-1}_{\text{out}} \cup \{2\})$, we show that there exist other sets of admissible perturbation signals and initial conditions for agents 2 and any agent in $\mathcal{N}^{2,-1}_{\text{out}}$ for which every agent $i \in \mathcal{V}$ still converges to $\frac{1}{N}\sum_{j=1}^N \mathsf{r}^j$ and also the output signal of every out-neighbor of agent 1 is exactly the same as the corresponding signals in the actual scenario. In the following, we show one of these possible cases. Without loss of generality, assume that $3 \in \mathcal{N}^{2,-1}_{\text{out}}$. Let $\bar{\mathcal{N}} = \mathcal{N}^3_{\text{in}} \cup \{3\}$ (note that $2 \in \bar{\mathcal{N}}$). Next, let $t \mapsto \bar{\mathbf{x}}(t)$ be trajectories of the modified average consensus algorithm (2), initialized according to $\bar{x}^i(0) = \mathsf{r}^i$, $i \in (\mathcal{V}\backslash\bar{\mathcal{N}})$, and $\bar{x}^j(0) \in \mathbb{R}$, $j \in \bar{\mathcal{N}}$ such that $\sum_{j\in\bar{\mathcal{N}}} \bar{x}^j(0) = \sum_{j\in\bar{\mathcal{N}}} \mathsf{r}^j$, and admis-

sible perturbation signals $\{\bar{f}^i \equiv 0, \bar{g}^i \sim (\bar{q}^i(0), \bar{h}^i(t))\}_{i\in\mathcal{V}}$. Since $\frac{1}{N}\sum_{j=1}^N \bar{x}^j(0) = \frac{1}{N}\sum_{j=1}^N \mathsf{r}^j$ and we are using admissible perturbations $\{\bar{f}^i \equiv 0, \bar{g}^i \sim (\bar{q}^i(0), \bar{h}^i(t))\}_{i\in\mathcal{V}}$, we have $\bar{x}^i(t) \to \frac{1}{N}\sum_{j=1}^N \mathsf{r}^j$, $i \in \mathcal{V}$, as $t \to \infty$. In this alternative case we let every agent $i \in (\mathcal{V}\backslash\bar{\mathcal{N}})$ use the same admissible perturbation signals $f^i \equiv 0$ and $g^i \sim (q^i(0), h^i(t))$ as in the actual scenario. Let $e^i_x(t) = x^i(t) - \bar{x}^i(t)$, $e^i_y(t) = y(t) - \bar{y}^i(t)$, $\bar{e}^i_{q(0)} = q^i(0) - \bar{q}^i(0)$, and $\bar{e}^i_h = h^i - \bar{h}^i$, $i \in \mathcal{V}$. Note that $\bar{e}^i_{q(0)} = 0$ and $\bar{e}^i_h(t) \equiv 0$ for $i \in (\mathcal{V}\backslash\bar{\mathcal{N}})$. Next, we show that there exists admissible perturbation signals $\{\bar{f}^j \equiv 0, \bar{g}^j \sim (\bar{q}^j(0), \bar{h}^j(t))\}_{j\in\bar{\mathcal{N}}}$ for which the output $\bar{y}^k(t) = y^k(t)$, $k \in \mathcal{N}^1_{\text{out}}$, i.e., agent 1 cannot distinguish between initial conditions $x^j(0)$ and $\bar{x}^j(0)$, $j \in \bar{\mathcal{N}}$. In the proof below we use the fact that given a set of initial conditions and integrable external signals, the solution of any linear ordinary differential equation is unique.

Our choice of perturbation signals should also result in $e^i_y(t) = 0$ for all $t \in \mathbb{R}_{\geq 0}$, for every $i \in \mathcal{V}\backslash\{3\}$. Let $\bar{q}^3(0) = q^3(0)$ and $\bar{h}^3(t) = h^3(t)$, $t \in \mathbb{R}_{>0}$. Then, we have

$$\dot{e}^i_x = -\mathsf{d}^i_{\text{out}}\, e^i_x, \quad i \in (\mathcal{V}\backslash\mathcal{N}^3_{\text{in}}) \cup \{3\}), \tag{10a}$$

$$\dot{e}^j_x = -\mathsf{d}^j_{\text{out}}\, e^j_x + a_{j3}\, e^3_x, \quad j \in \mathcal{N}^3_{\text{in}}. \tag{10b}$$

For $i \in \mathcal{V}\backslash\bar{\mathcal{N}}$, because $e^i_x(0) = 0$, from (10a) we obtain that $e^i_x(t) = 0$, for all $t \in \mathbb{R}_{\geq 0}$. Then, because $e^i_g(t) = 0$ for $t \in \mathbb{R}_{\geq 0}$, our assumption of $e^i_y(t) \equiv 0$ for $t \in \mathbb{R}_{\geq 0}$ is correct for $i \in \mathcal{V}\backslash\bar{\mathcal{N}}$. Since $\mathcal{N}^1_{\text{out}} \subset ((\mathcal{V}\backslash\bar{\mathcal{N}}))\cup\{2\}$, what remains to show is that $e^2_y \equiv 0$ for all $t \in \mathbb{R}_{\geq 0}$. In what follows recall that $2 \in \mathcal{N}^3_{\text{in}}$.

Because $e^3_x(0) \neq 0$, from (10a) we obtain that

$$e^3_x(t) = \mathrm{e}^{-\mathsf{d}^3_{\text{out}}\, t}\, e^3_x(0), \quad t \in \mathbb{R}_{\geq 0}.$$

Let the admissible perturbation signals for the in-neighbors of agent 3 be such that

$$\bar{h}^j(t) = h^j(t) - a_{j3}\, \mathrm{e}^{-\mathsf{d}^3_{\text{out}}\, t}\, e^3_x(0), \quad j \in \mathcal{N}^3_{\text{in}},$$

which gives

$$e^j_h(t) = -a_{j3}\, \mathrm{e}^{-\mathsf{d}^3_{\text{out}}\, t}\, e^3_x(0), \quad t \in \mathbb{R}_{\geq 0},\ j \in \mathcal{N}^3_{\text{in}},$$

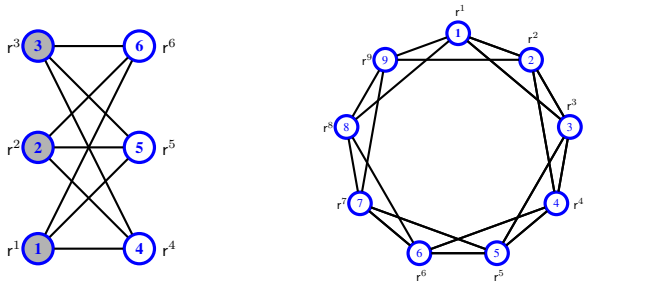
For this signal from (7a) we have

$$e^j_{q(0)} = \int_0^\infty e^j_h(t)\mathrm{d}t = \frac{a_{j3}}{\mathsf{d}^3_{\text{out}}}\, e^3_x(0), \qquad j \in \mathcal{N}^3_{\text{in}}.$$

Moreover, we can write $\dot{e}^j_q = -\mathrm{e}^{-\mathsf{d}^3_{\text{out}}\, t}\, e^3_x(0)$, for $j \in \mathcal{N}^3_{\text{in}}$. Note here that $\bar{h}^j(t)$, $j \in \mathcal{N}^3_{\text{in}}$, as defined above is admissible because it is bounded and continuous function which satisfies also $\lim_{t\to\infty} \bar{h}^j(t) = 0$. Define $e^j_p = p^i - \bar{p}^i$. Then for every agent $j \in \mathcal{N}^3_{\text{in}}$ we can write

$$\begin{aligned}\dot{e}^j_x + \dot{e}^j_p + \dot{e}^j_q =& -\mathsf{d}^j_{\text{out}}\,(e^j_x + e^j_p + e^j_q) + \\ & a_{j3}\, e^3_x - a_{j3}\, \mathrm{e}^{-\mathsf{d}^3_{\text{out}}\, t}\, e^3_x(0) \\ =& -\mathsf{d}^j_{\text{out}}\,(e^j_x + e^j_p + e^j_q).\end{aligned} \tag{11}$$

For $j \in \mathcal{N}^3_{\text{in}}$, let $e^j_x(0) = -\frac{a_{j3}}{\mathsf{d}^3_{\text{out}}}\, e^3_x(0)$ so that $e^j_x(0) + e^j_p(0) + e^j_q(0) = 0$. As a result from (11) we obtain for any agent $j \in \mathcal{N}^3_{\text{in}}$ that $e^j_y(t) = e^j_x(t) + e^j_p(t) + e^j_q(t) \equiv 0$ for $t \in \mathbb{R}_{\geq 0}$.



(a) An example of a cyclic bipartite undirected connected graph.

(b) A 4-regular ring lattice undirected connected graph on 12 vertices.

Fig. 2: Examples of graphs in which privacy of all agents implementing the modified static average consensus algorithm (2) with admissible perturbation signals $\{f^i \equiv 0, g^i \sim (q^i(0), h^i(t))\}_{i\in\mathcal{V}}$ is preserved.

To complete the proof, we show that the initial conditions described above for the alternative case satisfy $\sum_{j\in\bar{\mathcal{N}}} \bar{x}^j(0) = \sum_{j\in\bar{\mathcal{N}}} \mathsf{r}^j$. For this note that

$$\begin{aligned}\sum_{j\in\bar{\mathcal{N}}} e^j_x(0) =& e^3_x(0) + \sum_{j\in\mathcal{N}^3_{\text{in}}} e^j_x(0) = e^3_x(0) - \sum_{j\in\mathcal{N}^3_{\text{in}}} \frac{a_{j3}}{\mathsf{d}^3_{\text{out}}}\, e^3_x(0) \\ =& e^3_x(0) - \frac{\mathsf{d}^3_{\text{in}}}{\mathsf{d}^3_{\text{out}}}\, e^3_x(0) = e^3_x(0) - e^3_x(0) = 0.\end{aligned}$$

Here, we used the fact that $\mathcal{G}$ is weight-balanced, therefore, $\mathsf{d}^3_{\text{out}} = \mathsf{d}^3_{\text{in}} = \sum_{j\in\mathcal{N}^3_{\text{in}}} a_{j3}$. ■

Undirected cyclic bipartite graphs and 4-regular ring lattice undirected graphs with $N > 5$ are examples of network topologies that satisfy the relation mentioned for the adversarial node and its out-neighbors in Theorem 3.3 (see Fig. 2). Therefore, the privacy of the agents in these graphs are preserved when they implement the modified static average consensus algorithm (2) with the admissible perturbation signals $\{f^i \equiv 0, g^i \sim (q^i(0), h^i(t))\}_{i\in\mathcal{V}}$.

IV. Numerical Example

We demonstrate our results using an execution of the modified static average consensus (2) over the strongly connected and weight-balanced digraph in Fig. 3. Let the perturbation signals be such that $f^i(t) \equiv 0$ and $g^i(t)$ be defined according to (7). The local reference value of the agents as well the h^i component of the the perturbation signal g^i are specified in Fig. 3. The adversarial agent here is agent 1, which wants to obtain the reference values of its out-neighbors $\{2, 6, 5\}$. In regards to agent 5, as guaranteed in Theorem 3.2, agent 1 can employ the observer (6) to obtain $x^5(0) = \mathsf{r}^5 = -3$ (see Fig. 4). Agent 1 however, cannot uniquely identify r^2 and agent r^6, since each of these agents have out-neighbors that are not out-neighbors of agent 1. To show this, consider an alternative implementation of algorithm (2) with initial and admissible perturbation signals

$$\bar{x}^1(0) = 3, \quad \bar{x}^2(0) = -2, \ \bar{x}^3(0) = 9, \quad \bar{x}^4(0) = -3, \tag{12a}$$

$$\bar{x}^5(0) = -3, \ \bar{x}^6(0) = 11, \ \bar{x}^7(0) = 16, \ \bar{x}^8(0) = -3, \tag{12b}$$

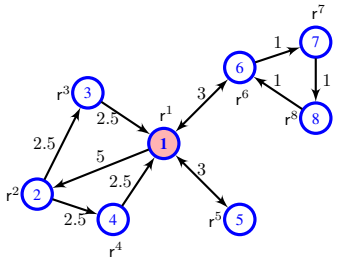


Fig. 3: A strongly connected and weight-balanced digraph $\mathcal{G}$ in which node 1 is the adversarial agent.

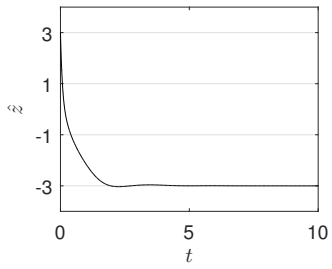


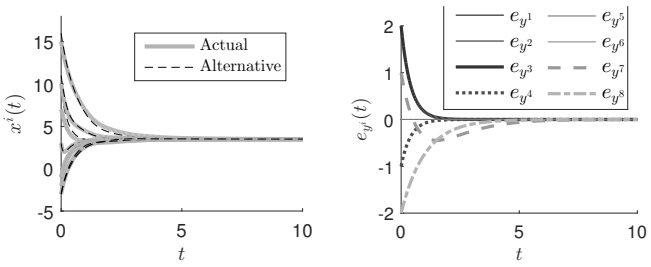
Fig. 4: Adversarial agent 1's estimate of r^5 using the observer (6). As seen, the adversary recovers the reference value of agent 5, i.e., $x^5(0) = \mathsf{r}^5 = -3$.

$$\bar{h}^i(t) = h^i(t), \qquad i \in \{1, 3, 4, 5, 7, 8\}, \tag{12c}$$
$$\bar{h}^2(t) = h^2(t) - 2.5\mathrm{e}^{-2.5t}, \tag{12d}$$
$$\bar{h}^6 = h^6(t) + 2t\,\mathrm{e}^{-t} - \mathrm{e}^{-t}, \tag{12e}$$

where $\frac{1}{8}\sum_{i=1}^{8} \bar{x}^i(0) = \frac{1}{8}\sum_{i=1}^{8} x^i(0) = \frac{1}{8}\sum_{i=1}^{8} \mathsf{r}^i = 3.5$. As Fig. 5(a) shows the execution of algorithm (2) using the initial conditions and perturbation signal specified in Fig. 3 (actual case) and those in (12) (alternative case) converges to the same final value of 3.5. Let $e_y^i = y^i - \bar{y}^i$, $i \in \{1, \ldots, 8\}$ be the error between the output of agents in the actual and alternative cases. As Fig. 5(b) $e_y^i \equiv 0$ for all $i \in \mathcal{N}_{\text{out}}^1 = \{2, 5, 6\}$. Therefore, agent 1 cannot distinguish between these two cases.

V. 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 necessary and sufficient conditions on admissible perturbation signals, which do not perturb the final convergence point of the algorithm. Then, we showed that despite employing additive perturbation signals, if an adversarial agent in the network has access to all the input and out signals of an agent, it can employ an asymptotic observer to obtain the initial value of the state equation of that agent, which is the reference value of the agent. Our next contribution was to identify the conditions under which an agent's privacy is preserved. In this paper, we only studied the problem of privacy preservation with respect to internal adversarial agents. Future work will focus on studying privacy preservation with respect to external adversaries. We will also extend our results to other multi-agent distributed algorithms such as dynamic average consensus and distributed optimization algorithms.



(a) State trajectories of the agents

(b) Output difference $e_y^i = y^i - \bar{y}^i$, $i \in \{1, \ldots, 8\}$

Fig. 5: Simulation results for the execution of algorithm (2) using the initial conditions and perturbation signal specified in Fig. 3 (actual case) and those in (12) (alternative case).

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

[14] J. Hale, *Ordinary Differential Equations*. Pure and Applied Mathematics - Marcel Dekker, R. E. Krieger Publishing Company, 1980.

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

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

[17] H. K. Khalil, *Nonlinear Systems*. Englewood Cliffs, NJ: Prentice Hall, 3 ed., 2002.