Secure and Privacy Preserving Average Consensus

Qipeng Liu*, Xiaoqiang Ren*, and Yilin Mo*

***Abstract*—Due to the wide application of consensus algorithms, its privacy and security problems have attracted significant attention. In this paper, we consider the average consensus scheme threatened by a set of cooperating malicious and curious attackers, who intend to not only estimate the initial states of other benign agents but also disturb the consensus (perturb the final consensus value or prevent consensus) via injecting malicious signal to the consensus process. A privacy-preserving average consensus algorithm equipped with an attack detector is designed for each benign agent. We show that the privacy of initial states of all benign agents are guaranteed under mild conditions. The false alarm rate of the detector can be arbitrarily low by choosing proper thresholds for the detector. To characterize the maximum disturbance the adversaries can introduce to the system, we calculate the reachable set of all possible system states under stealthy attacks, i.e., attacks the do not trigger an alarm. Numerical examples are given to validate the theoretical results.**

I. Introduction

Consensus is one of the most frequent tasks to be accomplished in networked systems, such as the synchronization of clock speed in wireless sensor networks, the load balance of parallel computers, and the unanimous rendezvous point selection for a team of autonomous vehicles [1]–[6]. In the literature, several consensus algorithms are proposed dealing with continuous-time systems [7], [8], discrete-time systems [9]–[12], communication bandwidth constrained systems [13], etc.

Most recently, some drawbacks of consensus algorithms have gradually attracted attention in the research community. One is the security problem. If there exist some malicious agents who intend to disrupt the consensus by injecting additional signals into the system instead of following the consensus scheme, the final consensus may be not achievable or its value is manipulated. Sundaram and Hadjicostis show the resilience property of the consensus scheme in the presence of malicious agents using the parity space method from fault detection and isolation theory [14]. Pasqualetti et al. consider the attack detection and identification problem based on unknown input observer method [15], [16].

Besides the security problem, the privacy problem of consensus algorithms has also be considered. If all agents update their states according to the consensus algorithm, under some observability conditions, a curious agent can infer the initial states of all the other agents. This may not be desirable in the privacy sensitive situations, like opinion survey on a private question among a group of individuals. To guarantee the privacy of the initial states of the agents, Huang et al. propose a consensus algorithm where an exponentially decaying noise process is added into the consensus computation, by which a randomized consensus value is obtained [17]. Manitara and Hadjicostis consider the privacy problem in the average consensus situation, where the final consensus value is required to be the exact average of the initial states of all the agents [18]. Also dealing with average consensus, Mo and Murray provide a quantitative result on how good the initial state can be estimated and obtain necessary and sufficient conditions under which the privacy of the benign agents are preserved [19].

In this paper, we consider the average consensus scheme threatened by a set of attackers who are both malicious and curious. We adopt the privacy preserving algorithm proposed in [19] and show that it can also guarantee the privacy in our situation. Note that, since noisy signals exist in our privacy preserving consensus algorithm, the detector design methods dealing with noiseless consensus algorithms in [14]–[16] cannot be applied to our situation. We design for each benign agent a detector with time-varying alarm thresholds. We exam the performance of the detector by computing its false alarm rate and characterizing the researchable region of the system under malicious attacks.

The rest of the paper is organized as follows: In Section II, we briefly describe the average consensus algorithm and introduce two kinds of attack models. In Section III, we adopt a privacy preserving average consensus algorithm equipped with an attack detector, and exam its performance. In Section IV, numerical examples are provided to illustrate the effect of attacks on the system. Finally, Section V concludes the paper.

Notations: $\mathbb{R}^n$ is the set of $n \times 1$ real vectors. $\mathbb{R}^{n\times m}$ is the set of $n \times m$ real matrices. $\operatorname{tr} M$ is the trace of matrix M. $\mathbf{1}$ is an all one vector of proper dimension. $\|v\|$ indicates the 2-norm of the vector v, while $\|M\|$ is the induced 2-norm of the matrix M. Denote by $\mathbb{S}_+^n$ the set of $n \times n$ positive semi-definite matrices. X^+ is the Moore–Penrose pseudoinverse of the matrix X. For any set $\mathcal{S} \subseteq \mathbb{R}^n$ and $\rho \in \mathbb{R}$, define the following set

$$\rho S \triangleq \{x : x = \rho x' \text{ for some } x' \in \mathcal{S}\}.$$

II. Problem Formulation

A. Average Consensus

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

Suppose that each agent $i \in V$ has an initial state $x_i(0)$. At any time k, agent i first broadcasts its state to all of its

*: Qipeng Liu, Xiaoqiang Ren, and Yilin Mo are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. Email: {qpliu, xren, ylmo}@ntu.edu.sg

Preprint submitted to 56th IEEE Conference on Decision and Control.
Received March 20, 2017.

neighbors and then updates its own state in the following linear combination manner:

$$x_i(k+1) = \sum_{j \in N_i} a_{ij} x_j(k). \tag{1}$$

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

$$x(k+1) = Ax(k). \tag{2}$$

We say the agents reach a consensus if

$$\lim_{k \to \infty} x(k) = c\mathbf{1}, \tag{3}$$

where c is an arbitrary scalar constant. If $c = \frac{1}{n}\sum_{i=1}^{n} x_i(0)$, then we say an average consensus is reached.

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

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

(A2) $A\mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ is a column vector with all ones.

B. Attack Models

In this subsection, we consider the average consensus in an adversarial environment. In particular, we consider two kinds of security problems in the system:

Malicious Attack: Some agents intend to disrupt the average consensus. To this end, they will add arbitrary input signals instead of following the updating rule (1), i.e.,

$$x_i(k+1) = \sum_{j \in N_i} a_{ij} x_j(k) + u_i(k), \tag{4}$$

where $u_i(k) \neq 0$ is the attack signal added by i at time k.

This kind of malicious attack can potentially either prevent benign agents from reaching a consensus or manipulate the final consensus value to be arbitrary.

Curious Attack: Some agents intend to estimate the initial states of other agents. This might not be desirable in opinion pool over social networks and other privacy sensitive situations.

In this paper, we deal with a set of attackers who are both malicious and curious. Our aim is to design a privacy preserving average consensus algorithm equipped with a detector such that

1) The privacy of initial states of benign agents is guaranteed under mild conditions;

2) The false alarm rate of the detector is low;

3) If the alarm is not triggered, the final consensus value is inside a bounded area around the average value of the initial states.

III. Main Results

A. Privacy Preserving Consensus Algorithm and Attack Detector Design

In order to protect the agents' privacy, we adopt the privacy preserving algorithm proposed in [19]. For the sake of completeness, we briefly describe the algorithm below.

Let $v_i(k) (i = 1, 2, \cdots, n, k = 0, 1, \cdots)$ be standard normal distributed random variables, which are independent across i and k. Denote $v(k) = [v_1(k), v_2(k), \cdots, v_n(k)]$. Based on $v(k)$ we can construct the following noisy signals

$$w(k) = \begin{cases} v(0), & \text{if } k = 0 \\ \varphi^k v(k) - \varphi^{k-1} v(k-1), & \text{otherwise} \end{cases} \tag{5}$$

where $0 < \varphi < 1$.

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

$$x^+(k) = x(k) + w(k).$$

Under this privacy preserving scheme, and consider a set of malicious attackers $\{i_1, i_2, \cdots, i_m\}$, the state updating rule is as follows:

$$\begin{aligned} x(k+1) &= Ax^+(k) + \Gamma u(k) \\ &= A(x(k) + w(k)) + \Gamma u(k), \end{aligned}$$

where $\Gamma = [e_{i_1}, e_{i_2}, \cdots, e_{i_m}]$ with e_i being the ith canonical basis vector in $\mathbb{R}^n$, and $u(k) = [u_{i_1}(k), u_{i_2}(k), \cdots, u_{i_m}(k)]^T$.

Next we turn to design attack detector for each benign agent. Without loss of generality, we assume that agent 1 is benign, and we focus on detector design for agent 1. Suppose the neighbors of agent 1 are $\{j_1, j_2, \cdots, j_p\}$. Define the measurement of agent 1 as

$$y(k) = C(x(k) + w(k)), \tag{6}$$

where $C = [e_1, e_{j_1}, e_{j_2}, \cdots, e_{j_p}]^T$.

We first propose a linear filter to compute state estimate $\hat{x}(k)$:

$$\hat{x}(k+1) = A\hat{x}(k) + K[y(k+1) - CA\hat{x}(k)], \ \hat{x}(0) = 0 \tag{7}$$

where K is a filter gain such that $A - KCA$ is stable and $A - AKC$ is diagonalizable.

Define the residue

$$r(k) = y(k) - CA\hat{x}(k-1). \tag{8}$$

We design the malicious attack detector which compares $\|r(k)\|$ with a time-varying threshold and triggers an alarm if and only if $\|r(k)\|$ is greater than the threshold. Mathematically, this can be represented as

$$\begin{cases} \|r(k)\| > c\rho^k & \text{Triggering an alarm} \\ \|r(k)\| \leq c\rho^k & \text{Not triggering an alarm} \end{cases}$$

where $c\rho^k$ is the time-varying threshold, and $c > 0$, $\rho < 1$.

Note that $r(k)$ can be decomposed into three parts due to the linearity of the system:

$$r(k) = r^0(k) + r^n(k) + r^a(k),$$

where $r^0(k)$ is caused by the nonzero initial state $x(0)$, $r^n(k)$ by the noise signal $w(k)$, and $r^a(k)$ by the possible attacks $u(k)$. In the absence of attacks, a false alarm is triggered if

$$\|r^0(k) + r^n(k)\| > c\rho^k.$$

Defined the false alarm rate α as the probability of triggering a false alarm at least once from initial time to infinity, i.e.,

$$\alpha \triangleq P\{\sup_k \rho^k \|r^0(k) + r^n(k)\| > c\}. \tag{9}$$

Preprint submitted to 56th IEEE Conference on Decision and Control.
Received March 20, 2017.

B. Privacy Guarantees

In this subsection, we will show that the proposed consensus algorithm in Sec. III-A can guarantee the privacy of the initial states of benign agents under certain mild conditions. The result is summarized in Theorem 1.

Theorem 1. *The initial condition of any benign agent i can be estimated by a set of collaborative attackers $\{i_1, i_2, \cdots, i_m\}$ if and only if $N_i \cup \{i\} \subseteq N_{i_1} \cup \cdots \cup N_{i_m} \cup \{i_1, i_2, \cdots, i_m\}$.*

Due to space limit, we do not provide the detailed proof here, which essentially follows the procedure of the same proof as in [19]. We can view the set of collaborative attackers as a single *super-agent* who inherits from the set of attackers all the communication relationships with benign agents. The only difference is that the attacker in our work can add arbitrary signal into the system. However, the reduced system which ignores the influence of the attacker is identical to that in [19].

C. False Alarm Rate

Our goal in this subsection is to show that the proposed detector has a low false alarm rate.

Let $e(k) \triangleq x(k) - \hat{x}(k)$. Define

$$\tilde{A} \triangleq A - KCA, \tilde{B} \triangleq \Gamma - KC\Gamma,$$
$$\tilde{C} \triangleq CA, D \triangleq C\Gamma.$$

Similar to the residue $r(k)$, $x(k)$ and $e(k)$ can also be decomposed into three terms, respectively: $x^0(k)$ and $e^0(k)$, the state and estimate error driven by the initial value, $x^n(k)$ and $e^n(k)$, driven by noise, and $x^a(k)$ and $e^a(k)$, driven by the attacker's input. Then one verifies that

$$x^0(k+1) = Ax^0(k), \tag{10}$$
$$e^0(k+1) = \tilde{A}e^0(k), \tag{11}$$
$$r^0(k+1) = \tilde{C}e^0(k), \tag{12}$$

$$x^n(k+1) = Ax^n(k) + Aw(k) \tag{13}$$
$$e^n(k+1) = \tilde{A}e^n(k) + \tilde{A}w(k) - KCw(k+1) \tag{14}$$
$$r^n(k+1) = \tilde{C}e^n(k) + \tilde{C}w(k) + Cw(k+1) \tag{15}$$

$$x^a(k+1) = Ax^a(k) + \Gamma u(k), \tag{16}$$
$$e^a(k+1) = \tilde{A}e^a(k) + \tilde{B}u(k), \tag{17}$$
$$r^a(k+1) = \tilde{C}e^a(k) + Du(k), \tag{18}$$

where $x^0(0) = x(0)$, $e^0(0) = x(0) - KC(x(0) + w(0))$, $x^n(0) = e^n(0) = 0$ and $x^a(0) = e^a(0) = 0$.

By (11) and (12), we have

$$r^0(k) = C(A - AKC)^k x(0). \tag{19}$$

And, by (14) and (15), we have

$$r^n(k) = C\sum_{t=0}^{k}(A - AKC)^{k-t}w(t). \tag{20}$$

Define $\bar{A} \triangleq A - AKC$ and $\gamma(\bar{A})$ as the spectral radius of $\bar{A}$. Since the filter gain K is chosen such that $\bar{A}$ can be diagonalizable, we have $\bar{A} = V\Lambda V^{-1}$, where Λ is a diagonal matrix. Without loss of generality, the initial state $x(0)$ can be scaled such that $\|x(0)\| \leq 1$. We are now ready to state the main result in this subsection, the proof of which is reported in the appendix for the sake of legibility.

Theorem 2. *If $\rho > \max\{\varphi, \gamma(\bar{A})\}$, then the false alarm rate of the residue detector is*

$$\alpha \leq \frac{\beta}{(c - c_0)^2}$$

where

$$c_0 = \|V\| \, \|V^{-1}\|$$

and

$$\beta = \frac{\rho^2}{\rho^2 - \varphi^2} \operatorname{tr}\left\{C\left[I + \rho^{-2}(\bar{A} - I)X(\bar{A} - I)^T\right]C^T\right\}$$

X is the unique and finite solution of the Lyapunov equation $(\rho^{-1}\bar{A})X(\rho^{-1}\bar{A})^T - X + I = 0$.

From Theorem 2, we know that the false alarm rate can be arbitrarily small by tuning c.

D. Characterizing the Influence of Malicious Attacker via Reachable Set Computation

In this section, we characterize the effectiveness of our detector when the system is under attack. In particular, we use the concept of reachable set to provide a sufficient and necessary condition under which the consensus is reached, if the detector proposed in Sec. III-A, running on one benign agent, does not raise an alarm along the whole horizon.

To proceed, notice that we have the following result on the evolution of $x^0(k)$ and $x^n(k)$. Let

$$\bar{x} \triangleq \frac{\mathbf{1}^T x(0)}{n}\mathbf{1}.$$

Theorem 3 ([20])**.** *For any initial condition $x(0)$,*

$$\lim_{k\to\infty} \mathbb{E}[x^0(k) + x^n(k) - \bar{x}]^T[x^0(k) + x^n(k) - \bar{x}] = 0.$$

Furthermore, this convergence is exponentially fast.

1) Reachable Set of $e^a(k)$: By triangular inequality, if the following inequality holds

$$\sup_k \rho^{-k}\|r^a(k)\| > c + \sup_k \rho^{-k}\|r^0(k)\| + \sup_k \rho^{-k}\|r^n(k)\|,$$

then an alarm will be triggered at some time k. Therefore, in order to avoid detection, the adversary must enforce $\sup_k \rho^{-k}\|r^a(k)\|$ to be bounded, which implies that

$$\|r^a(k)\| \leq c_a\rho^k, \, \forall k,$$

where c_a is a design parameter for the adversary. A smaller c_a will result in lower detection rate, but it implies that the influence of the adversary on the consensus will also be smaller. [1]

Due to linearity, we assume $c_a = 1$ in order to simplify the notation.

[1] By Theorem 2, one verifies that the probability that the detector does not raise an alarm along the horizon is greater than $1 - \beta/(c - c_0 - c_a)^2$.

Preprint submitted to 56th IEEE Conference on Decision and Control.
Received March 20, 2017.

The attacker's strategy can be modeled as a constrained control problem in which the state is given in (16) and the constraint is imposed by

$$\|r^a(k)\| \leq \rho^k. \tag{21}$$

Let $e^a(:) \triangleq (e^a(0), e^a(1), \ldots)$ and $u(:) \triangleq (u(0), u(1), \ldots)$. We define $e^a(:), u(:)$ to be feasible if $e^a(:), u(:)$ satisfy (16) and (21) for all $k = 0, 1, \ldots$. We then define the reachable region of $\tilde{x}^a(k)$ as:

Definition 1. *The reachable region $\mathcal{R}_k$ of $e^a(k)$ is defined as*

$$\mathcal{R}_k = \{e \in \mathbb{R}^n : e = e^a(k) \text{ for some feasible } e^a(:) \text{ and } u(:)\}.$$

In the following, we focus on the characterization of $\mathcal{R}_k$, especially $\mathcal{R}_\infty$. To this end, define

Definition 2. *A set $\mathcal{S} \subseteq \mathbb{R}^n$ is ρ-invariant if for any $e \in \mathcal{S}$, there exist an $u \in \mathbb{R}^n$, such that*

$$\tilde{A}e + \tilde{B}u \in \rho\mathcal{S},\ \|\tilde{C}e + Du\| \leq 1.$$

The following theorem presents some important properties of a ρ-invariant set. In particular, any feasible $e^a(:), u(:)$ is related to the largest ρ-invariant set.

Theorem 4. *The following statements are true:*

1) *The union of any collection of ρ-invariant sets is still a ρ-invariant set.*
2) *There exists a largest ρ-invariant set, denoted as S_*.*
3) *There exists at least one ρ-invariant set.*
4) *Any feasible $e^a(:), u(:)$ satisfies*

$$e^a(k) \in \rho^k \mathcal{S}_*.$$

Proof: The first two statements are trivial. The third one can be easily proved by noticing that $\{0\}$ is ρ-invariant. We thus focus on the fourth one. Suppose that $e^a(:), u(:)$ is feasible. Let us define the following set

$$\mathcal{X}_k = \{e^a(t),\ t \geq k\},$$

Consider the following set

$$\mathcal{X} = \bigcup_{k=0}^{\infty} \rho^{-k} \mathcal{X}_k.$$

For any $e \in \mathcal{X}$, it must be of the following form:

$$e = \rho^{-k} e^a(t),$$

where $0 \leq k \leq t$. Now let us define $u = \rho^{-k} u(t)$. Hence,

$$\tilde{A}e + \tilde{B}u = \rho^{-k} e^a(t+1) = \rho\left(\rho^{-k-1} e^a(t+1)\right) \in \rho\mathcal{X},$$

and

$$\|\tilde{C}e + Du\| = \|\rho^{-k} r^a(t+1)\| \leq \rho^{-k}\rho^{t+1} \leq 1.$$

Therefore, $\mathcal{X}$ is ρ-invariant. Since $\mathcal{S}_*$ is the largest ρ-invariant set, $\rho^{-k} e^a(k) \in \mathcal{S}_*$, which implies that $e^a(k) \in \rho^k \mathcal{S}_*$. ∎

We then show how we can compute $\mathcal{S}_*$ recursively. To this end, define

Definition 3. *Given a set $\mathcal{S} \subset \mathbb{R}^n$, define* $\mathrm{Pre}(\mathcal{S})$ *as*

$$\mathrm{Pre}(\mathcal{S}) = \{e \in \mathbb{R}^n : \exists u \in \mathbb{R}^n \text{ such that } \tilde{A}e + \tilde{B}u \in \rho\mathcal{S},\ \|\tilde{C}e + Du\| \leq 1\}.$$

Then similar to Proposition 1 in [20], we have the following result:

Lemma 1. *Let $\mathcal{S}_0 = \mathbb{R}^n$ and $\mathcal{S}_{k+1} = \mathrm{Pre}(\mathcal{S}_k)$. Then the following holds:*

$$\mathcal{S}_\infty = \mathcal{S}_*. \tag{22}$$

To compute $\mathcal{R}_k$, we need the following definition.

Definition 4. *Given a set $\mathcal{S} \subset \mathbb{R}^n$, define* $\mathrm{Rch}(\mathcal{S}, k)$ *with $k = 0, 1, \ldots$ as*

$$\mathrm{Rch}(\mathcal{S}, k) = \{e' \in \mathbb{R}^n : \exists u \in \mathbb{R}^n, e \in \mathcal{S} \text{ such that } \tilde{A}e + Bu = e',\ \|\tilde{C}e + Du\| \leq \rho^k\}.$$

Then similar to Theorem 2 in [20], we have

Theorem 5. *The followings hold:*

- $\mathcal{R}_0 = \{0\}$.
- $\mathcal{R}_{k+1} = \mathrm{Rch}(\mathcal{R}_k, k+1) \cap \rho^{k+1}\mathcal{S}_*$, *for every* $k = 0, 1, \ldots$

2) Ellipsoidal Approximation: Since it is not tractable to numerically compute the $\mathcal{S}_*$ and $\mathcal{R}_k$ as k goes to infinity, in the following we consider their ellipsoidal approximations.

Given a positive semi-definite matrix $M \in \mathbb{S}_+^n$, define the ellipsoid $\mathcal{E}(M)$ as

$$\mathcal{E}(M) = \{x \in \mathbb{R}^n : x^T M x \leq 1\}.$$

Then similar to [20], we have the following result on the inner and outer ellipsoidal approximations of the operators Pre.

Theorem 6. *Let $M \in \mathbb{S}_+^n$, then the followings hold:*

$$\mathcal{E}(M_{in}^p) \subset \mathrm{Pre}(\mathcal{E}(M)) \subset \mathcal{E}(M_{out}^p), \tag{23}$$
$$\mathcal{E}(M_{in}^r) \subset \mathrm{Rch}(\mathcal{E}(M), k) \subset \mathcal{E}(M_{out}^r), \tag{24}$$

where

$$M_{in}^p = f(M), M_{out}^p = f(M)/2, \tag{25}$$
$$M_{in}^r = g(M,k), M_{out}^r = g(M,k)/2 \tag{26}$$

with

$$f(M) \triangleq \rho^{-2}\tilde{A}^T M \tilde{A} + \tilde{C}^T\tilde{C} - (\rho^{-2}\tilde{A}^T M\tilde{B} + \tilde{C}^T D)(\rho^{-2}\tilde{B}^T M\tilde{B} + D^T D)^+(\rho^{-2}\tilde{B}M\tilde{A} + D^T\tilde{C}),$$
$$g(M,k) \triangleq \hat{A}^T M\hat{A} + \rho^{-k}\hat{C}^T\hat{C} - \hat{A}^T M\hat{B} + \rho^{-k}\hat{C}^T\hat{D})(\hat{B}^T M\hat{B} + \rho^{-k}\hat{D}^T\hat{D})^+(\hat{B}M\hat{A} + \rho^{-k}\hat{D}^T\hat{C}).$$

The matrices $\hat{A}, \hat{B}, \hat{C}, \hat{D}$ are given by

$$\hat{A} = \tilde{A}^+,\ \hat{B} = [-\tilde{A}^+\tilde{B}, I_{2n} - \hat{A}^+\hat{A}],$$
$$\hat{C} = \tilde{C}\tilde{A}^+,\ \hat{D} = [D - \tilde{C}\tilde{A}^+B, \tilde{C} - \tilde{C}\tilde{A}^+\tilde{A}].$$

We then define recursively the following inner and outer ellipsoidal approximations of $\mathcal{S}_k$.

$$\mathcal{S}_{in}(k+1) = f(\mathcal{S}_{in}(k)),$$
$$\mathcal{S}_{out}(k+1) = f(\mathcal{S}_{out}(k))/2,$$
$$\mathcal{R}_{in}(k+1) = g(\mathcal{R}_{in}(k), k+1) + \mathcal{S}_{in}(\infty),$$
$$\mathcal{R}_{out}(k+1) = [g(\mathcal{R}_{out}(k), k+1)/2 + \mathcal{S}_{in}(\infty)]/2$$

Preprint submitted to 56th IEEE Conference on Decision and Control.
Received March 20, 2017.

with $\mathcal{S}_{in}(0) = \mathcal{S}_{out}(0) = 0$. and $\mathcal{R}_{in}(0) = \mathcal{R}_{out}(0) = \infty$. It can be verified that the followings hold:

$$\mathcal{E}(\mathcal{S}_{in}(k)) \subset \mathcal{S}_k \subset \mathcal{E}(\mathcal{S}_{out}(k)), \tag{27}$$

$$\mathcal{E}(\mathcal{R}_{in}(k)) \subset \mathcal{R}_k \subset \mathcal{E}(\mathcal{R}_{out}(k)). \tag{28}$$

3) Analysis of Malicious Effects on Consensus: We first have the following proposition relating triviality of $\mathcal{R}_\infty$ to the matrices $\tilde{A}, \tilde{B}, \tilde{C}, D$. Define the following pencil

$$P(z) = \begin{bmatrix} \tilde{A} - zI & \tilde{B} \\ \tilde{C} & D \end{bmatrix} \tag{29}$$

where z is a complex number.

Theorem 7. *If the pencil $P(z)$ has full column rank for all $|z| \geq 1$, then there holds*

$$\lim_{k\to\infty} \mathcal{R}_k = \{0\}. \tag{30}$$

Proof: Due to space limitation, we do not present the proof here, which, however, can be found in [21]. ■

The following theorem states that, as long as the detector is not triggered for all k, the state estimate $\hat{x}(k)$ will converge to consensus.

Theorem 8. *If $\|r(k)\| < c\rho^k$ for all k, we have*

$$\lim_{k\to\infty} (\hat{x}(k) - \frac{\mathbf{1}\mathbf{1}^T}{n}\hat{x}(k)) = 0$$

Proof:

$$\begin{aligned}\hat{x}(k) =& A\hat{x}(k-1) + Kr(k) \\ =& A^k\hat{x}(0) + \sum_{t=1}^{k} A^{k-t}Kr(t)\end{aligned}$$

Multiplying both sides by $I - \frac{\mathbf{1}\mathbf{1}^T}{n}$ yields

$$\begin{aligned}&\|(I - \frac{\mathbf{1}\mathbf{1}^T}{n})\hat{x}(k)\| \\ =& \|(A - \frac{\mathbf{1}\mathbf{1}^T}{n})^k\hat{x}(0) + \sum_{t=1}^{k}(A - \frac{\mathbf{1}\mathbf{1}^T}{n})^{k-t}Kr(t)\| \\ \leq& c\sum_{t=1}^{k}\|\rho^t(A - \frac{\mathbf{1}\mathbf{1}^T}{n})^{k-t}\|\ \|K\| \\ \leq& c(k-1)[\max(\rho, |\lambda_2|, |\lambda_n|)]^k\|K\|.\end{aligned}$$

Thus, we have

$$\lim_{k\to\infty} \|(I - \frac{\mathbf{1}\mathbf{1}^T}{n})\hat{x}(k)\| = 0$$

which implies

$$\lim_{k\to\infty} (\hat{x}(k) - \frac{\mathbf{1}\mathbf{1}^T}{n}\hat{x}(k)) = 0$$

■

We then have the following result on the consensus of the system state $x(k)$.

Theorem 9. *If the detector proposed in Sec. III-A does not raise an alarm along the whole horizon and the pencil $P(z)$ has full column rank for all $|z| \geq 1$, $x(k)$ achieves consensus in the mean squared sense, i.e.,*

$$\lim_{k\to\infty} \mathbb{E}\left\|(I - \frac{\mathbf{1}\mathbf{1}^T}{n})x(k)\right\|^2 = 0. \tag{31}$$

Proof: By Theorem 7, the two conditions **NoAlarm** and **NoUnstableZeros** imply that

$$\lim_{k\to\infty} e^a(k) = 0.$$

Using similar argument as the proof of Theorem 2, we can conclude that $\lim_{k\to\infty} e^0(k) = 0$ and $\lim_{k\to\infty} \mathbb{E}\|e^n(k)\|^2 = 0$. Thus, $e(k)$ converges to 0 in the mean squared sense. By Theorem 8, $\hat{x}(k)$ will achieve consensus if no alarm is triggered, which implies that $x(k)$ must achieve consensus since $e(k)$ converges to 0. ■

IV. Numerical Examples

Consider the following network

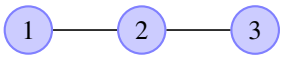

Fig. 1. Network Topology

Suppose that Agent 1 is an attacker, and Agent 2 is running the detector. The system is as follows:

$$A = \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0.25 & 0.25 \\ 0 & 0.25 & 0.75 \end{bmatrix}, C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The filter gain is chosen as

$$K = \begin{bmatrix} 0.9 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.5 \end{bmatrix}$$

We first verify the result that the false alarm rate can be very low by tuning the threshold $c\rho^k$. Suppose $\varphi = 0.2$. By Theorem 2, we know that $\rho > \max\{\varphi, \gamma(\bar{A})\} = 0.43$ and $c > 8.82$ can lead to a false alarm rate less than one. To obtain a low false alarm rate, here we choose $\rho = 0.5$ and $c = 15$. Through simulations, we find the false alarm rate is around 1%. One snapshot of the comparison between $\|r(k)\|$ and $c\rho^k$ is shown in Fig. 2.

In Fig. 3, we plot the minimum eigenvalue of the outer ellipsoid approximation $\mathcal{R}_{out}$. It can be seen that this minimum eigenvalue is monotonically increasing. Therefore[2], one concludes that $\mathcal{R}_k \to \{0\}$.

V. Conclusion

In this paper, we deal with curious and malicious attacks in an average consensus system. A privacy-preserving average consensus algorithm equipped with a malicious attack detector is designed for each benign agent. We show that the privacy of initial states of all benign agents are guaranteed under mild conditions. Moreover, the false alarm rate of the detector can

[2]Notice that the radius of an ellipsoid $\mathcal{E}(M)$ is the inverse of the minimum eigenvalue of M.

Preprint submitted to 56th IEEE Conference on Decision and Control.
Received March 20, 2017.

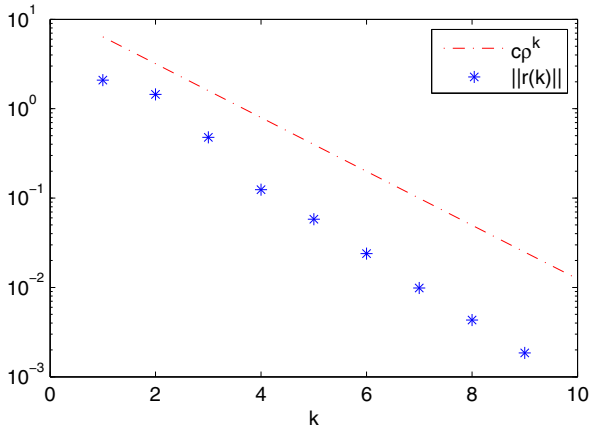

Fig. 2. A snapshot of comparison between $\|r(k)\|$ and $c\rho^k$

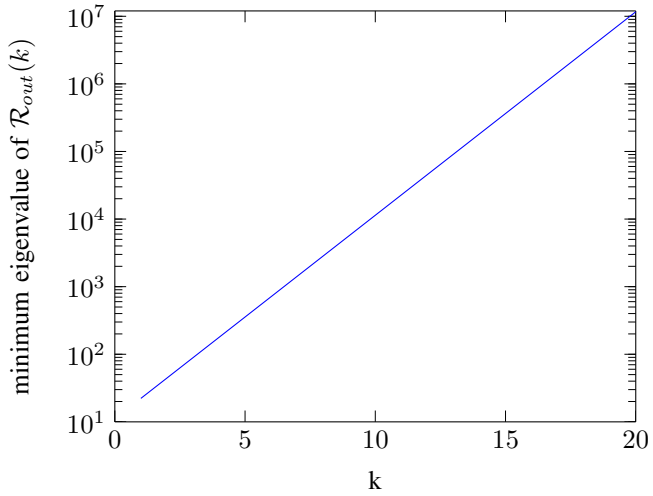

Fig. 3. Minimum eigenvalue of $\mathcal{R}_{out}(k)$.

be arbitrarily low by choosing proper time-varying thresholds. We also prove that the states of all agents can reach a consensus if the alarm is not triggered. As a future research direction, we will study accuracy of the final consensus value versus the average of initial states.

REFERENCES

[1] L. Schenato and G. Gamba, "A distributed consensus protocol for clock synchronization in wireless sensor network," in *Proceedings of the 46th IEEE Conference on Decision and Control*, 2007, pp. 2289–2294.

[2] G. Cybenko, "Dynamic load balancing for distributed memory multiprocessors," *Journal of parallel and distributed computing*, vol. 7, no. 2, pp. 279–301, 1989.

[3] J. Lin, A. S. Morse, and B. D. O. Anderson, "The multi-agent rendezvous problem," in *Proceedings of the 42th IEEE Conference on Decision and Control*, 2003, pp. 1508–1513.

[4] W. Ren, R. W. Beard, and E. M. Atkins, "A survey of consensus problems in multi-agent coordination," in *Proceedings of the American Control Conference*, 2005, pp. 1859–1864.

[5] 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.

[6] W. Ren and R. W. Bear, *Distributed consensus in multi-vehicle cooperative control*. Springer, 2008.

[7] 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.

[8] W. Ren and R. W. Beard, "Consensus seeking in multiagent systems under dynamically changing interaction topologies," *IEEE Transactions on Automatic Control*, vol. 50, no. 5, pp. 655–661, 2005.

[9] M. H. DeGroot, "Reaching a consensus," *Journal of the American Statistical Association*, vol. 69, no. 345, pp. 118–121, 1974.

[10] J. N. Tsitsiklis, "Problems in decentralized decision making and computation," Ph.D. dissertation, Massachusetts Institute of Technology, 1984.

[11] A. Jadbabaie, J. Lin, and A. Morse, "Coordination of groups of mobile autonomous agents using nearest neighbor rules," *IEEE Transactions on Automatic Control*, vol. 48, no. 6, pp. 988–1001, 2003.

[12] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Systems & Control Letters*, vol. 53, no. 1, pp. 65–78, 2004.

[13] A. Nedic, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, "On distributed averaging algorithms and quantization effects," *IEEE Transactions on Automatic Control*, vol. 54, no. 11, pp. 2506–2517, 2009.

[14] S. Sundaram and C. N. Hadjicostis, "Distributed function calculation via linear iterative strategies in the presence of malicious agents," *IEEE Transactions on Automatic Control*, vol. 56, no. 7, pp. 1495 – 1508, 2011.

[15] F. Pasqualetti, A. Bicchi, and F. Bullo, "Distributed intrusion detection for secure consensus computations," in *Proceedings of the 46th IEEE Conference on Decision and Control*, 2007, pp. 5594–5599.

[16] ——, "Consensus computation in unreliable networks: A system theoretic approach," *IEEE Transactions on Automatic Control*, vol. 57, no. 1, pp. 90–104, 2012.

[17] Z. Huang, S. Mitra, and G. Dullerud, "Differentially private iterative synchronous consensus," in *Proceedings of the 2012 ACM workshop on Privacy in the electronic society*, 2012, pp. 81–90.

[18] N. Manitara and C. N. Hadjicostis, "Privacy-preserving asymptotic average consensus," in *Proceedings of the 12th European Control Conference*, 2013, pp. 760–765.

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

[20] Y. Mo and B. Sinopoli, "On the performance degradation of cyber-physical systems under stealthy integrity attacks," *IEEE Transactions on Automatic Control*, vol. 61, no. 9, pp. 2618 – 2624, 2016.

[21] Q. Liu, X. Ren, and Y. Mo, "Secure and privacy preserving average consensus (http://yilinmo.github.io/papers/cdc17-2.html)," Tech. Rep.

APPENDIX A
PROOF OF THEOREM 2

In order to prove Theorem 2, we first need the following lemmas.

Lemma 2. *For all* $k \in \{0, 1, \cdots\}$,

$$\sup_k \|\rho^{-k} r^0(k)\| \leq c_0$$

Proof:

$$\begin{aligned}\sup_k \|\rho^{-k} r^0(k)\| &\leq \sup_k \|\rho^{-k} C(A - AKC)^k x(0)\| \\ &\leq \sup_k \|\rho^{-k} \bar{A}^k\| \ \|C\| \ \|x(0)\|.\end{aligned}$$

Note that $\|x(0)\|$ is scaled to be 1. The matrix C defined in (6) also has its 2-norm as 1.

$$\begin{aligned}\sup_k \|\rho^{-k} \bar{A}^k\| &= \sup_k \|\rho^{-k} V \Lambda^k V^{-1}\| \\ &\leq \sup_k \|(\rho^{-1}\Lambda)^k\| \ \|V\| \ \|V^{-1}\| \\ &\leq \|V\| \ \|V^{-1}\| = c_0.\end{aligned}$$

The last inequality holds because $\rho > \gamma(\bar{A})$. Thus, the final result is obtained. ■

Lemma 3. *For all* $k \in \{0, 1, \cdots\}$,

$$P\left(\sup_k \|\rho^{-k} r^n(k)\| \geq \epsilon\right) \leq \frac{1}{\epsilon^2}\beta$$

Preprint submitted to 56th IEEE Conference on Decision and Control.
Received March 20, 2017.

Proof: The residue generated by noise signals can be rewritten as:

$$\begin{aligned}
&r^n(k)\\
&= C\{\bar{A}^k v(0) + \bar{A}^{k-1}[\varphi v(1) - v(0)] + \cdots\\
&\quad + [\varphi^k v(k) - \varphi^{k-1} v(k-1)]\}\\
&= C[\sum_{t=1}^{k} \varphi^{k-t}(\bar{A}^t - \bar{A}^{t-1} v(k-t)) + \varphi^k v(k)],\ (k \geq 1)\\
&\quad r^n(0) = Cv(0).
\end{aligned}$$

Note that $r^n(k)$ is a random variable. Next we compute its upper bound associated with a probability, using Markov inequality.

$$\begin{aligned}
&P\left(\|\rho^{-k} r^n(k)\| \geq \epsilon\right)\\
&= P\left(\rho^{-2k}(r^n(k))^T r^n(k) \geq \epsilon^2\right)\\
&\leq \frac{1}{\epsilon^2} E\left[\rho^{-2k}(r^n(k))^T r^n(k)\right]\\
&= \frac{1}{\epsilon^2}\left(\frac{\varphi}{\rho}\right)^{2k} \operatorname{tr}\{C[\sum_{t=1}^{k} \varphi^{-2t}(\bar{A}^t - \bar{A}^{t-1})(\bar{A}^t - \bar{A}^{t-1})^T\\
&\quad + I]C^T\},\ (k > 0)\\
&\quad P\left(\|r_n(0)\| \geq \epsilon\right) \frac{1}{\epsilon^2} \operatorname{tr}\left\{CC^T\right\}
\end{aligned}$$

Then for all k,

$$\begin{aligned}
&P\left(\sup_k \|\rho^{-k} r^n(k)\| \geq \epsilon\right)\\
&= P\left(\bigcup_k \left\{\|\rho^{-k} r^n(k)\| \geq \epsilon\right\}\right)\\
&\leq \frac{1}{\epsilon^2} \operatorname{tr}\left\{C\left[I + (\rho^{-2}(\bar{A} - I)(\bar{A} - I)^T + (\rho^{-1}\varphi)^2 I)\right.\right.\\
&\quad \left.\left. + \cdots\right] C^T\right\}\\
&= \frac{1}{\epsilon^2} \frac{\rho^2}{\rho^2 - \varphi^2} \operatorname{tr}\left\{C\left[I + \rho^{-2}(\bar{A} - I)(I + \rho^{-2}\bar{A}\bar{A}^T\right.\right.\\
&\quad \left.\left. + \rho^{-4}\bar{A}^2(\bar{A}^T)^2 + \cdots)(\bar{A} - I)^T\right] C^T\right\}\\
&= \frac{1}{\epsilon^2} \frac{\rho^2}{\rho^2 - \varphi^2} \operatorname{tr}\left\{C\left[I + \rho^{-2}(\bar{A} - I)X(\bar{A} - I)^T\right] C^T\right\}\\
&= \frac{\beta}{\epsilon^2}
\end{aligned}$$

■

We are now ready to prove Theorem 2:

Proof:

$$\begin{aligned}
\alpha &= P\left(\sup_k \|\rho^{-k} r^0(k) + \rho^{-k} r^n(k)\| \geq c\right)\\
&\leq P\left(\sup_k \|\rho^{-k} r^n(k)\| \geq c - c_0\right)\\
&\leq \frac{\beta}{(c - c_0)^2}
\end{aligned}$$

■

Preprint submitted to 56th IEEE Conference on Decision and Control.
Received March 20, 2017.