This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSIPN.2016.2614898, IEEE Transactions on Signal and Information Processing over Networks

Data injection attacks in randomized gossiping

Reinhard Gentz, Sissi Xiaoxiao Wu, Hoi-To Wai, Anna Scaglione, Amir Leshem

***Abstract*—The subject of this paper is the detection and mitigation of data injection attacks in randomized average consensus gossip algorithms. It is broadly known that the main advantages of randomized average consensus gossip are its fault tolerance and distributed nature. Unfortunately, the flat architecture of the algorithm also increases the attack surface for a data injection attack. Even though we cast our problem in the context of sensor network security, the attack scenario is identical to existing models for opinion dynamics (the so called DeGroot model) with stubborn agents steering the opinions of the group towards a final state that is not the average of the network initial states. We specifically propose two novel strategies for detecting and locating attackers, and study their detection and localization performance numerically and analytically. Our detection and localization methods are completely decentralized and, therefore nodes can directly act on their conclusions and stop receiving information from nodes identified as attackers. As we show by simulation the network can often recover in this fashion, leveraging the resilience of randomized gossiping to reduced network connectivity.**

***Index Terms*—data injection attack, attack detection, decentralized learning, randomized gossip protocol**

I. Introduction

The key advantage of gossip-based algorithms is the built-in fault tolerance to node failures, as nodes can reorganize themselves automatically. To prevent interference from unauthorized nodes, authentication and encryption methods can be used (see e.g. [2], [3]). However, in the case of an *insider attack* gossip-based algorithms are highly vulnerable, even if only one node is compromised. In fact, the flat, self-organizing architecture, which is the selling feature for these algorithms, can become a liability.

In this paper, we consider the randomized average consensus gossiping protocol introduced in [4] and focus on the insider attack scenario, where authentication and encryption have failed. Noting that average consensus gossiping is equivalent to the DeGroot model dynamics [5], the attack scenarios are identical to models that have emerged for the study of stubborn agents or zealots in social networks (see e.g. [6]–[11]). We propose decentralized strategies which aim at detecting and localizing insider attackers by analyzing the statistics of the nodes' states, as the nodes in the network perform the algorithm several times starting from different initial conditions. While in general data injection from stubborn agents will not lead to a consensus [6], [12], in our paper we model the attackers in the randomized consensus algorithm as a group of coordinated agents that are trying to steer the consensus state to a value of their choosing, while hiding their nature by judiciously preserving the expected exponential convergence rate [13] and leading the network to consensus towards a desired final state (see Section II-A). This is the worst case scenario, since the network will still converge to consensus, but to the wrong state.

Specifically, we refer to the proposed strategies as the *time difference* and *spatial difference* methods respectively and refer to nodes that are not attackers as *normal* nodes. Our methods are fully decentralized and hence each normal node can detect and localize neighboring attackers independently. Moreover, the spatial strategy can even detect an attack which is not a direct neighbor of the sensing node. Once a normal node detects and localizes an attacker, it can report the anomaly to a central authority or, alternatively, cut future communication with the attacker. Eventually, the proposed algorithms isolate all the attackers from the network, thereby preventing future harm to the whole system. It is worth mentioning that each node only needs to collect its local statistical information by evaluating messages transmitted by nodes in the neighborhood as the protocol is executed. Therefore, no additional communication overhead is required in the two proposed strategies.

A. Prior art on data-injection in gossip based algorithms

To the best of our knowledge, algorithms to detect and mitigate insider attacks in gossip-based networks have received limited attention so far. Exceptions are [14], [15] and the very recent submission [19]. In [15] the authors propose to detect injection attacks using a likelihood ratio test that is appropriate for synchronous average consensus, but not for its more popular asynchronous implementation, while [19] proposes to discard neighbors state values that are extreme (maxima or minima), given that malicious agents do not average their state with that of normal nodes. Reference [14], instead, proposes two protection schemes for randomized consensus algorithms. The first one is motivated by the fact that the convergence speed is usually slower in the presence of an attacker. Thus, a data injection attack can be spotted by detecting possible anomalies in the convergence speed, which has an exponential trend. Note that, the "normal" convergence speed can be estimated only if we have prior knowledge of the underlying physical model; e.g., see [4], [13], [16]. The second scheme in [14] is based on using cryptographic signatures. However, for the detection of the attacker a node needs to be surrounded by a majority of normal neighbors. Furthermore, the cryptographic solution does not detect unintentional bias at a node, which can result in similar catastrophic results. In this work, we consider a data injection attack model in which the

This material is based upon work supported by NSF CCF-1011811, CREDC DOE grant DE-OE0000780 and ISF grant 903/2013. Preliminary results have been presented in [1].

R. Gentz, S. X. Wu, H.-T. Wai and A. Scaglione are with the School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85281, USA. E-mails: {rgentz,xiaoxi12,htwai,Anna.Scaglione}@asu.edu. A. Leshem is with Faculty of Engineering, Bar-Ilan University, Ramat-Gan, Israel. Email: leshema@eng.biu.ac.il.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSIPN.2016.2614898, IEEE Transactions on Signal and Information Processing over Networks

attackers are insiders, they are possibly coordinated and also deceive their neighbors by following an expected convergence rate. Therefore, the first detection scheme in [14] is not even applicable to our target problem. We propose two detection schemes that are based on computing two metrics at each node whose high values are indicators of a possible attack. The metrics are computed at each node locally, overhearing the messages exchanged in the neighborhood over several instances of average consensus.

This paper is organized as follows. In Section II, we describe the pairwise randomized consensus algorithm and introduce the data injection attack model. In Section III, we propose the detection and localizing strategies for eliminating the attackers. The performance analysis for the proposed strategies are analyzed in Section IV. We conclude with simulation results in Section V.

Notations: We use boldfaced letter to denote vector/matrix. For a vector $\boldsymbol{x}$, $[\boldsymbol{x}]_i$ denotes its ith element; similarly, for a matrix $\boldsymbol{A}$, $[\boldsymbol{A}]_{ij}$ denotes its (i,j)th element.

II. Consensus Network Model

Let us consider a sensor network, which is described by a connected, undirected graph $G = (V, E)$, where $V = \{1, ..., n\}$ denotes a set of nodes and $E \subseteq V \times V$ denotes the connections between the nodes. Assume that the sensor nodes continuously perform a randomized consensus algorithm. We assume that at each iteration of the detection algorithm, a total of K instances of the consensus algorithm have taken place, either running in parallel or sequentially, and each node has overheard many, if not all, transmissions in its neighborhood and accrued historical data of K runs of the consensus algorithm from its vantage point, which we enumerate in this paper with the superscript k, with $k \in \{1, \dots, K\}$. In our notation $k \in \mathbb{N}$ is used as a superscript in reference to the instance of the consensus algorithm and the time index $t \in \mathbb{N}$ denotes the specific iteration. Correspondingly, the random vector $\boldsymbol{x}^k(t) = (x_1^k(t), ..., x_n^k(t))^\top \in \mathbb{R}^n$ represent the states at the tth consensus iteration[1].

Let the initial state of node $i \in V$ be $x_i^k(0) = \gamma_i^k$. The goal of the consensus algorithm is to compute the network initial states' average

$$x_{av}^k := \frac{1}{n}\mathbf{1}^\top \boldsymbol{x}^k(0) = \frac{1}{n}\mathbf{1}^\top \boldsymbol{\gamma}^k, \tag{1}$$

where $\mathbf{1}$ is an all-one vector. The consensus algorithm we consider in this work is the random pairwise exchange algorithm [4], shown in Algorithm 1. We remark that the non-negative parameter P_{ij}, which is probability that node i selects node j to update with, in Algorithm 1, satisfies $\sum_{j=1}^n P_{ij} = 1$ and Algorithm 1 can be implemented asynchronously. Each sensor node does not need to know the iteration index t of the protocol. The updates in Algorithm 1 can be conveniently expressed as:

$$\boldsymbol{x}^k(t) = \boldsymbol{W}(t-1)\boldsymbol{x}^k(t-1), \tag{4}$$

[1]One can assume a random waiting time between updates, for instance draw from i.i.d. exponential distributions [4]

Algorithm 1: Randomized consensus protocol

Input: no. of iterations T, initial states: $x_i^k(0)\ \forall\ i \in V$.

for $t = 1 : T$ **do**

- Uniformly wake up a random node $i \in V$.
- Node i selects node j from its neighborhood with the probability

$$P_{ij}, \text{ where } j \in \mathcal{N}_i \text{ and } \mathcal{N}_i := \{j : (i,j) \in E\}. \tag{2}$$

- Node i and j update their states as follows

$$x_i^k(t+1) = x_j^k(t+1) = \frac{x_i^k(t) + x_j^k(t)}{2}; \tag{3}$$

Other nodes keep their original states, i.e., $x_v^k(t+1) = x_v^k(t)$ for all $v \neq i, j$.

where $\boldsymbol{W}(t)$ is the transition matrix at instance k and time t. Define $[\boldsymbol{P}]_{ij} = P_{ij}$ and $\boldsymbol{\Sigma}$ as a diagonal matrix with $[\boldsymbol{\Sigma}]_{ii} = \sum_{j=1}^n (P_{ij} + P_{ji})$, the expected transition matrix can be written as

$$\overline{\boldsymbol{W}} = \mathbb{E}\left[\boldsymbol{W}(t)\right] = \boldsymbol{I} - \frac{1}{2n}\boldsymbol{\Sigma} + \frac{\boldsymbol{P} + \boldsymbol{P}^\top}{2n}. \tag{5}$$

It can be verified that $\overline{\boldsymbol{W}}$ is non-negative, symmetric and doubly stochastic. We have the expected states

$$\mathbb{E}[\boldsymbol{x}^k(t)|\boldsymbol{x}^k(0)] = \overline{\boldsymbol{W}}\mathbb{E}\left[\boldsymbol{x}^k(t-1)|\boldsymbol{x}^k(0)\right] = \overline{\boldsymbol{W}}^t\boldsymbol{x}^k(0),$$

Under some mild assumptions, the protocol above finds the true average x_{av}^k. Denote $\lambda_2(\overline{\boldsymbol{W}})$ as the second largest eigenvalue of $\overline{\boldsymbol{W}}$, we have

Fact 1 *For each k, the state at every sensor $i \in V$ converges to a Δ-neighborhood of x_{av}^k with a high probability, i.e.*

$$\mathbb{P}\Big(|x_i^k(t) - x_{av}^k| < \Delta \max_{j \in V} |x_j^k(0)|\Big) \geq 1 - \Delta, \tag{6}$$

for all $\Delta \geq 0$ and $t \geq 3\log\Delta^{-1}/\log\lambda_2(\overline{\boldsymbol{W}})^{-1}$.

The detailed proof and conditions of Fact 1 can be found in [4, Theorem 3]. Notice that the lower bound on t is finite only if $\lambda_2(\overline{\boldsymbol{W}}) < 1$, which depends on the design of $\boldsymbol{P}$ and thus can be satisfied when G is a connected graph; see [13] for further discussions.

A. Data Injection Attack Model

The data injection attack model we consider in this paper is analogous to the *stubborn agent* model studied under the framework of DeGroot opinion dynamics in social learning [5], whose average convergence properties were studied in [6]. We assume that the sensor network is compromised by a set of attackers, denoted by $V_s \subseteq V$. For simplicity, we set $V_s = \{1, ..., n_s\}$ and $n_s \leq n$. The remaining normal nodes form the set $V_r = V \setminus V_s$. The goal of the attackers (or malicious nodes) is to steer the consensus result of the network to a certain target value of their choice $\alpha^k \neq x_{av}^k$, so that the states converge to

$$\lim_{t \to \infty} \boldsymbol{x}^k(t) = \alpha^k \mathbf{1}. \tag{7}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSIPN.2016.2614898, IEEE Transactions on Signal and Information Processing over Networks

To stage the attack, the malicious nodes follow a modified update rule. That is, under the consensus protocol of Algorithm 1, if a malicious node $j \in V_s$ is selected at iteration t, in lieu of (3), the node's state will be generated as

$$x_j^k(t) = \alpha^k + m_j^k(t), \tag{8}$$

where $m_j^k(t)$ is a zero-mean *artificial noise* generated by the attackers to hide their malicious intent from the normal nodes. Notice that a normal node could easily detect a malicious agent with no artificial noise as the node's state over time would be constant, $x_s^k(t) = x_s^k(0)$, and thus easily detectable.

We claim that under the modified update rule (8) and the assumptions that (i) the attackers are not isolated from the network and (ii) the induced sub-graph $G[V_r]$ is strongly connected, the attackers can successfully steer the consensus result of the network:

$$\lim_{t\to\infty} \mathbb{E}[\boldsymbol{x}^k(t)|\alpha^k] = \alpha^k \mathbf{1}. \tag{9}$$

To show (9), let us first define

$$\boldsymbol{x}^k(t) = \left(\begin{array}{cc} \boldsymbol{s}^k(t)^\top, & \boldsymbol{r}^k(t)^\top \end{array} \right)^\top, \tag{10}$$

where $\boldsymbol{s}^k(t) \in \mathbb{R}^{n_s}, \boldsymbol{r}^k(t) \in \mathbb{R}^{n-n_s}$ correspond to the states of the malicious nodes and normal nodes, respectively. As a consequence of the update rules in (3) and (8), we have

$$\mathbb{E}[\boldsymbol{s}^k(t)|\alpha^k] = \alpha^k \mathbf{1}, \ \forall\ t \geq 1. \tag{11}$$

Moreover, the expected transition matrix $\overline{\boldsymbol{W}}$ becomes

$$\overline{\boldsymbol{W}} = \mathbb{E}[W(t)] = \left(\begin{array}{cc} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{B} & \boldsymbol{D} \end{array} \right), \tag{12}$$

where the malicious nodes correspond to the identity matrix in $\overline{\boldsymbol{W}}$ as they are never affected by the other nodes. It is well known that [17, Theorem 1.1, Chapter 2]:

Fact 2 *If $\boldsymbol{D}$ is sub-stochastic and irreducible, then it holds that $\|\boldsymbol{D}\|_2 < 1$.*

Notice that $\boldsymbol{B} \neq \boldsymbol{0}$ as the attackers are not isolated, thus $\boldsymbol{D}$ is sub-stochastic. Moreover, $\boldsymbol{D}$ is irreducible as $G[V_r]$ is strongly connected. It can be verified that

$$\mathbb{E}[\boldsymbol{r}^k(t)|\boldsymbol{\gamma}^k, \alpha^k] = \alpha^k \textstyle\sum_{s=0}^t \boldsymbol{D}^{t-s}\boldsymbol{B}\mathbf{1} + \boldsymbol{D}^t\boldsymbol{\gamma}^k. \tag{13}$$

As $\|\boldsymbol{D}\|_2 < 1$ and using the identities $\sum_{t=1}^{n-1} \boldsymbol{D}^t = (\boldsymbol{I} - \boldsymbol{D}^n)(\boldsymbol{I} - \boldsymbol{D})^{-1}$ and $\boldsymbol{B}\mathbf{1} + \boldsymbol{D}\mathbf{1} = \mathbf{1}$,

$$\textstyle\sum_{s=0}^t \boldsymbol{D}^{t-s}\boldsymbol{B}\mathbf{1} = \mathbf{1} - \boldsymbol{D}^{t+1}\mathbf{1},$$

As $\boldsymbol{D}^t$ decays to zero, we have $\lim_{t\to\infty} \mathbb{E}[\boldsymbol{r}^k(t)|\boldsymbol{\gamma}^k, \alpha^k] = \alpha^k\mathbf{1}$. Combining this with (11) yields (9).

Remark 1 *From the normal nodes' point of view, the attackers appear to make progress towards the final value $\boldsymbol{x}^k(\infty)$. If there is no attacker, $\boldsymbol{x}^k(\infty)$ would be the true average of all nodes; in the presence of coordinated attackers, it will tend to α^k. At the same time, the attackers states converge with the expected convergence speed to α^k.*

Remark 2 *When there are multiple attackers in the network, we assume that they are coordinated such that all the malicious agents bias their state with the same value α^k. Otherwise, the network almost surely will not reach consensus [6], [18], and thus attacks can be detected by spotting different final states. Interestingly, a recent submission [19] proposed an alternative defense mechanism against attackers for* synchronous *gossiping, which is based on discarding extreme values in the consensus iteration. It is useful to notice that their method does not generally work with random gossiping and against the noisy coordinated attack technique we consider. Our methods, in contrast, can be applied very successfully against the attack model in [19] in which the malicious nodes simply do not average with their neighbors, as well as the more insidious one we consider, as we show by simulation in Section V.*

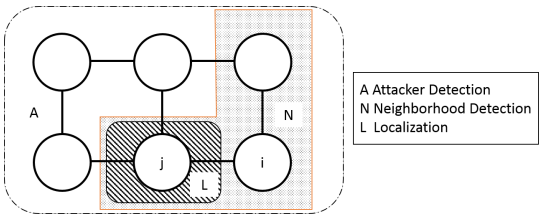

Fig. 1: Different tasks involved in the attack detection scheme.

III. Detecting Data Injection Attack

We consider *three* detection tasks, to be performed in a decentralized fashion by the normal nodes $i \in V_r$ in order.

(I) Attacker Detection Task — The first test checks if the presence of attacker(s) in the network:

$\mathcal{H}_0$: No attacker in the network, i.e., $V_s = \emptyset$,

$\mathcal{H}_1$: At least one attacker in the network, i.e., $V_s \neq \emptyset$.

(II) Neighborhood Detection Task — The second test checks if an attacker is present in the neighborhood of node i:

$\mathcal{H}_0^i$: No attacker in the neighborhood, i.e., $V_s \cap \mathcal{N}_i = \emptyset$,

$\mathcal{H}_1^i$: Attacker in the neighborhood, i.e., $V_s \cap \mathcal{N}_i \neq \emptyset$.

(III) Localization Task — The third test attempts to locate the malicious node in the neighborhood of node i. For all $j \in \mathcal{N}_i$, we check:

$\mathcal{H}_0^{ij}$: node j is not an attacker, i.e., $j \notin V_s$.

$\mathcal{H}_1^{ij}$: node j is an attacker, i.e., $j \in V_s$.

Note that the localization task is performed only if the neighborhood detection task decides that $\mathcal{H}_1^i$ is true. We depict the detection/localization targets of the three tasks in Figure 1.

In the case when a central authority (CA) exists, node i can report the tests' results to the CA. The CA will take appropriate actions, possibly fusing the information of multiple nodes. We also propose the following decentralized protection scheme which does not require the existence of a CA. Specifically, upon the completion of task II and task III, node i shall cut all future communication to the located malicious nodes, i.e., $\mathcal{E}_{cut}^i = \{ij \in E : \hat{\mathcal{H}}^{ij} = \mathcal{H}_1^{ij}\}$ where $\hat{\mathcal{H}}^{ij}$ is the outcome from the localization task. If completed successfully by all nodes, we can effectively isolate the attackers and prevent any future harm to the network. An illustration of the detection

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSIPN.2016.2614898, IEEE Transactions on Signal and Information Processing over Networks

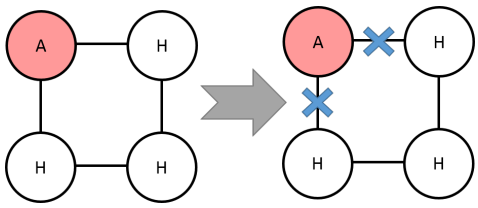

Fig. 2: Each normal node H will perform detection and localization independently, therefore isolating attacker A from the network.

and localization steps is depicted in Figure 2. We mention that this protection scheme is successful if *all* edges to an attacker are disconnected and the attacker cannot do any harm, i.e. $\|\boldsymbol{B}\|_2 = 0$. False detections, as long as they do not disconnect the graph, only slow down convergence.

Our detection schemes rely on finding the statistical anomalies through statistics computed at the normal nodes. In particular, we make the following assumptions:

$$\text{(A1)}: \ \bar{\gamma} := \mathbb{E}[\gamma_i^k], \ \forall\ i \in V_r. \quad \text{(A2)}: \ \bar{\gamma} \neq \bar{\alpha} := \mathbb{E}[\alpha^k],$$

where we emphasize that the expectation above is taken w.r.t. k. The first assumption (A1) states that the initial values for the normal nodes have the same mean $\bar{\gamma}$ for all normal nodes; and the second assumption (A2) states that $\bar{\gamma}$ is different from $\bar{\alpha}$, i.e., the initial value for the attacker.

A. Detection through the temporal difference

We introduce a strategy which detects the anomalies caused by the malicious agent by evaluating the (average) *temporal difference* of the values held by normal nodes. To explain the intuition, observe that the expected initial value of a malicious agent $s \in V_s$ is different from that a normal agent $j \in V_r$, i.e.,

$$\mathbb{E}[x_s^k(0)] = \bar{\alpha} \neq \bar{\gamma} = \mathbb{E}[x_j^k(0)]. \tag{14}$$

While when $t \to \infty$, the network will be misled by the malicious nodes, i.e.,

$$\mathbb{E}[x_s^k(\infty)] = \bar{\alpha} = \mathbb{E}[x_j^k(\infty)]. \tag{15}$$

This implies that the quantity $|x_i^k(\infty) - x_i^k(0)|$ will be close to zero if $i \in V_s$ or be large otherwise, indicating an anomaly.

The *temporal difference* method is developed from the observation above. Consider a normal node $i \in V_r$, let $x_j^k(T_{ij}), x_j^k(0)$ be respectively the last and the first observed state value for a node j in the neighborhood of node i. The following metric can be evaluated:

$$\xi_{ij} := \frac{1}{K}\sum_{k=0}^{K}(x_j^k(T_{ij}) - x_j^k(0)), \tag{16}$$

for all $j \in \mathcal{N}_i$. Notice that if T_{ij} is sufficiently large and node j is not malicious, then ξ_{ij} tends to be large.

We propose the following detection criterion for the neighborhood detection task (which implies attacker detection):

$$\sum_{m\in\mathcal{N}_i} |\xi_{im} - \bar{\xi}_i| \underset{\mathcal{H}_1^i}{\overset{\mathcal{H}_0^i}{\lessgtr}} \delta_I, \tag{17}$$

where $\bar{\xi}_i := (1/|\mathcal{N}_i|)\sum_{m\in\mathcal{N}_i}\xi_{im}$ and $\delta_I > 0$ is some predesigned threshold. The detection criterion in (17) finds if there is an outlier in $\mathcal{N}_i$ for the set of statistics $\{\xi_{im}\}_{m\in\mathcal{N}_i}$. This, however, implies that a node that has no attacker in its neighborhood cannot detect that an attack is present in the network. This can also be seen mathematically as $\mathbb{E}[\xi_{im} - \bar{\xi}_i] = 0$ for both $\mathcal{H}_0^i \cap \mathcal{H}_0$ and $\mathcal{H}_0^i \cap \mathcal{H}_1$ (17), where the $\cap$ operator returns true if both events are true. Note that we require that there is at least one normal neighbor to detect an attack.

For the localization task, we propose the following criterion:

$$|\xi_{ij}| \underset{\mathcal{H}_0^{ij}}{\overset{\mathcal{H}_1^{ij}}{\lessgtr}} \epsilon_I \tag{18}$$

for all $j \in \mathcal{N}_i$. The intuition behind this criterion was given at the beginning of this subsection. We remark that the localization task is performed only if the neighborhood detection task returns $\mathcal{H}_1^i$.

B. Detection through the spatial difference

This subsection describes a *spatial difference* strategy for data injection attack detection. Herein, our main idea is to exploit the fact that a malicious node, if it exists, always tries to influence and steers the nodes away from their true average; if there is no malicious node in the network, the average state of all nodes are identical. Mathematically, if $0 < t < \infty$,

$$\mathbb{E}[x_m^k(t) - x_j^k(t)|\mathcal{H}_0] = 0, \ \mathbb{E}[x_m^k(t) - x_j^k(t)|\mathcal{H}_1] \neq 0, \tag{19}$$

i.e., anomalies can be found in the *spatial difference* of states.

Define $\mathcal{T}_k \subseteq \mathbb{N}$ as the set of sampling times observed by a normal node i at the kth instance of the consensus algorithm. We consider the following metric for all $m \in \mathcal{N}_i \cup \{i\}$:

$$X_{im}^k := \sum_{t\in\mathcal{T}_k}\Big(x_m^k(t) - \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i} x_j^k(t)\Big). \tag{20}$$

Notice that $|X_{im}^k|$ is the difference between the value held by a neighboring node m and the sum of all the nodes in the same neighborhood (excluding node i itself), and then sum up this difference from all the observed consensus iterations. Compared to the temporal method, the spatial difference method registers an anomaly even if attacks are not staged directly in the neighborhood of node i. This is a double-edged-sword because while it indicates that one can attain situation awareness throughout the network, not just in the immediate proximity of attackers, it complicates the localization task.

Based on X_{im}^k, we have the following criterion for *both* attacker detection and neighborhood detection tasks:

$$S_1^i := \frac{1}{|\mathcal{N}_i|}\sum_{m\in\mathcal{N}_i}\Big(\frac{1}{K}\sum_{k=1}^{K} X_{im}^k\Big)^2 \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\lessgtr}} \delta_{II}, \tag{21}$$

where we shall use a different threshold $\delta'_{II} > \delta_{II}$ for the neighborhood detection task. Furthermore, we require that node i which performs (21) to have at least 2 neighbors, from which at least one is normal.

For the localization task, we define the following metric:

$$\tilde{X}_{ij}^k := \sum_{t\in\mathcal{T}_k}\big(x_j^k(t) - x_i^k(t)\big) - X_{ii}^k, \tag{22}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSIPN.2016.2614898, IEEE Transactions on Signal and Information Processing over Networks

which has a similar interpretation as X_{im}^k. This metric compares a neighboring node j to the node i itself and the neighborhood average with respect to the node itself. The localization task is performed by the following test:

$$S_2^{ij} := \Big(\frac{1}{K}\sum_{k=1}^{K}\tilde{X}_{ij}^k\Big)^2 \underset{\mathcal{H}_1^{ij}}{\overset{\mathcal{H}_0^{ij}}{\lessgtr}} \epsilon_{II} \tag{23}$$

The next section analyzes the performances of both temporal difference and spatial difference methods.

IV. Performance Analysis

We first define the following performance metric — for the *attacker detection task* (performed by the ith agent):

$$P_{\text{ad}}^i := P(\hat{\mathcal{H}} = \mathcal{H}_1|\mathcal{H}_1),\ P_{\text{af}}^i := P(\hat{\mathcal{H}} = \mathcal{H}_1|\mathcal{H}_0),$$

for the *neighborhood detection task*:

$$P_{\text{nd}}^i := P(\hat{\mathcal{H}}^i = \mathcal{H}_1^i|\mathcal{H}_1^i),\ P_{\text{nf}}^i := P(\hat{\mathcal{H}}^i = \mathcal{H}_1^i|\mathcal{H}_0^i),$$

for the *localization task*:

$$P_{\text{ld}}^{ij} := P(\hat{\mathcal{H}}^{ij} = \mathcal{H}_1^{ij}|\mathcal{H}_1^{ij}),\ P_{\text{lf}}^{ij} := P(\hat{\mathcal{H}}^{ij} = \mathcal{H}_1^{ij}|\mathcal{H}_0^{ij}),$$

Our analysis holds under the following assumptions on the statistics of the attacker and normal nodes:

- The initial state for normal nodes, $x_i^k(0) = \gamma_i^k$, is identically independently distributed (i.i.d.) with mean $\bar{\gamma}$ with sub-Gaussian parameter σ_γ^2.
- The initial state for malicious nodes, α^k, is i.i.d. with mean $\bar{\alpha}$ with sub-Gaussian parameter σ_α^2.
- The artificial noise for malicious nodes, $m_i^k(t)$, is i.i.d. with zero mean and sub-Gaussian parameter σ_m^2.

Notice that a random variable (r.v.) z with mean $\bar{z}$ is said to have sub-Gaussian parameter σ_z^2 if

$$\mathbb{E}\big[e^{\lambda(z-\bar{z})}\big] \le e^{\sigma_z^2\lambda^2/2},\ \forall\ \lambda \in \mathbb{R}.$$

If z is also Gaussian, then σ_z^2 is the variance of z.

Remark 3 *Even when these assumptions are violated, the metrics can be applied and loose bounds (e.g., using generalized Markov inequality) are obtainable as long as the distribution of the attackers states has finite moments. In the case of infinite moments the metrics however will fail. In addition, our simulation results test the same metrics in scenarios that violate the assumptions above to show their effectiveness.*

A. Analysis for the temporal difference strategy

Observe that the metric ξ_{ij} is evaluated as a finite sum of independent sub-Gaussian r.v.s. Define the following constants:

$$\mu_i = \frac{|\mathcal{N}_{i,r}|}{|\mathcal{N}_i|}(\bar{\alpha}-\bar{\gamma}),\ \mathcal{N}_{i,r} = V_r \cap \mathcal{N}_i,\ \mathcal{N}_{i,s} = V_s \cap \mathcal{N}_i, \tag{24}$$

$$\sigma_i^2 = \left(\frac{|\mathcal{N}_i|^2 - 2|\mathcal{N}_i| + |\mathcal{N}_{i,s}|}{|\mathcal{N}_i|^2}\right)\sigma_m^2 + \frac{|\mathcal{N}_{i,r}|^2}{|\mathcal{N}_i|^2}\sigma_\alpha^2 + \frac{|\mathcal{N}_{i,r}|}{|\mathcal{N}_i|^2}\sigma_\gamma^2,$$

We have the following performance guarantees:

Theorem 1 *Let $T_{ij} \to \infty$. We have*

$$P_{nf}^i \le 2|\mathcal{N}_i| \cdot \exp\Big(-K\delta_I^2/(2\sigma_\gamma^2|\mathcal{N}_i|(|\mathcal{N}_i|-1))\Big), \tag{25}$$

$$P_{nd}^i \ge 1 - \exp\Big(-K(\max\{0, -\delta_I + |\mu_i|\})^2/(2\sigma_i^2)\Big), \tag{26}$$

When the initial states and artificial noise are Gaussian, we have

$$P_{nf}^i \le 2|\mathcal{N}_i| \cdot Q\left(\sqrt{K}\delta_I/(\sigma_\gamma\sqrt{(|\mathcal{N}_i|-1)|\mathcal{N}_i|})\right), \tag{27}$$

$$P_{nd}^i \ge Q\left(\sqrt{K}(\delta_I + \mu_i)/\sigma_i\right) + Q\left(\sqrt{K}(\delta_I - \mu_i)/\sigma_i\right). \tag{28}$$

The result in Theorem 1 is proven in Appendix A for the sub-Gaussian case. The analysis above shows the impact of the variance on the detection (17) performance. We see that the false alarm rate (25) depends solely on σ_γ^2, while the miss detection rate (26) depends on the other parameters as well.

From Theorem 1, we observe that δ_I should be chosen to be smaller than $|\mu_i|$ to yield a non-trivial bound. In this case, let $\underline{P}_{\text{nd}}^i$ be the minimum required detection rate, the false alarm rate can be bounded as:

$$P_{\text{nf}}^i \le 2|\mathcal{N}_i| \cdot \exp\Big(-\frac{K}{|\mathcal{N}_i|}\frac{\Big(|\mu_i| - \sqrt{\frac{2\sigma_i^2}{K}\log\frac{1}{1-\underline{P}_{\text{nd}}^i}}\Big)^2}{2\sigma_\gamma^2(|\mathcal{N}_i|-1)}\Big).$$

This implies that for a given requirement on the detection rate probability, the false alarm probability will only be influenced by the choice of K.

Following the same line of reasoning we can prove the following performance bounds for the localization task:

Lemma 1 *Let $T_{ij} \to \infty$. We have*

$$P_{lf}^{ij} \le \exp\Big(-K(\max\{0, -\epsilon_I + |\bar{\alpha}-\bar{\gamma}|\})^2/(2(\sigma_\alpha^2+\sigma_\gamma^2))\Big), \tag{29}$$

$$P_{ld}^{ij} \ge 1 - 2\cdot\exp\Big(-K\epsilon_I^2/(2\sigma_m^2)\Big). \tag{30}$$

For the case of Gaussian random initial states and attackers noise we have:

$$P_{lf}^{ij} = Q\left(\frac{-\epsilon_I + |\bar{\alpha}-\bar{\gamma}|}{\sqrt{(\sigma_\alpha^2+\sigma_\gamma^2)/K}}\right) - Q\left(\frac{\epsilon_I + |\bar{\alpha}-\bar{\gamma}|}{\sqrt{(\sigma_\alpha^2+\sigma_\gamma^2)/K}}\right) \tag{31}$$

$$P_{ld}^{ij} = 1 - 2Q\left(\sqrt{K}\epsilon_I/\sigma_m\right). \tag{32}$$

Lemma 1 is proven in Appendix B for the sub-Gaussian case. Notice that the formulas in (31) and (32) are *exact*. Once again, if K is sufficiently large, (31) and (32) are good approximations for the sub-Gaussian case as well as we verified by simulation.

Remark 4 *The requirement that $T_{ij} \to \infty$ is imposed for the sake of obtaining constants that can be evaluated in closed form. Without such assumption, we can still obtain bounds similar to Theorem 1 and Lemma 1 and the exponential scaling with K will remain valid.*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSIPN.2016.2614898, IEEE Transactions on Signal and Information Processing over Networks

B. Analysis for the spatial difference strategy

We perform the following analysis under the assumption that the initial states $\boldsymbol{\gamma}^k, \alpha^k$ and attackers' noise $\boldsymbol{m}^k(t)$ are Gaussian distributed. Our first result is the following characterization of the random variable X_{im}^k. Let $\boldsymbol{e}_i$ be the unit vector with 1 being in the ith entry. We define

$$\eta_{im} = (\boldsymbol{e}_m - \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i}\boldsymbol{e}_j)^\top \begin{bmatrix} \boldsymbol{0} \\ (\bar{\gamma}-\bar{\alpha})\sum_{t\in\mathcal{T}_k}\boldsymbol{D}^t\boldsymbol{1}\end{bmatrix}. \quad (33)$$

Also, τ_{im} and β_{im} are constants that are bounded as

$$\tau_{im}^2 = \mathcal{O}\left(\sigma_\gamma^2\frac{\lambda_2(\overline{\boldsymbol{W}})}{(1-\lambda_2(\overline{\boldsymbol{W}}))^2}\right), \quad (34)$$

$$\beta_{im}^2 = \mathcal{O}\left((\sigma_\gamma^2+\sigma_\alpha^2)\frac{3\tilde{\lambda}-1}{(1-\tilde{\lambda})^2} + \sigma_m^2\frac{3\tilde{\lambda}-\tilde{\lambda}^2}{(1-\tilde{\lambda})^3}\right), \quad (35)$$

where $\tilde{\lambda} = \max\{\hat{\lambda}^2, \lambda_1(\boldsymbol{D})\}$. It can be proven that:

Theorem 2 *Assume that $\boldsymbol{\gamma}^k, \alpha^k$ and $\boldsymbol{m}^k(t)$ are Gaussian, the random variable X_{im}^k is also Gaussian with statistics*

$$\mathcal{H}_0 : X_{im}^k \sim \mathcal{N}(0, \tau_{im}^2) \quad \text{and} \quad \mathcal{H}_1 : X_{im}^k \sim \mathcal{N}\left(\eta_{im}, \beta_{im}^2\right).$$

The results hold for all $m \in \mathcal{N}_i \cup \{i\}$ and $i \in V_r$.

We remark that the hypothesis $\mathcal{H}_1$ refers to the scenario when an attacker exists somewhere in the network (not necessarily in $\mathcal{N}_i$). In other words, the statistics of X_{im}^k changes whenever at least one attacker exists therefore, this metric is a suitable candidate for performing the attacker detection task.

In fact, the difficulty in establishing Theorem 2 lies on the fact that X_{im}^k is an infinite sum of *correlated* random variables. It is not obvious whether its variance is bounded. In our proof, we exploit the sub-stochasticity of $\boldsymbol{D}$ and that the infinite sum may be treated as a converging geometric series. The proof to the proposition can be found in Appendix C:

Our main result is summarized as follows.

Theorem 3 *Assume that $\boldsymbol{\gamma}^k, \alpha^k, \boldsymbol{m}^k(t)$ are Gaussian. The attacker detection performance of the spatial difference strategy is given as:*

$$P_{af}^i \le \exp\big(-K\max\{0, |\mathcal{N}_i|\delta_{II} - c_0\}/c_1\big). \quad (36)$$

for some c_0, c_1 that scale with τ_{im}. Also,

$$1 - P_{ad}^i \le \exp\left(-K\max\{0, -|\mathcal{N}_i|\delta_{II} + c_2\}\right),$$

for some $c_2 > 0$ that scales with β_{im}, η_{im}.

The proof is relegated to Appendix D.

Next we characterize the localization performance. In the following, we shall assume that at least one attacker is present in the network, i.e., $\mathcal{H}_1$ holds. Our first step is to study the statistics of $\tilde{X}_{ij}^k$. Observe that

$$\tilde{\eta}_{ij} = \mathbb{E}[\tilde{X}_{ij}^k] = (\boldsymbol{e}_j - \boldsymbol{e}_i)^\top \begin{bmatrix} \boldsymbol{0} \\ (\bar{\gamma}-\bar{\alpha})\sum_{t\in\mathcal{T}_k}\boldsymbol{D}^t\boldsymbol{1}\end{bmatrix} - \eta_{ii}.$$

Moreover, the variance can be bounded as

$$\tilde{\beta}_{ij}^2 := \mathrm{var}(\tilde{X}_{ij}^k) \le 4\max\{\mathrm{var}\big(\textstyle\sum_{t\in\mathcal{T}_k}(x_j^k(t) - x_i^k(t))\big), \beta_{ii}^2\}$$

In the same spirit as the proof of Proposition 2, it can be verified that $\mathrm{var}\big(\sum_{t\in\mathcal{T}_k}(x_j^k(t) - x_i^k(t))\big) < \infty$ and thus $\tilde{\beta}_{ij}^2$ is bounded from above. We remark that the values of $\tilde{\eta}_{ij}, \tilde{\beta}_{ij}^2$ are dependent on the cases $(\mathcal{H}_0^{ij}, \mathcal{H}_1^{ij})$ we are in. For instance, it can be seen that $|\tilde{\eta}_{ij}|$ is larger in $\mathcal{H}_1^{ij}$ than in $\mathcal{H}_0^{ij}$.

Knowing the statistics above, the localization performance can be evaluated straightforwardly as:

Lemma 2 *Assume that $\boldsymbol{\gamma}^k, \alpha^k, \boldsymbol{m}^k(t)$ are Gaussian, the localization performance of the spatial difference strategy is given as:*

$$P_{lf}^{ij} \le Q\left(\sqrt{K}(\sqrt{\epsilon_{II}} - \tilde{\eta}_{ij})/\tilde{\beta}_{ij}\right) + Q\left(\sqrt{K}(\sqrt{\epsilon_{II}} + \tilde{\eta}_{ij})/\tilde{\beta}_{ij}\right)$$

$$1 - P_{ld}^{ij} \le Q\left(\sqrt{K}(\tilde{\eta}_{ij} - \sqrt{\epsilon_{II}})/\tilde{\beta}_{ij}\right) - Q\left(\sqrt{K}(\sqrt{\epsilon_{II}} + \tilde{\eta}_{ij})/\tilde{\beta}_{ij}\right).$$

The proof can be found in Appendix E.

Compared to the analysis of the temporal strategy, we observe a similar improvement in the performance that decays exponentially in K. Moreover, the bounds obtained for the attack detection task using the spatial difference method depend explicitly on the network topology, while the respective bound for neighborhood detection task with the temporal difference method only depends on the neighborhood size.

Remark 5 *We remark that the analysis above can be extended to sub-Gaussian initial states, e.g., by applying the general results from [20], [21]. We omit such extensions in the interest of space limitation.*

C. Optimal Attacker's Strategy

We consider a scenario when the attacker optimizes his/her strategy to maximize the damages caused to the consensus network. Specifically, we focus on the defense strategy employing the temporal difference detection (cf. Section III-A) and assume that the attacker is aware of the strategy employed by the network, including the parameter δ_I.

The attacker's goal is to introduce the maximum perturbation $|\bar{\alpha} - \bar{\gamma}|$ to the network's final state, while avoiding being detected. For simplicity, we assume that $\sigma_\alpha^2 = 0$ and the attacker optimizes its attack statistics by:

$$\max_{\bar{\alpha}} \quad |\bar{\alpha} - \bar{\gamma}| \quad \text{s.t.} \quad P_{nd}^i(\bar{\alpha}) \le \Pi, \ \forall\, i \in V_r, \quad (37)$$

where $\Pi \in (0,1)$ is the detection probability threshold for the attacker.

Due to the intractability of the constraint on $P_{nd}^i(\bar{\alpha})$ in Problem (37), we bound the detection probability and derive a conservative approximation to (37). The following can be derived as an extension to Theorem 1:

Lemma 3 *Let $T_{ij} \to \infty$. We have*

$$P_{nd}^i(\boldsymbol{\theta}_{att}) \le 2|\mathcal{N}_i| \cdot \exp\Big(-K\frac{(\max\{0, |\mathcal{N}_i|^{-1}\delta_I - |\mu_i|\})^2}{2\sigma_i^2}\Big), \quad (38)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSIPN.2016.2614898, IEEE Transactions on Signal and Information Processing over Networks

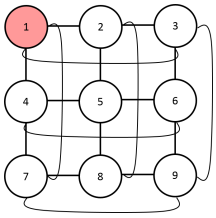

Fig. 3: The Manhattan network topology considered. First we select only node 1 as an attacker, while all other nodes are normal. Then, in the second set of experiments, we consider and increasing number of nodes to be an attacker i.e node $[1], [1, 2], [1, 2, 3] \ldots$.

When the initial states and artificial noise are Gaussian,

$$P_{nd}^{i}(\boldsymbol{\theta}_{att}) \leq 2|\mathcal{N}_i| \cdot Q\Big(\sqrt{K}\big(|\mathcal{N}_i|^{-1}\delta_I - |\mu_i|\big)/\sigma_i\Big), \quad (39)$$

see the definitions of the constants in (24).

The proof can be found in Appendix F. Notice that the bounds above are non-trivial only when $|\mathcal{N}_i|^{-1}\delta_I \geq |\mu_i|$, since otherwise the bounds become equal to 1. Since $|\mu_i| \propto |\bar{\alpha} - \bar{\gamma}|$, this limits the maximum deviation that the attacker can introduce to the network. Based on the bound from Theorem 1, we observe that $-\delta_I + |\mu_i| \leq 0$, then $P_{nd}^{i} \geq 1$ and any attack will be detected. Applying the results above, an approximate optimal attack strategy can be found by replacing $P_{nd}^{i}(\boldsymbol{\theta}_{att})$ in (37) with the right hand side of (38) or (39).

To obtain an optimal solution to the approximated (37), we observe that the bounds in (38) (or (39)) are monotonically increasing with $|\mu_i|$. Now, to maximize $|\bar{\alpha} - \bar{\gamma}|$, the right hand side of (38) (or (39)) must equal to Π. Taking the sub-Gaussian case as an example, suppose the consensus network designs a threshold δ_I such that the false alarm probability is no bigger than $\bar{P}_{\text{nf}}^{i}$, it can be verified that the maximum perturbation subject to a detection probability Π is:

$$|\bar{\alpha}^{\star} - \bar{\gamma}| = \max\Big\{0, \sqrt{\frac{2\sigma_{\gamma}^2|\mathcal{N}_i|(|\mathcal{N}_i|-1)}{K|\mathcal{N}_{i,r}|^2}\log\big(\frac{2|\mathcal{N}_i|}{\bar{P}_{\text{nf}}^{i}}\big)} - \sqrt{\frac{2\sigma_i^2|\mathcal{N}_i|^2}{K|\mathcal{N}_{i,r}|^2}\log\big(\frac{2|\mathcal{N}_i|}{\Pi}\big)}\Big\}.$$

Note that as $K \to \infty$, the maximum perturbation goes to zero.

V. Numerical Results

In this section, we numerically evaluate the performance of the proposed methods. For the simulation results that follow, we consider a Manhattan topology with $n = 9$ nodes, as shown in Figure 3. The randomized gossip-based consensus protocol (cf. Algorithm 1) is run with $P_{ij} = 1/|\mathcal{N}_i|$ (cf. (2)), and is terminated with $T = 500$. We have $\alpha^k \sim \mathcal{N}(0, 1)$, $\gamma_i^k \sim \mathcal{U}[-0.5, 1.5]$, $m_i^k(t) \sim \mathcal{U}[-\hat{\lambda}^t, \hat{\lambda}^t]$. The Monte Carlo simulation is run with 10^3 trials.

Before we present the performance evaluations, let us describe a few observations that motivated us to develop our methods. In Figure 4, we show the evolution of the states of all nodes in an instance of the average consensus algorithm when an attacker is present. Recalling that $\xi_{ij} \approx x_j^k(T_{ij}) - x_j^k(0)$, i.e., the difference between the terminal and initial state values,

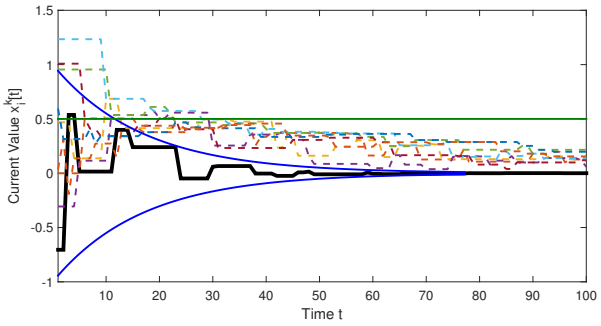

Fig. 4: State evolution in a single random consensus run. The dashed lines are the state trajectories for the normal nodes. The malicious node (black) is forcing all normal nodes (dashed) to its target value $\alpha = 0$, while the true $x_{av} = 0.5$ (green). Furthermore the noise of the malicious agent is given by the true λ_2 of the network without attackers (blue).

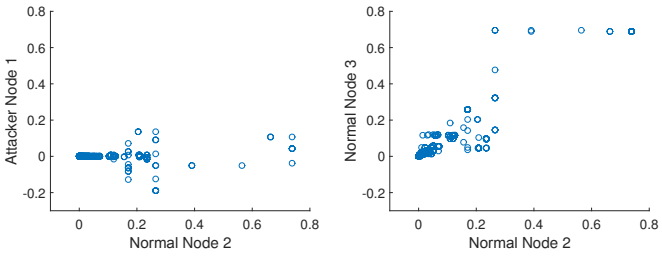

Fig. 5: (Left) Scatter plot of normal node and attacker $(x_2[k], x_1[k])$; (Right) Scatter plot of 2 normal Node $(x_2[k], x_3[k])$.

we see that ξ_{ij} tends to be larger if node j is normal, i.e., $j \in V_r$. On the other hand, Figure 5 presents a scatter plot for the value of $(x_i^k(t), x_j^k(t))$ for two pairs of adjacent nodes, one with a malicious node and one without. We observe that in the former case, the scatter plot is tilted towards horizontal, indicating a larger spatial difference, i.e., V_{ij} (or $\tilde{V}_{ij}$).

A. Detection and Localization with one attacker

We first simulate the performance of the proposed schemes when the network is under attack from $|V_s| = 1$ malicious node. As the network topology is symmetrical, without loss of generality we set node 1 to be the attacker. Notice that there are 4 nodes located directly next to the attacker.

1) Temporal Difference Method: We present the receiver operating characteristic (ROC) curves for the temporal difference method in Figure 6. First, we consider the performance of the neighborhood detection task in Figure 6 (Left). As we only focus on the case when the evaluating node i is located next to the attacker, the ROC curves also correspond to the attacker detection task. The false alarm and detection probabilities are evaluated by taking an average of the probabilities of all the four neighbors of the attacker. From the figure, we notice that the detection performance improve as K increases, as predicted in Theorem 1. Accruing statistics from $K \approx 100$ instances seems to provide a reliable detection.

For the localization task in Figure 6 (Right), we assume that the neighborhood detection test was completed without errors (by an 'Oracle'). Similar to the attacker detection, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSIPN.2016.2614898, IEEE Transactions on Signal and Information Processing over Networks

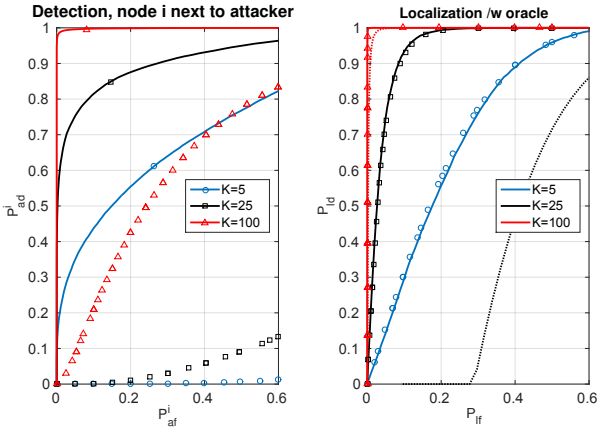

Fig. 6: Temporal method detector performance: (Left) ROCs for attacker detection. For the considered Ks, the theoretical bounds (25) & (26) are trivial and therefore omitted. (Right) ROCs for localization of attacker. Dotted lines show the theoretical bounds in (31) & (32). Markers show the bounds obtained by applying Gaussian approximation.

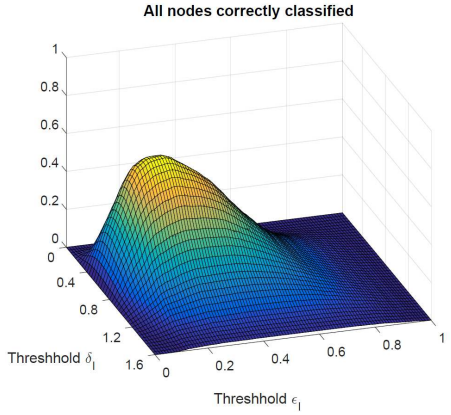

Fig. 7: Temporal method: Probability of correct localization of all nodes for $K = 25$

performance of the localization task improves with K. Moreover, with the same K, the performance of the detection task is worse than that of the localization task. This corroborates our observations in Theorem 1 and Lemma 1.

We also compare the theoretical bounds with the actual performance for the two tasks. Here, the performance bounds predicted for the Gaussian case are plotted. We observe that the bounds (27) & (28) are generally loose in the case of attacker detection, yet (31) & (32) nearly match the actual performance.

We then investigate the optimal thresholds ϵ_I, δ_I for the temporal difference method. Fixing at $K = 25$, Figure 7 shows the probability when *all* neighbors are classified correctly. Since all neighboring nodes have to be classified correctly, the nodes next to the attacker have to both detect and localize the attackers, as well as classifying the normal nodes. Nodes that are not next to an attacker, i.e. with only normal neighbors, must *not* detect a neighborhood attack. We find that for the thresholds (δ_I, ϵ_I) chosen as (0.6,0.25), we classify *all* neighbors correctly with a probability of 63%.

2) Spatial Difference Method: For the spatial difference method, the ROC curves of the attacker detection are shown in Figure 8. Contrary to the temporal difference case, the spatial difference method can also detect an attack if a node is more than one hop away from the attacker. Therefore, in addition to performing attacker detection on nodes that are directly next to the attacker, we also compare the detection performance on nodes that, for our example, are two hops away from the attacker seen in Figure 8 (Right). From the figure, the nodes that are directly next to an attacker are clearly more sensitive then the nodes far from an attacker.

In Figure 9, we show the ROC curves for the localization and neighborhood detection tasks. Specifically, in the neighborhood detection task we evaluate the false alarm/detection probabilities conditioned on $\mathcal{H}_1$, i.e., when the attacker is actually present in the network. Observe that we can get the neighborhood detection to work, however it is the worst performing test for the spatial difference method. For the localization task in Figure 9 (Right), similar to the temporal case, we assume that the neighborhood detection test was completed without errors. Also in this case, the tests improve with K in a way that is more pronounced than with the temporal difference method, and under the same K, the performance of the neighborhood detection task is worse than that of the localization task. Nevertheless, the spatial method has a drastic advantage over the temporal method in spotting attacks, as it leverages information of the entire dynamic, while the temporal method only uses the initial and terminal states.

We now investigate the optimal thresholds $\epsilon_{II}, \delta_{II}$ for the spatial difference method by studying the case with $K = 25$. In Figure 10 we plot the probability when *all* nodes are classified correctly, using similar settings as Figure 8. We find that for the thresholds $(\delta_{II}, \epsilon_{II})$ chosen as (50,1100) we classify all nodes correctly with a probability of 87%. Comparing the performance of the temporal and spatial difference methods, we see that the spatial difference method is outperforming the former. However, we notice that the computational complexity requirement of applying the spatial difference method is higher.

3) Non Sub-Gaussian Distribution: Next, we evaluate the performance of the data injection attack methods when the states' distributions are not sub-Gaussian. In particular, we repeat the simulations above with the attackers and normal nodes' states generated with a Laplacian distribution, with unit variance and mean $\bar{\gamma} = 0.5$ for normal nodes, and mean $\bar{\alpha} = 0$ for the attacker. The simulation results are shown in Figure 11. As seen, the detection/localization performances are almost identical to the cases considered with Gaussian initialization. This shows that the proposed methods are robust to the distribution of the nodes' states.

4) Correlated Attackers: Finally, we consider a scenario when the attacker's target values are correlated across instances. We assume that the attacker's states follow an autoregressive model, i.e., $\alpha^{k+1} = 0.9\gamma^{k+1} + 0.1\alpha^k$ with $\alpha^0 = \gamma^0$; while the other settings remain the same. Under this attack, the detector performance is shown in Figure 12. As seen in the figure, the two proposed method achieve similar performance as in the case with i.i.d. attacker's statistics.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSIPN.2016.2614898, IEEE Transactions on Signal and Information Processing over Networks

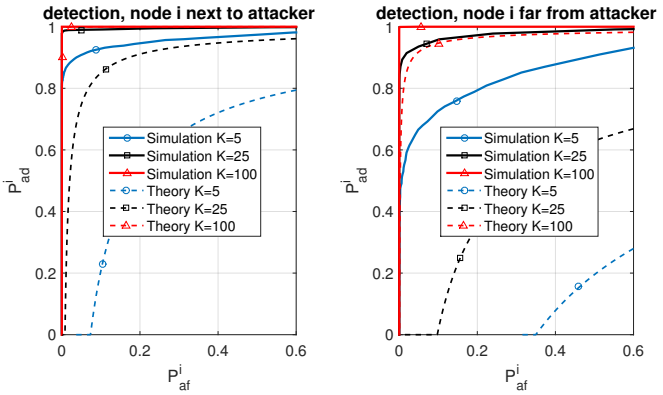

Fig. 8: ROCs spatial difference detection performance (Left) of nodes next to an attacker.(Right) of nodes not next to an attacker. Dashed lines show the theoretical bounds for Gaussian approximation

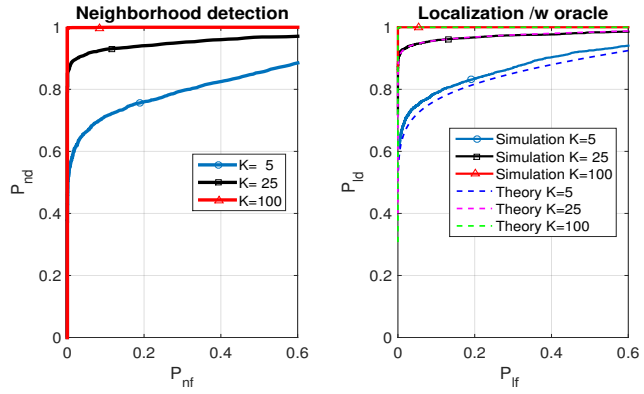

Fig. 9: ROCs spatial difference performance: (Left) for neighborhood detection. (Right) Localization of the attacker with oracle for neighborhood attack detection. Dashed lines show the theoretical bounds for Gaussian approximation

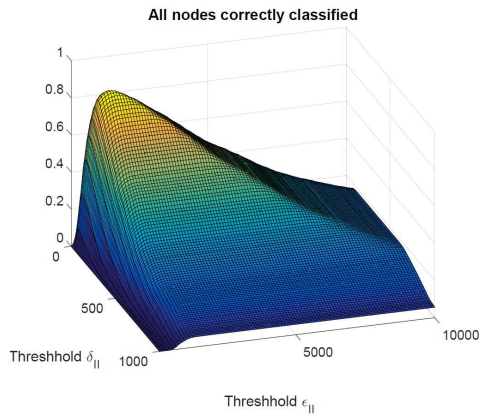

Fig. 10: Spatial method: Probability of correct localization of all nodes for $K = 25$

B. Detection & localization of multiple attackers

We now consider the case when the network is under the coordinated attack from multiple nodes. We consider the same topology and parameters as before and fix $K = 25$ for all experiments. The attackers share the same α^k, but each of them adds a random and independent series of noise samples. In the experiments, we assign the first d nodes as the attackers, i.e., we set nodes $\{1, ..., d\}$ as the attackers when considering

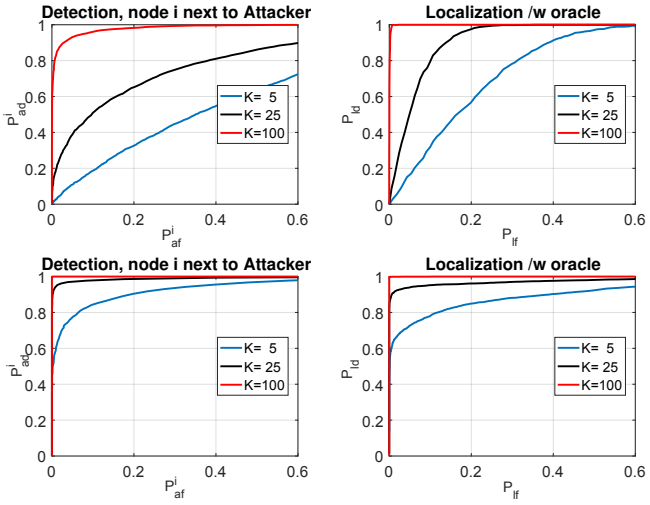

Fig. 11: ROCs with the Laplacian distribution via the (Top) temporal method and (Bottom) spatial method.

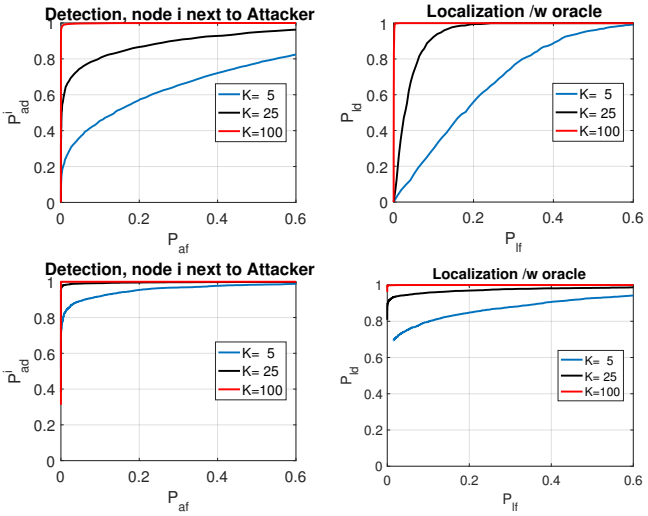

Fig. 12: ROCs with an autoregressive attacker's target value via the (Top) temporal method and (Bottom) spatial method.

a scenario with d attackers (cf. Figure 3).

Figure 13 plots the ROC curves for the attacker detection task with up to 7 attackers. We notice that the performance for both methods depends on the number of attacking neighbors. At this point we recall that in the chosen topology each node has 4 neighbors. For the temporal method, we observe that the detection rate is best with 2 attackers and 2 normal neighboring nodes. For the spatial method, we notice that the attack detection performance degrades with the amount of attackers increasing in the network, but the attacker localization performance is the best with 2 neighboring attackers and 2 normal nodes. This result makes sense as in this case, (20) will be maximized, thus giving rise to a higher detection rate. The performance also seems to be identical for nodes with either 1 or 3 attacking neighbors, given the same total number of attackers in the network. This is due to the fact that each node is comparing with all its neighbors (21). As there are more attackers, the value of $\frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i} x_j^k(t)$ becomes more biased by the attackers themselves, which therefore become

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSIPN.2016.2614898, IEEE Transactions on Signal and Information Processing over Networks

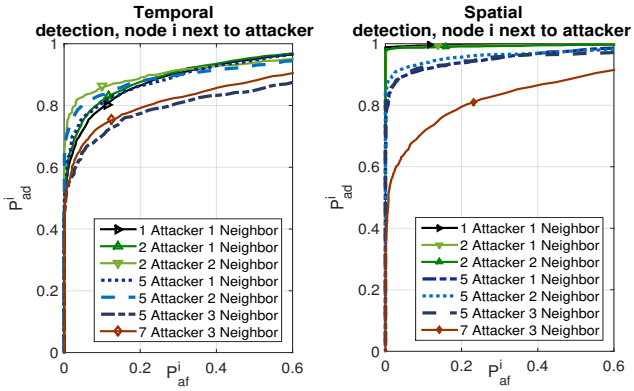

Fig. 13: Detection performance, depending on how many attackers are present, averaged over all nodes with the same amount of malicious neighbors: (Left) ROCs for the temporal method (Right) ROCs for the spatial method.

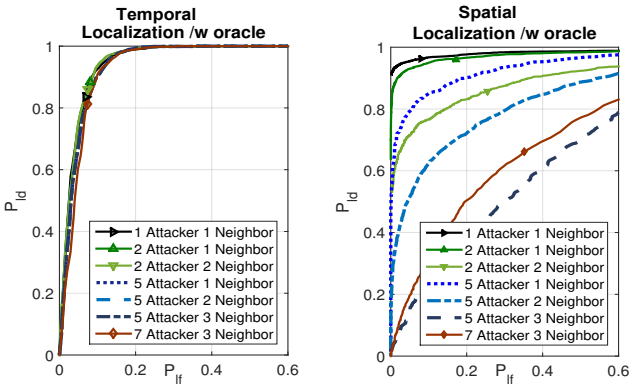

Fig. 14: Localization performance, depending on how many attackers are present, averaged over all nodes with the same amount of malicious neighbors: (Left) ROCs for the temporal method (Right) ROCs for the spatial method.

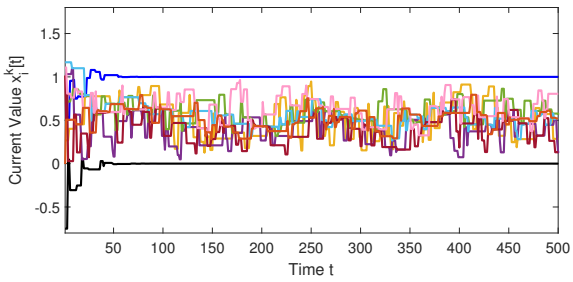

Fig. 15: Trajectories of nodes' states with two non-agreeing attackers in the network. The attackers' states are in black and blue

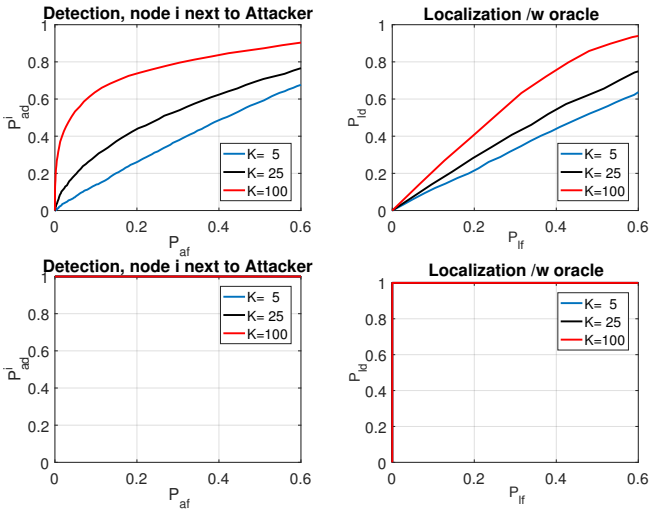

Fig. 16: ROCs with non agreeing attackers via the (Top) temporal method; (Bottom) spatial method.

harder to detect.

In Figure 14, we show the ROC curves for the localization task with up to 7 attackers. For the temporal method, the localization performance is independent of the number of attackers and neighboring attackers. Meanwhile, the spatial method's localization performance, degrades with both the number of attackers and the number of neighboring attackers. We speculate that this is due to the fact that in (23), when there are more neighboring attackers, the influence of an individual attacker becomes less pronounced. Furthermore with more attackers in the network, more normal nodes will be affected directly, thus the localization task becomes more difficult with the spatial difference method. Nevertheless, the spatial method performs the best from Figure 14. With an increasing amount of attackers, however, the temporal difference method provides better performance.

Lastly, we consider a scenario when the attacker nodes, i.e., node 1 & 2, do not share the same target value. In Figure 15, we show the trajectories of the nodes' states of the consensus algorithm when we set $[\boldsymbol{\alpha}^k]_1 = 1$ and $[\boldsymbol{\alpha}^k]_2 = 0$. We observe that the two attackers settles at their individual target value (in black and blue), while the states of normal nodes fluctuate between 0 and 1. The detection and localization performances of the proposed methods are shown in Figure 16. From the top figure, we observe the temporal metric is able to the attack with worsened performance than in the previous sections. On the other hand, from the bottom figure, the spatial metric achieves an almost perfect performance. The reason for that is that the network is never converging, thus the spatial metric diverges, i.e., $X_{im}^k \to \infty$, whenever the attackers are present (cf. compare Figure 15 with Figure 4).

C. Decentralized Disconnection

In Figure 17, we show the performance of decentralized disconnection method discussed in Section III using the spatial metric, for the same network topology as in the previous experiments with one attacker and setting $K = 25$.

We show the expected number of residual attackers in the network in Figure 17 (Left), from which we observe that after 4 iterations of the detection and disconnection algorithm we remove the attacker 100% of the times.

In Figure 17 (Middle) we show the probability that normal nodes get disconnected. Because we set a low probability of false alarm we rarely have disconnections and rarely they are in excess of the attacker. This explains the fact that the average algebraic connectivity, shown in Figure 17 (Right), hardly changes over the iterations.

D. Sequential change detection

We consider extending the attacker detection method to a sequential change detection. The idea is compute the log-likelihood of each sample (of temporal or spatial difference

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSIPN.2016.2614898, IEEE Transactions on Signal and Information Processing over Networks

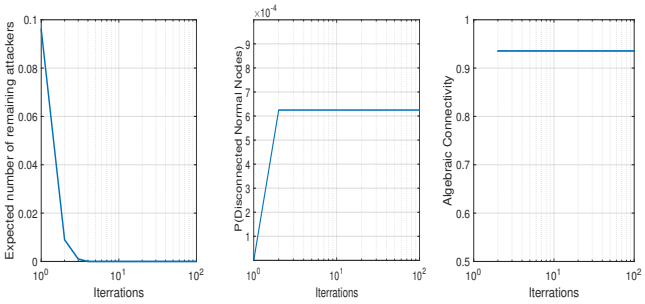

Fig. 17: (Left) Expected number of residual attackers, (Middle) probability of disconnected nodes and (Right) average algebraic connectivity versus iteration number.

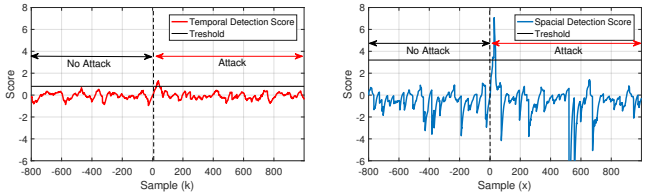

Fig. 18: Trajectories of the log likelihood using (Left) temporal difference metric and (Right) spatial difference metric.

metrics) assuming a change in mean and variance of a Gaussian distribution (approximating the true distribution) and decide if a significant change in distribution has occurred due to attacks.

Let $\bar{\mu}^k, (\bar{\sigma}^k)^2$ be estimated mean and variance respectively in the *absence* of an attack, adaptively updated at every k as follows

$$\bar{\mu}^{k+1} = 0.98\bar{\mu}^k + 0.02 z_i^k, \tag{40}$$

$$\left(\bar{\sigma}^{k+1}\right)^2 = 0.98\left(\bar{\sigma}^k\right)^2 + 0.02 \cdot (z_i^k - \bar{\mu}^k)^2. \tag{41}$$

where z_i^k is the temporal or spatial's metrics output. Moreover, at instance k, we compute the sample mean $\hat{\mu}^k$ and variance $(\hat{\sigma}^k)^2$ from a window of 25 samples of z_i^k into the future.

If the network is under attack, we expect to see a significant difference between the two pairs of sample mean/variance. Assuming a Gaussian prior, we can compute the log-likelihood:

$$L^k = \log(\hat{\sigma}^k/\bar{\sigma}^k) - ((z_i^k - \hat{\mu}^k)^2 - (z_i^k - \bar{\mu}^k)^2)/2 \tag{42}$$

We further smooth out the log-likelihood using an AR model with a forgetting factor of 0.1. Figure 18 shows an example trajectory of L^k with the temporal and spatial difference metric, where the attack began at sample 0 and the first 200 samples are used to initialize the estimators. Both detectors show a peak after the attack started, and fall under the threshold after the change, as the attack continue persist and the mean/variance estimated is now tracking the high value corresponding to the persisting attack. The classification of the state of the network can be done by comparing the sample mean/variance prior and after the change since the metric has an increased mean/variance when the network is under attack. Note that the spatial metric's peak is higher than the temporal one, indicating once again that the spatial method has better detection performance. In Figure 19 we show the probability that an attack/change is detected before sample k, with an

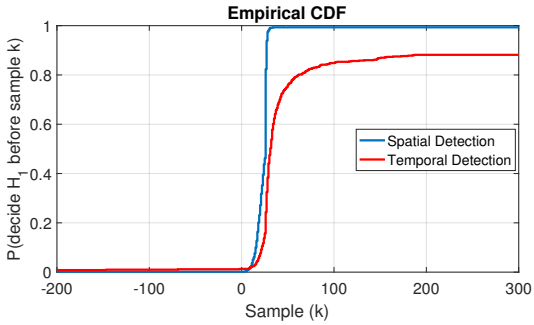

Fig. 19: Performance of sequential change attack detection.

attack happening at sample $k = 0$, averaged over 600 Monte-Carlo trials. We can see that the temporal method does not detect all the time the starting time of the attack but provides a reasonable low amount of false alarm. The spatial method on the other hand is very accurate.

VI. Conclusion

To conclude, we have proposed two independent novel strategies to detect and localize malicious nodes in a randomized consensus algorithm. Each strategy can be performed at each individual node in a completely decentralized manner and without any communication overhead. The performance bounds of the algorithm are analyzed, and the simulation results confirmed our findings.

VII. Acknowledgement

We thank the anonymous reviewers for their constructive comments on the paper.

Appendix A
Proof of Theorem 1

Observe the following chain for the false alarm rate:

$$P\left(\hat{\mathcal{H}} = \mathcal{H}_1^i \mid \mathcal{H}_0^i\right) = P\left(\sum_{m\in\mathcal{N}_i} |\xi_{im} - \bar{\xi}_i| \geq \delta \mid \mathcal{H}_0^i\right)$$
$$\leq |\mathcal{N}_i| P\left(|\xi_{im} - \bar{\xi}_i| \geq \frac{\delta}{|\mathcal{N}_i|} \mid \mathcal{H}_0^i\right), \text{ for some } m \in \mathcal{N}_i, \tag{43}$$

where we have applied the union bound in the last inequality. We have

$$\xi_{im} - \bar{\xi}_i = \frac{1}{K}\sum_{k=1}^{K}\left(\left(-1 + \frac{1}{|\mathcal{N}_i|}\right)\gamma_m^k + \sum_{j\in\mathcal{N}_i\setminus\{m\}} \frac{1}{|\mathcal{N}_i|}\gamma_j^k\right).$$

The quantity above is a zero mean r.v. with sub-Gaussian parameter $\sigma_\gamma^2(|\mathcal{N}_i| - 1)/(K|\mathcal{N}_i|)$. Applying the Hoeffding's inequality [22] to the last term of (43) yields the desired result.

For the miss detection rate, we have

$$P\left(\hat{\mathcal{H}} = \mathcal{H}_0^i \mid \mathcal{H}_1^i\right) = P\left(\sum_{m\in\mathcal{N}_i} |\xi_{im} - \bar{\xi}_i| \leq \delta \mid \mathcal{H}_1^i\right)$$
$$\leq P\left(|\xi_{im} - \bar{\xi}_i| \leq \delta \mid \mathcal{H}_1^i\right) \ \forall\ m \in \mathcal{N}_i. \tag{44}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSIPN.2016.2614898, IEEE Transactions on Signal and Information Processing over Networks

Observe that

$$\xi_{ij} = \begin{cases} (1/K)\sum_{k=1}^{K}(m_j^k(0)), & j \in V_s, \\ (1/K)\sum_{k=1}^{K}(\alpha^k - \gamma_j^k), & j \notin V_s. \end{cases}$$

and

$$\bar{\xi}_i = \frac{1}{K|\mathcal{N}_i|}\sum_{k=1}^{K}\left(\sum_{j\in V_s\cap\mathcal{N}_i} m_j^k(0) + \sum_{j\in V_r\cap\mathcal{N}_i}(\alpha^k - \gamma_j^k)\right)$$

We observe that $\xi_{im} - \bar{\xi}_i$ is a r.v. with mean μ_i and sub-Gaussian parameter σ_i^2/K for $m \in V_s$. Let us write $\xi_{im} - \bar{\xi}_i = \tilde{\xi}_{im} + \mu_i$. We can upper bound the last term in (44) as:

$$P\left(|\xi_{im} - \bar{\xi}_i| \le \delta \mid \mathcal{H}_1^i\right) \le P\left(\tilde{\xi}_{im} \ge -\delta + |\mu_i| \mid \mathcal{H}_1^i\right) \quad (45)$$

Consequently, the desired inequality can be obtained by applying Hoefding's inequality.

Appendix B
Proof of Lemma 1

Under $\mathcal{H}_1^{ij}$, we have $\xi_{ij} = (1/K)\sum_{k=1}^{K} m_j^k(0)$, where $m_j^k(0)$ are zero mean, independent r.v.s with sub-Gaussian parameter σ_M^2. Under $\mathcal{H}_0^{ij}$, we have

$$\xi_{ij} = \frac{1}{K}\sum_{k=1}^{K}(\alpha^k - \gamma_j^k), \quad (46)$$

note that the terms inside the summation have mean $\bar{\alpha} - \bar{\gamma}$ and are independent with sub-Gaussian parameter $\sigma_\alpha^2 + \sigma_\gamma^2$. Similar to Theorem 1. the desired inequalities can be obtained by applying Hoeffding's inequality.

Appendix C
Proof of Theorem 2

For the ease of presentation, we ignore the index i throughout this proof. Throughout this section, we use $\bullet$ to denotes the inner product between matrices, i.e., $\boldsymbol{A} \bullet \boldsymbol{B} := \mathrm{Tr}(\boldsymbol{A}^\top \boldsymbol{B})$.

A. First-order statistics

Under $\mathcal{H}_0$, it is obvious that:

$$\mathbb{E}[X_m^k] = \mathbb{E}\left[\sum_{t\in\mathcal{T}_k}\left(x_m^k(t) - \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i} x_j^k(t)\right)\right] = 0. \quad (47)$$

Under $\mathcal{H}_1$, we observe the following chain:

$$\begin{aligned}\eta_{im} &= \mathbb{E}[X_m^k] = \mathbb{E}\left[\sum_{t\in\mathcal{T}_k}\left(x_m^k(t) - \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i} x_j^k(t)\right)\right] \\ &= (\boldsymbol{e}_m - \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i}\boldsymbol{e}_j)^\top \sum_{t\in\mathcal{T}_k}\overline{\boldsymbol{W}}^t\begin{bmatrix}\bar{\alpha}\mathbf{1}\\ \bar{\gamma}\mathbf{1}\end{bmatrix} \\ &= (\boldsymbol{e}_m - \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i}\boldsymbol{e}_j)^\top \sum_{t\in\mathcal{T}_k}\overline{\boldsymbol{W}}^t\left(\bar{\alpha}\mathbf{1} + (\bar{\gamma}-\bar{\alpha})\begin{bmatrix}\mathbf{0}\\ \mathbf{1}\end{bmatrix}\right) \\ &= (\boldsymbol{e}_m - \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i}\boldsymbol{e}_j)^\top \begin{bmatrix}\mathbf{0}\\ (\bar{\gamma}-\bar{\alpha})\sum_{t\in\mathcal{T}_k}\boldsymbol{D}^t\mathbf{1}\end{bmatrix}.\end{aligned}$$

where the last equality is due to the stochasticity of $\overline{\boldsymbol{W}}$ and the fact that $(\boldsymbol{e}_m - \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i}\boldsymbol{e}_j)^\top\mathbf{1} = 0$.

B. Second-order statistics

Define the following quantities

$$\boldsymbol{A}_{t_1,t_2}^k := \left(\boldsymbol{x}^k(t_1) - \mathbb{E}[\boldsymbol{x}^k(t_1)]\right)\left(\boldsymbol{x}^k(t_2) - \mathbb{E}[\boldsymbol{x}^k(t_2)]\right)^\top. \quad (48)$$

$$\boldsymbol{F}_{n,m} := (\boldsymbol{e}_n - \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i}\boldsymbol{e}_j)(\boldsymbol{e}_m - \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i}\boldsymbol{e}_j)^\top. \quad (49)$$

Note that the variance of X_m^k can be written as $\mathbb{E}[(X_m^k - \mathbb{E}[X_m^k])^2] = \boldsymbol{F}_{m,m} \bullet (\sum_{t_1,t_2}\mathbb{E}[\boldsymbol{A}_{t_1,t_2}^k])$.

1) Under hypothesis $\mathcal{H}_0$ — In this case, we observe that $\mathbb{E}[\boldsymbol{x}^k(t)] = \bar{\gamma}\mathbf{1}$. Using $\boldsymbol{W}(t)\mathbf{1} = \mathbf{1}$ for all t, we have:

$$\boldsymbol{x}^k(t) - \mathbb{E}[\boldsymbol{x}^k(t)] = \boldsymbol{W}(t)\cdots\boldsymbol{W}(1)(\boldsymbol{\gamma}^k - \bar{\gamma}\mathbf{1}),$$

where $(\boldsymbol{\gamma}^k - \bar{\gamma}\mathbf{1}) \sim \mathcal{N}(0, \sigma_\gamma^2\boldsymbol{I})$ is independent of $\boldsymbol{W}(t), ..., \boldsymbol{W}(1)$. We can evaluate:

$$\begin{aligned}\mathbb{E}[\boldsymbol{A}_{t_1,t_2}^k] &= \sigma_\gamma^2\mathbb{E}[\boldsymbol{W}(t_1)\cdots\boldsymbol{W}(1)\boldsymbol{W}(1)^\top\cdots\boldsymbol{W}(t_2)^\top] \\ &= \sigma_\gamma^2\overline{\boldsymbol{W}}^{\max\{t_1,t_2\}},\end{aligned} \quad (50)$$

where the first equality is due to $\mathbb{E}\left[\boldsymbol{W}(t)\boldsymbol{W}^\top(t)\right] = \mathbb{E}\left[\boldsymbol{W}^2(t)\right] = \mathbb{E}\left[\boldsymbol{W}(t)\right] = \overline{\boldsymbol{W}}$, since $\boldsymbol{W}(t)$ is a projection matrix under the current hypothesis. The variance of X_m^k under $\mathcal{H}_0$ can be evaluated as:

$$\begin{aligned}&\mathbb{E}[(X_m^k - \mathbb{E}[X_m^k])^2|\mathcal{H}_0] = \boldsymbol{F}_{m,m} \bullet \Big(\sum_{t_1,t_2}\mathbb{E}[\boldsymbol{A}_{t_1,t_2}^k]\Big) \\ &= \boldsymbol{F}_{m,m} \bullet \Big(\sum_t \mathbb{E}[\boldsymbol{A}_{t,t}^k] + 2\sum_{t_2,t_1>t_2}\mathbb{E}[\boldsymbol{A}_{t_1,t_2}^k]\Big) \\ &= \sigma_\gamma^2\boldsymbol{F}_{m,m} \bullet \Big(\sum_t \overline{\boldsymbol{W}}^t + 2\sum_{t_2,t_1>t_2}\overline{\boldsymbol{W}}^{t_1}\Big) \\ &= \sigma_\gamma^2\boldsymbol{F}_{m,m} \bullet \Big(\sum_t (2t-1)\big(\mathbf{1}\mathbf{1}^\top + \sum_{i=2}^{n}\lambda_i^t(\overline{\boldsymbol{W}})\boldsymbol{v}_i\boldsymbol{v}_i^\top\big)\Big),\end{aligned} \quad (51)$$

where $\lambda_i(\overline{\boldsymbol{W}})$ is the ith largest eigenvalue of $\overline{\boldsymbol{W}}$ and $\boldsymbol{v}_i$ is the associated eigenvector. To show that the variance is bounded, we observe the following fact: (i) $\boldsymbol{F}_{m,m} \bullet \mathbf{1}\mathbf{1}^\top = 0$ for all m; (ii) $\lambda_i(\overline{\boldsymbol{W}}) < 1$ for all $i \ge 2$ and thus the associated sum is bounded by

$$\sum_{t=1}^{\infty}(2t-1)\lambda_i^t(\overline{\boldsymbol{W}}) = \frac{3\lambda_i(\overline{\boldsymbol{W}}) - 1}{(1-\lambda_i(\overline{\boldsymbol{W}}))^2} < \infty. \quad (52)$$

We conclude that under $\mathcal{H}_0$, $\mathbb{E}[(X_m^k - \mathbb{E}[X_m^k])^2] < \infty$ for all m and it can be analytically calculated by (51) & (52). Furthermore, the variance grows as

$$\tau_{im}^2 = \mathcal{O}\Big(\sigma_\gamma^2\frac{\lambda_2(\overline{\boldsymbol{W}})}{(1-\lambda_2(\overline{\boldsymbol{W}}))^2}\Big). \quad (53)$$

2) Under hypothesis $\mathcal{H}_1$ — Define $\hat{\boldsymbol{x}}^k(t) = \boldsymbol{x}^k(t) - \mathbb{E}[\boldsymbol{x}^k(t)]$ and its partition as $\hat{\boldsymbol{x}}^k(t) = (\hat{\boldsymbol{s}}^k(t)^\top, \hat{\boldsymbol{r}}^k(t)^\top)^\top$ such that $\hat{\boldsymbol{s}}^k(t), \hat{\boldsymbol{r}}^k(t)$ correspond to the malicious nodes and normal nodes, respectively. Our goal is to evaluate:

$$\mathbb{E}[\boldsymbol{A}_{t_1,t_2}^k] = \begin{bmatrix}\mathbb{E}[\hat{\boldsymbol{s}}^k(t_1)\hat{\boldsymbol{s}}^k(t_2)^\top] & \mathbb{E}[\hat{\boldsymbol{s}}^k(t_1)\hat{\boldsymbol{r}}^k(t_2)^\top] \\ \mathbb{E}[\hat{\boldsymbol{r}}^k(t_1)\hat{\boldsymbol{s}}_k(t_2)^\top] & \mathbb{E}[\hat{\boldsymbol{r}}^k(t_1)\hat{\boldsymbol{r}}^k(t_2)^\top]\end{bmatrix}. \quad (54)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSIPN.2016.2614898, IEEE Transactions on Signal and Information Processing over Networks

In Appendix G, we show:

Lemma 4 *Under $\mathcal{H}_1$ and the same settings as in Proposition 2. The expectation of $\boldsymbol{A}^k_{t_1,t_2}$ is:*

$$\mathbb{E}[\boldsymbol{A}^k_{t_1,t_2}] = \sigma_\alpha^2 \mathbf{1}\mathbf{1}^\top + \boldsymbol{\Theta}(\sigma_m^2 \min\{t_1,t_2\} \cdot \tilde{\lambda}^{\max\{t_1,t_2\}}) + \boldsymbol{\Theta}((\sigma_\gamma^2+\sigma_\alpha^2)\tilde{\lambda}^{\max\{t_1,t_2\}}) - \boldsymbol{\Xi}_{t_1,t_2}, \tag{55}$$

where $\tilde{\lambda} = \max\{\hat{\lambda}^2, \lambda_1(\boldsymbol{D})\} < 1$ and

$$\boldsymbol{\Xi}_{t_1,t_2} = \sigma_\alpha^2 \begin{pmatrix} \mathbf{0} & \mathbf{1}\mathbf{1}^\top (\boldsymbol{D}^{t_2})^\top \\ \boldsymbol{D}^{t_1}\mathbf{1}\mathbf{1}^\top & \boldsymbol{D}^{t_1}\mathbf{1}\mathbf{1}^\top + \mathbf{1}\mathbf{1}^\top(\boldsymbol{D}^{t_2})^\top \end{pmatrix}. \tag{56}$$

Using the fact that $\boldsymbol{F}_{m,m} \bullet \mathbf{1}\mathbf{1}^\top = 0$ and $\boldsymbol{F}_{m,m} \bullet \boldsymbol{\Xi}_{t_1,t_2} = 0$, the variance of X^k_m depends on the latter two terms in (55). The second last term with $\boldsymbol{\Theta}(\min\{t_1,t_2\}\tilde{\lambda}^{\max\{t_1,t_2\}})$ is bounded as

$$\begin{aligned}
&\sum_{t_1,t_2} \min\{t_1,t_2\}\tilde{\lambda}^{\max\{t_1,t_2\}} \\
&= \sum_t t\tilde{\lambda}^t + 2\sum_{t_1>t_2} t_2\tilde{\lambda}^{t_1} = \frac{\tilde{\lambda}}{(1-\tilde{\lambda})^2} + 2\sum_t\sum_\tau t\tilde{\lambda}^{t+\tau} \\
&= \frac{\tilde{\lambda}}{(1-\tilde{\lambda})^2} + 2\sum_t \frac{t\tilde{\lambda}^t}{1-\tilde{\lambda}} = \frac{\tilde{\lambda}(3-\tilde{\lambda})}{(1-\tilde{\lambda})^3} < \infty.
\end{aligned}$$

In particular, the variance of X^k_m has the following order

$$\beta^2_{im} = \mathcal{O}\left((\sigma_\gamma^2+\sigma_\alpha^2)\frac{3\tilde{\lambda}-1}{(1-\tilde{\lambda})^2} + \sigma_m^2\frac{3\tilde{\lambda}-\tilde{\lambda}^2}{(1-\tilde{\lambda})^3}\right). \tag{57}$$

This concludes the proof. It is worth mentioning that S_2^{ij} can be similarly expressed as a Chi-square r.v. with bounded variance.

Appendix D
Proof of Theorem 3

To facilitate our analysis, we define $\bar{X}_{im} := K^{-1}\sum_{k=1}^K X^k_{im}$ as the averaged statistics over the K observed instances of consensus. Moreover, let $\bar{\boldsymbol{X}}_i := (\bar{X}_{im})_{m\in\mathcal{N}_i}$ be an $|\mathcal{N}_i|$-dimensional random vector. Note that $S_1^i = |\mathcal{N}_i|^{-1}\|\bar{\boldsymbol{X}}_i\|_2^2$, whose concentration inequalities are derived below.

1) Under $\mathcal{H}_0$ — In this case, we observe that:

$$\bar{\boldsymbol{X}}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{i0}), \text{ where } \operatorname{diag}(\boldsymbol{\Sigma}_{i0}) = (\tau^2_{im}/K)_{m\in\mathcal{N}_i}. \tag{58}$$

The K^{-1} scaling in the diagonal of $\boldsymbol{\Sigma}_{i0}$ is due to the fact that X^k_{im} are independent across k. In fact, as $\boldsymbol{\Sigma}_{i0}$ must be positive semidefinite, there is also a K^{-1} scaling for every element in the matrix $\boldsymbol{\Sigma}_{i0}$. Let $\boldsymbol{\Sigma}_{i0} = \boldsymbol{Q}\operatorname{Diag}(\bar{\boldsymbol{\tau}}_i^2/K)\boldsymbol{Q}^\top$ with[2] $\bar{\boldsymbol{\tau}}_i^2 := (\bar{\tau}^2_{im})_{m\in\mathcal{N}_i}$, we can express S_1^i as:

$$S_1^i = |\mathcal{N}_i|^{-1}\sum_{m\in\mathcal{N}_i}\frac{\bar{\tau}^2_{im}}{K}(\tilde{X}_{im})^2, \tag{59}$$

where $\tilde{X}_{im} \sim \mathcal{N}(0,1)$ are independent across $m\in\mathcal{N}_i$. We have

$$\begin{aligned}
&P(S_1^i \geq \delta_{II}|\mathcal{H}_0) = P\big(\textstyle\sum_{m\in\mathcal{N}_i}\bar{\tau}^2_{im}\tilde{X}^2_{im} \geq K|\mathcal{N}_i|\delta_{II}\big) \\
&= P\big(\sum_{m\in\mathcal{N}_i}\bar{\tau}^2_{im}\tilde{X}^2_{im} \geq \|\bar{\boldsymbol{\tau}}_i^2\|_1 - 2\|\bar{\boldsymbol{\tau}}_i^2\|_2\sqrt{t^\star} + 2\|\bar{\boldsymbol{\tau}}_i^2\|_\infty t^\star\big)
\end{aligned}$$

[2]We remark that $\bar{\tau}^2_{im}$ are at the same order of τ^2_{im}.

The last term can be bounded by $\exp(-t^\star)$ using Proposition 1.1 in [21], which is due to Laurent and Massart [23]. Hence, we have:

$$\sqrt{t^\star} = \frac{\|\bar{\boldsymbol{\tau}}_i^2\|_2}{2\|\bar{\boldsymbol{\tau}}_i^2\|_\infty}\left(1+\sqrt{\frac{2\|\bar{\boldsymbol{\tau}}_i^2\|_\infty}{\|\bar{\boldsymbol{\tau}}_i^2\|_2^2}(K|\mathcal{N}_i|\delta_{II} - \|\bar{\boldsymbol{\tau}}_i^2\|_1) + 1}\right) \tag{60}$$

We consider the case when $K|\mathcal{N}_i|\delta_{II} \geq \|\bar{\boldsymbol{\tau}}_i^2\|_1$. If $K\to\infty$, then $t^\star \approx K|\mathcal{N}_i|\delta_{II}/(2\|\bar{\boldsymbol{\tau}}_i^2\|_\infty)$. Finally:

$$P(S_1^i \geq \delta_{II}|\mathcal{H}_0) \leq \exp(-K|\mathcal{N}_i|\delta_{II}/(2\|\bar{\boldsymbol{\tau}}_i^2\|_\infty)). \tag{61}$$

2) Under $\mathcal{H}_1$ — In this case, we observe that:

$$\bar{\boldsymbol{X}}_i \sim \mathcal{N}(\boldsymbol{\eta}_i, \boldsymbol{\Sigma}_{i1}), \quad \text{where} \quad \operatorname{diag}(\boldsymbol{\Sigma}_{i1}) = \left(\frac{\beta^2_{im}}{K}\right), m\in\mathcal{N}_i, \tag{62}$$

where we obtained a scaling of K^{-1} using the independence of X^k_{im} across k. Let $\boldsymbol{\Sigma}_{i1} = \boldsymbol{Q}\operatorname{Diag}(\bar{\boldsymbol{\beta}}_i^2/K)\boldsymbol{Q}^\top$ with $\bar{\boldsymbol{\beta}}_i^2 := (\bar{\beta}^2_{im})_{m=1}^{|\mathcal{N}_i|}$, $\tilde{\boldsymbol{X}}_i \sim \mathcal{N}(\mathbf{0},\boldsymbol{I})$ be an isotropic Gaussian vector with elements $\tilde{\boldsymbol{X}}_i := (\tilde{X}_{im})_{m=1}^{|\mathcal{N}_i|}$ and $\bar{\boldsymbol{\eta}}_i = \boldsymbol{Q}^\top\boldsymbol{\eta}_i$, we observe the following chain:

$$\begin{aligned}
S_1^i &= |\mathcal{N}_i|^{-1}\|\operatorname{Diag}\left(\bar{\boldsymbol{\beta}}_i/\sqrt{K}\right)\tilde{\boldsymbol{X}}_i + \bar{\boldsymbol{\eta}}_i\|_2^2 \\
&= |\mathcal{N}_i|^{-1}\sum_{m=1}^{|\mathcal{N}_i|}\left(\frac{\bar{\beta}_{im}}{\sqrt{K}}\tilde{X}_{im} + \bar{\eta}_{im}\right)^2
\end{aligned} \tag{63}$$

Let N_{i1} be the rank of $\boldsymbol{\Sigma}_{i1}$ such that $\bar{\beta}_{im} = 0$ for all $m \geq N_{i1}+1$. Observe that S_1^i can now be regarded as a sum of N_{i1} independent random variables. Recalling the generic Chernoff's bound which states that for independent random variables $X_1,...,X_n$, it follows:

$$P(X_1+\ldots+X_n \leq a) \leq e^{at}\prod_{i=1}^n \mathbb{E}\left[e^{-tX_i}\right]. \tag{64}$$

Moreover, we observe that for $\bar{\beta}^2_{im} > 0$ and all $t < 1/2$:

$$\mathbb{E}\left[e^{t(\tilde{X}_{im}+\frac{\sqrt{K}}{\bar{\beta}_{im}}\bar{\eta}_{im})}\right] = (1-2t)^{-\frac{1}{2}}\exp\left(\frac{K\bar{\eta}^2_{im}t}{(1-2t)\bar{\beta}^2_{im}}\right).$$

Plugging the above into (64), we can upper bound $P(S_1^i \leq \delta_{II}|\mathcal{H}_1)$ by:

$$\exp\left(\delta\tilde{t} - \sum_{m=1}^{N_{i1}}\frac{\bar{\eta}^2_{im}\tilde{t}}{1+2\bar{\beta}^2_{im}\tilde{t}/K}\right)\prod_{m=1}^{N_{i1}}(1+2\bar{\beta}^2_{im}\tilde{t}/K)^{-\frac{1}{2}}, \tag{65}$$

for any $\tilde{t} > 0$ and we have set $\delta = |\mathcal{N}_i|\delta_{II} - \sum_{m=N_{i1}}^{|\mathcal{N}_i|}\bar{\eta}^2_{im}$. In particular, setting $t = K/2\bar{\beta}^2_{i^{\min}}$ yields:

$$\begin{aligned}
&P(S_1^i \leq \delta_{II}|\mathcal{H}_1) \leq \\
&\exp\left(-K\max\left\{0, -\delta + \sum_{m=1}^{N_{i1}}\frac{\bar{\eta}^2_{im}}{1+\bar{\beta}^2_{im}/\bar{\beta}^2_{i^{\min}}}\right\}\right).
\end{aligned} \tag{66}$$

Appendix E
Proof of Lemma 2

The test statistics S_2^{ij} can be written as

$$S_2^{ij} = (\tilde{\eta}_{ij} + \tilde{X}_{ij})^2,$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSIPN.2016.2614898, IEEE Transactions on Signal and Information Processing over Networks

where $\tilde{X}_{ij}$ is a zero-mean Gaussian r.v. with variance bounded by $\tilde{\beta}_{ij}^2/K$. The false alarm and miss detection probability can be bounded by evaluating:

$$P_{lf}^{ij} = P(S_2^{ij} \geq \epsilon_{II} | \mathcal{H}_0^{ij}),\ 1 - P_{ld}^{ij} = P(S_2^{ij} \leq \epsilon_{II} | \mathcal{H}_1^{ij}).$$

The desirable bounds can be obtained straightforwardly using the definition of the Q-function.

Appendix F
Proof of Lemma 3

Our goal is to bound the following probability:

$$P\Big(\hat{\mathcal{H}} = \mathcal{H}_1^i \mid \mathcal{H}_1^i\Big) = P\Big(\sum_{m\in\mathcal{N}_i} |\xi_{im} - \bar{\xi}_i| \geq \delta_I \mid \mathcal{H}_1^i\Big) \quad (67)$$

Under $\mathcal{H}_1^i$, each of $\xi_{im} - \bar{\xi}_i$ is an r.v. with mean μ_i and sub-Gaussian parameter of σ_i^2/K. Similar to Appendix A, we can write $\xi_{im} - \bar{\xi}_i = \tilde{\xi}_{im} + \mu_i$. Moreover, using the inequality $\sum_{m\in\mathcal{N}_i} |\xi_{im} - \bar{\xi}_i| \leq \sum_{m\in\mathcal{N}_i} |\xi_{im}| + |\mathcal{N}_i||\mu_i|$, we obtain the upper bound:

$$\begin{aligned} P\Big(\hat{\mathcal{H}} = \mathcal{H}_1^i \mid \mathcal{H}_1^i\Big) &\leq P\Big(\sum_{m\in\mathcal{N}_i} |\xi_{im}| \geq \delta_I - |\mathcal{N}_i||\mu_i|\Big) \\ &\leq |\mathcal{N}_i| \cdot P\big(|\xi_{im}| \geq |\mathcal{N}_i|^{-1}\delta_I - |\mu_i|\big) \\ &\leq 2|\mathcal{N}_i| \cdot \exp\Big(-K\frac{(\max\{0, |\mathcal{N}_i|^{-1}\delta_I - |\mu_i|\})^2}{2\sigma_i^2}\Big), \end{aligned} \quad (68)$$

where the last inequality is due to Chernoff's inequality. This concludes our proof.

Appendix G
Proof of Lemma 4

Recall that

$$\hat{\boldsymbol{s}}^k(t) = z_\alpha^k \mathbf{1} + \boldsymbol{m}^k(t),$$

where $z_\alpha^k := \alpha^k - \bar{\alpha} \sim \mathcal{N}(0, \sigma_\alpha^2)$. Denote $\boldsymbol{\Phi}(t,s) = \boldsymbol{D}(t)\boldsymbol{D}(t-1)...\boldsymbol{D}(s)$ and $\boldsymbol{z}_\gamma^k := \boldsymbol{\gamma}^k - \bar{\gamma}\mathbf{1}$, we can write

$$\hat{\boldsymbol{r}}^k(t) = \sum_{s=0}^{t-1} \boldsymbol{\Phi}(t-1, s+1)\boldsymbol{B}(s)(z_\alpha^k \mathbf{1} + \boldsymbol{m}^k(s)) + \boldsymbol{\Phi}(t-1,0)\boldsymbol{z}_\gamma^k.$$

The top-left block in (54) can be evaluated as:

$$\mathbb{E}[\hat{\boldsymbol{s}}^k(t_1)\hat{\boldsymbol{s}}^k(t_2)^\top] = \sigma_\alpha^2 \mathbf{1}\mathbf{1}^\top + \delta(t_1 - t_2)(\hat{\lambda}^{t_1}\sigma_m)^2 \mathbf{I}, \quad (69)$$

Then, the top-right and bottom-left blocks are decomposed as:

$$\mathbb{E}[\hat{\boldsymbol{r}}^k(t_1)\hat{\boldsymbol{s}}^k(t_2)^\top] = \mathbb{E}[z_\alpha^k \hat{\boldsymbol{r}}^k(t_1)\mathbf{1}^\top] + \mathbb{E}[\hat{\boldsymbol{r}}^k(t_1)\boldsymbol{m}^k(t_2)^\top].$$

Using the fact $\boldsymbol{B}(s)\mathbf{1} = (\boldsymbol{I} - \boldsymbol{D}(s))\mathbf{1}$, we note that

$$\begin{aligned} &\textstyle\sum_{s=0}^{t_1-1} \boldsymbol{\Phi}(t_1 - 1, s+1)\boldsymbol{B}(s)\mathbf{1} \\ &= \textstyle\sum_{s=0}^{t_1-1} (\boldsymbol{\Phi}(t_1 - 1, s+1) - \boldsymbol{\Phi}(t_1 - 1, s))\mathbf{1} \\ &= \mathbf{1} - \boldsymbol{\Phi}(t_1 - 1, 0)\mathbf{1}. \end{aligned} \quad (70)$$

As $z_\alpha^k, \boldsymbol{z}_\gamma^k, \boldsymbol{m}^k(t)$ are mutually independent, taking the expectation gives

$$\mathbb{E}[z_\alpha^k \hat{\boldsymbol{r}}^k(t_1)\mathbf{1}^\top] = \sigma_\alpha^2(\boldsymbol{I} - \boldsymbol{D}^{t_1})\mathbf{1}\mathbf{1}^\top, \quad (71)$$

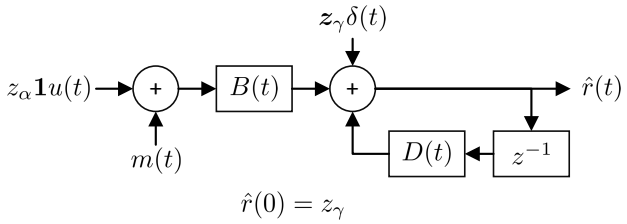

Fig. 20: Linear System for $\hat{\boldsymbol{r}}(t)$.

Moreover, if $t_1 \geq t_2$,

$$\begin{aligned} &\mathbb{E}[\hat{\boldsymbol{r}}^k(t_1)\boldsymbol{m}^k(t_2)^\top] \\ &= (\hat{\lambda}^{t_2}\sigma_m)^2 u(t_1 - t_2 - 1)\mathbb{E}[\boldsymbol{\Phi}(t_1 - 1, t_2 + 1)\boldsymbol{B}(t_2)] \\ &= (\hat{\lambda}^{t_2}\sigma_m)^2 u(t_1 - t_2 - 1)\boldsymbol{D}^{t_1 - t_2 - 1}\boldsymbol{B}, \end{aligned} \quad (72)$$

where $u(t)$ is the unit step function such that $u(t) = 1$ for all $t \geq 0$ and is zero otherwise. In general, the term above can be bounded by $\mathcal{O}(\sigma_m^2 \lambda_1(\boldsymbol{D})^{\max\{t_1,t_2\}})$. As such,

$$\begin{aligned} \mathbb{E}[\hat{\boldsymbol{r}}^k(t_1)\hat{\boldsymbol{s}}^k(t_2)^\top] &= \sigma_\alpha^2 \mathbf{1}\mathbf{1}^\top - \sigma_\alpha^2 \boldsymbol{D}^{t_1}\mathbf{1}\mathbf{1}^\top \\ &\quad + \Theta(\sigma_m^2 \lambda_1(\boldsymbol{D})^{\max\{t_1,t_2\}}) \end{aligned} \quad (73)$$

Finally, we compute the bottom-right block $\mathbb{E}[\hat{\boldsymbol{r}}^k(t_1)\hat{\boldsymbol{r}}^k(t_2)^\top]$, i.e., the covariance of $\hat{\boldsymbol{r}}(t)$. Observe that $\hat{\boldsymbol{r}}(t)$ can be viewed as the output of a linear system as shown in Figure 20, with the input:

$$\delta(t)\boldsymbol{z}_\gamma^k + z_\alpha^k \boldsymbol{B}(t)\mathbf{1} + \boldsymbol{B}(t)\boldsymbol{m}^k(t) \quad (74)$$

Importantly, $\hat{\boldsymbol{r}}^k(t)$ can be expressed as the superposition of the responses to the three input signals above. For $t \geq 1$:

$$\begin{aligned} \hat{\boldsymbol{r}}^k(t) = &\underbrace{\boldsymbol{\Phi}(t-1,0)\boldsymbol{z}_\gamma^k}_{\hat{\boldsymbol{r}}_1^k(t)} + \underbrace{z_\alpha^k \sum_{s=0}^{t-1} \boldsymbol{\Phi}(t-1, s+1)\boldsymbol{B}(s)\mathbf{1}}_{\hat{\boldsymbol{r}}_2^k(t)} \\ &+ \underbrace{\sum_{s=0}^{t-1} \boldsymbol{\Phi}(t-1, s+1)\boldsymbol{B}(s)\boldsymbol{m}^k(s)}_{\hat{\boldsymbol{r}}_3^k(t)}, \end{aligned} \quad (75)$$

and $\hat{\boldsymbol{r}}^k(0) = \mathbf{0}$. The output signals $\hat{\boldsymbol{r}}_1^k(t), \hat{\boldsymbol{r}}_2^k(t), \hat{\boldsymbol{r}}_3^k(t)$ correspond to the input signal $\delta(t)\boldsymbol{z}_\gamma^k$, $z_\alpha^k \boldsymbol{B}(t)\mathbf{1}$ and $\boldsymbol{B}(t)\boldsymbol{m}^k(t)$, respectively. It is obvious that $\hat{\boldsymbol{r}}_1^k(t), \hat{\boldsymbol{r}}_2^k(t), \hat{\boldsymbol{r}}_3^k(t)$ are mutually independent. As such, the covariance can be decomposed as

$$\begin{aligned} &\mathbb{E}[\hat{\boldsymbol{r}}^k(t_1)\hat{\boldsymbol{r}}^k(t_2)^\top] \\ =&\mathbb{E}[\hat{\boldsymbol{r}}_1^k(t_1)\hat{\boldsymbol{r}}_1^k(t_2)^\top] + \mathbb{E}[\hat{\boldsymbol{r}}_2^k(t_1)\hat{\boldsymbol{r}}_2^k(t_2)^\top] + \mathbb{E}[\hat{\boldsymbol{r}}_3^k(t_1)\hat{\boldsymbol{r}}_3^k(t_2)^\top]. \end{aligned} \quad (76)$$

Consider the following chain for $\mathbb{E}[\hat{\boldsymbol{r}}_1^k(t_1)(\hat{\boldsymbol{r}}_1^k(t_2))^\top]$:

$$\begin{aligned} &\mathrm{vec}\left(\mathbb{E}[\hat{\boldsymbol{r}}_1^k(t_1)(\hat{\boldsymbol{r}}_1^k(t_2))^\top]\right) \\ &= \sigma_\gamma^2 \left(\boldsymbol{I} \otimes \boldsymbol{D}\right)^{t_1 - t_2} \left(\mathbb{E}\left[\boldsymbol{D}(t) \otimes \boldsymbol{D}(t)\right]\right)^{t_2} \mathrm{vec}\left(\boldsymbol{I}\right) \\ &= \boldsymbol{\Theta}(\sigma_\gamma^2 \lambda_1(\boldsymbol{D})^{\max\{t_1,t_2\}}). \end{aligned} \quad (77)$$

where we have used the identity $\mathrm{vec}(\boldsymbol{AXB}) = (\boldsymbol{B}^\top \otimes \boldsymbol{A})\mathrm{vec}(\boldsymbol{X})$ recursively and the fact that $\lambda_1\left(\mathbb{E}\left[\boldsymbol{D}(t) \otimes \boldsymbol{D}(t)\right]\right) \leq \lambda_1(\boldsymbol{I} \otimes \boldsymbol{D}) = \lambda_1(\boldsymbol{D}) < 1$.

Next, using (70), we have

$$\hat{\boldsymbol{r}}_2^k(t) = z_\alpha^k(\mathbf{1} - \boldsymbol{\Phi}(t-1,0)\mathbf{1}). \quad (78)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSIPN.2016.2614898, IEEE Transactions on Signal and Information Processing over Networks

Therefore,

$$\begin{aligned}&\mathbb{E}[\hat{\boldsymbol{r}}_2^k(t_1)\hat{\boldsymbol{r}}_2^k(t_2)^\top]\\&=\sigma_\alpha^2\mathbb{E}[(\mathbf{1}-\boldsymbol{\Phi}(t_1-1,0)\mathbf{1})(\mathbf{1}-\boldsymbol{\Phi}(t_2-1,0)\mathbf{1})^\top]\\&=\sigma_\alpha^2\big(\mathbf{1}\mathbf{1}^\top-\boldsymbol{D}^{t_1}\mathbf{1}\mathbf{1}^\top-\mathbf{1}(\boldsymbol{D}^{t_2}\mathbf{1})^\top\\&\qquad+\mathbb{E}[\boldsymbol{\Phi}(t_1-1,0)\mathbf{1}\mathbf{1}^\top\boldsymbol{\Phi}(t_2-1,0)^\top]\big)\end{aligned}\tag{79}$$

Similar to (77), the last term above can be bounded as

$$\begin{aligned}&\operatorname{vec}\left(\mathbb{E}[\boldsymbol{\Phi}(t_1-1,0)\mathbf{1}\mathbf{1}^\top\boldsymbol{\Phi}(t_2-1,0)^\top]\right)\\&=(\boldsymbol{I}\otimes\boldsymbol{D})^{t_1-t_2}\left(\mathbb{E}\left[\boldsymbol{D}(t)\otimes\boldsymbol{D}(t)\right]\right)^{t_2}\operatorname{vec}\left(\mathbf{1}\mathbf{1}^\top\right)\\&=\boldsymbol{\Theta}(\lambda_1(\boldsymbol{D})^{\max\{t_1,t_2\}}).\end{aligned}\tag{80}$$

We finally consider the covariance of output due to $\boldsymbol{m}^k(t)$:

$$\mathbb{E}[\hat{\boldsymbol{r}}_3^k(t_1+1)(\hat{\boldsymbol{r}}_3^k(t_2+1))^\top]=$$
$$\mathbb{E}\left[\sum_{s=0}^{t_2}(\hat{\lambda}^s\sigma_m)^2\boldsymbol{\Phi}(t_1,s+1)\boldsymbol{B}(s)\boldsymbol{B}^\top(s)\boldsymbol{\Phi}^\top(t_2,s+1)\right],$$

where we have used the fact that $\boldsymbol{m}^k(s)$ is independent of $\boldsymbol{m}^k(s')$ for $s\neq s'$. Again, vectorizing the term above yields

$$\begin{aligned}&\sum_{s=0}^{t_2-1}(\hat{\lambda}^s\sigma_m)^2\left(\boldsymbol{I}\otimes\boldsymbol{D}\right)^{t_1-t_2}\left(\mathbb{E}\left[\boldsymbol{D}(t)\otimes\boldsymbol{D}(t)\right]\right)^{t_2-s-1}\tilde{\boldsymbol{b}}\\&=\boldsymbol{\Theta}(\sigma_m^2\min\{t_1,t_2\}\cdot\max\{\hat{\lambda}^2,\lambda_1(\boldsymbol{D})\}^{\max\{t_1,t_2\}})\end{aligned}\tag{81}$$

where $\tilde{\boldsymbol{b}}=\operatorname{vec}(\mathbb{E}[\boldsymbol{B}(s)\otimes\boldsymbol{B}(s)])$. Combining (77), (79) and (81) give:

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{r}}^k(t_1)\hat{\boldsymbol{r}}^k(t_2)^\top]&=\sigma_\alpha^2(\mathbf{1}\mathbf{1}^\top-\boldsymbol{D}^{t_1}\mathbf{1}\mathbf{1}^\top-\mathbf{1}(\boldsymbol{D}^{t_2}\mathbf{1})^\top)\\&+\boldsymbol{\Theta}((\sigma_\alpha^2+\sigma_\gamma^2)\lambda_1(\boldsymbol{D})^{\max\{t_1,t_2\}})\\&+\boldsymbol{\Theta}(\sigma_m^2\min\{t_1,t_2\}\cdot\max\{\hat{\lambda}^2,\lambda_1(\boldsymbol{D})\}^{\max\{t_1,t_2\}}).\end{aligned}$$

REFERENCES

[1] R. Gentz, H.-T. Wai, A. Scaglione, and A. Leshem, "Detection of data-injection attacks in decentralized learning," *Asilomar Conf*, 2015.

[2] A. Perrig, R. Szewczyk, J. D. Tygar, V. Wen, and D. E. Culler, "Spins: Security protocols for sensor networks," *Wireless Networks*, vol. 8, no. 5, pp. 521–534, Sep. 2002.

[3] S. Zhu, S. Setia, and S. Jajodia, "Leap: Efficient security mechanisms for large-scale distributed sensor networks," in *Proc CCS '03*, 2003, pp. 62–72.

[4] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2508–2530, Jun. 2006.

[5] M. DeGroot, "Reaching a consensus," in *Journal of American Statistcal Association*, vol. 69, 1974, pp. 118–121.

[6] M. E. Yildiz and A. Scaglione, "Computing along routes via gossiping," *IEEE Trans. on Signal Process.*, vol. 58, no. 6, pp. 3313–3327, 2010.

[7] A. Waagen, G. Verma, K. Chan, A. Swami, and R. D'Souza, "Effect of zealotry in high-dimensional opinion dynamics models," *Physical Review E*, February 2015.

[8] M. Ramos, J. Shao, S. D. S. Reis, C. Anteneodo, J. S. A. Jr, S. Havlin, and H. A. Makse, "How does public opinion become extreme?" *Sci. Rep.*, no. 10032, May 2015.

[9] M. Mobilia, "Does a single zealot affect an infinite group of voters ?" *Physical Review Letters*, July 2003.

[10] E. Yildiz, D. Acemoglu, A. Ozdaglar, A. Saberi, and A. Scaglione, "Discrete opinion dynamics with stubborn agents," *SSRN eLibrary*, 2011.

[11] H.-T. Wai, A. Scaglione, and A. Leshem, "Active Sensing of Social Networks," accepted by *IEEE Trans. Sig. and Inf. Proc. over Networks*, Mar. 2016.

[12] D. Acemoglu, G. Como, F. Fagnani, and A. Ozdaglar, "Opinion Fluctuations and Disagreement in Social Networks," *Mathematics of Operations Research*, vol. 38, no. 1, pp. 1–27, Feb. 2013.

[13] A. Dimakis, S. Kar, J. Moura, M. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing," *Proceedings of the IEEE*, vol. 98, no. 11, pp. 1847–1864, Nov 2010.

[14] Q. Yan, M. Li, T. Jiang, W. Lou, and Y. Hou, "Vulnerability and protection for distributed consensus-based spectrum sensing in cognitive radio networks," in *Proc INFOCOM 2012*, March 2012, pp. 900–908.

[15] B. Kailkhura, S. Brahma, and P. K. Varshney, "Consensus based detection in the presence of data falsification attacks," *arXiv preprint arXiv:1504.03413*, 2015.

[16] T. C. Aysal, M. E. Yildiz, A. D. Sarwate, and A. Scaglione, "Broadcast gossip algorithms for consensus," *IEEE Trans. on Signal Process.*, vol. 57, no. 7, pp. 2748–2761, 2009.

[17] H. Minc, *Nonnegative Matrices*. Wiley, 1974.

[18] W. Ben-Ameur, P. Bianchi, and J. Jakubowicz, "Robust Average Consensus using Total Variation Gossip Algorithm," in *VALUETOOLS*, 2012, pp. 99–106.

[19] S. Sundaram and B. Gharesifard, "Distributed optimization under adversarial nodes," 2016. [Online]. Available: http://arxiv.org/abs/1606.08939

[20] M. Rudelson and R. Vershynin, "Hanson-wright inequality and sub-gaussian concentration," *Electron. Commun. Probab.*, vol. 18, no. 82, 2013.

[21] D. Hsu, S. M. Kakade, and T. Zhang, "A tail inequality for quadratic forms of subgaussian random vectors," *Electron. Commun. Probab.*, vol. 17, no. 52, 2012.

[22] P. Massart, *Concentration Inequalities and Model Selection*. Springer, 2003.

[23] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Annals of Statistics*, vol. 28, no. 5, pp. 1302–1338, 2000.