Citation for the original published paper:

Vuković, O., Dán, G. (2013)
Detection and Localization of Targeted Attacks on Fully Distributed Power System State Estimation.
In: *Proceedings of IEEE SmartGridComm, October 2010*

Detection and Localization of Targeted Attacks on Fully Distributed Power System State Estimation

Ognjen Vuković and György Dán
Laboratory for Communication Networks
School of Electrical Engineering
KTH Royal Institute of Technology, Stockholm, Sweden
Email: {vukovic,gyuri}@ee.kth.se

***Abstract*—Distributed state estimation will play a central role in the efficient and reliable operation of interconnected power systems. Therefore, its security is of major concern. In this work we show that an attacker that compromises a single control center in an interconnected system could launch a denial of service attack against state-of-the-art distributed state estimation by injecting false data, and consequently, it could blind the entire system. We propose a fully distributed attack detection scheme based on local measurements to detect such a denial of service attack. We then propose a fully distributed attack localization scheme that relies on the regions' beliefs about the attack location, and performs inference on the power system topology to identify the most likely attack location. We validate both algorithms on the IEEE 118 bus power system.**

I. INTRODUCTION

Modern power systems are becoming increasingly interconnected in order to improve their operational efficiency. In the future smart grid, it is expected that interconnected power systems become even more prevalent, and that their control and supervision becomes fully distributed, without any central coordinator. The goal of each independent operator of such an interconnected system is to operate its system in an efficient and reliable way, and to preserve the system's stability. Since the stability of an entire interconnected power system depends on the stability of its constituent regions, e.g., as shown by the cascading failures that led to the 2003 North-East blackout in the U.S., it is important that the operators cooperate and exchange timely and accurate information about their systems. However, to safeguard confidential information, the cooperation and information exchange between the operators is limited.

One of the most fundamental applications that requires coordination between operators is power system state estimation (SE). SE is essential for maintaining the power system stable and for operating it efficiently. The SE uses noisy field measurements acquired by the Supervisory Control and Data Acquisition (SCADA) system and a steady-state model of the power flows in the physical system to provide an accurate estimate of the state of the system [1], [2]. The SE is a core component of the Energy Management System (EMS): the output of the SE is used by various EMS applications. Examples include contingency analysis, which evaluates how an outage would affect the system, and optimal power flow, which computes the optimal generation profile based on some criteria such as minimization of generation costs.

In the case of distributed SE, each independent operator performs the SE of its region, but it needs to cooperate with the operators of neighboring regions so that a consistent and correct estimate of power flows on the lines connecting two regions can be obtained. The resulting fully distributed SE (DSE) [3], [4], [5], [6] are effectively extensions of the basic SE algorithm [1], [2], and they obtain a consistent state estimate for the entire interconnected power system. The DSE typically requires that every operator exchanges only partial information about the state of its system.

The central role that SE plays in power system operations makes its security a major concern. The security of standalone SEs against so called stealth attacks on the measurements acquired by the SCADA system has been widely studied [7], [8], [9], [10], [11], [12], [13], [14], [15]. Stealth attacks are false measurement data injection attacks that bypass the model-based bad data detection used in the SE [7]. Various mitigation schemes against stealth attacks have been proposed: protection of individual data [9], changes to the model-based detection [10], and the protection of the SCADA infrastructure [11], [12]. Detection of stealth attacks along with state recovery has been studied in [15].

The security of DSE against false data injection attacks on the exchanged data between neighboring operators has been studied in [16]. It was shown that such attacks can disable the DSE (prevent it from finding a correct estimate). Furthermore, a detection scheme was proposed to detect such attacks along with a simple mitigation scheme where the operators perform a local state estimation upon detecting the attacks. However, by using such a mitigation scheme, the power flows connecting any two regions cannot be correctly estimated.

In this work we address the detection and localization of false data injection attacks on the DSE. We consider an attacker that compromises a single control center so that it can manipulate the data exchanged between the control center and its neighbors. We apply the attack to one of the most recent DSE algorithms [6] (outlined in Section II), and show that the attack can effectively disable the DSE (Section III). We propose an algorithm to detect the attacks by identifying discrepancies between the evolution of manipulated data (Section IV). Furthermore, we propose a distributed algorithm to locate the attacks, which relies on the regions' beliefs about the attack location, and performs graphical inference on the power system topology to identify the most likely attack location. To the best of our knowledge this is the first work to propose a mitigation scheme for denial of service attacks against power system DSE.

II. System Model and State Estimation

We consider an inter-connected power system that consists of several control areas, which we call regions. We denote the set of regions by $\mathcal{R}$, and use $|\mathcal{R}| = R$. A region $r \in \mathcal{R}$ includes a subset of all buses, and a subset of the transmission lines. Regions have no common buses, but there are shared transmission lines, which connect two regions. We refer to the shared transmission lines as *tie lines*, and to the buses connected by these lines as *border buses*.

We consider models of the active power injections at every bus, and active power flows on transmission lines [1], [2]. The active power injection and flow measurements taken in region r are denoted by the vector $z_r \in \mathbb{R}^{M_r}$, where M_r is the number of measurements in region r. The measurements equal to the actual power injections/flows plus independent random measurement noise. The noise is usually assumed to have a Gaussian distribution of zero mean. We denote by W_r the diagonal measurement covariance matrix.

The state-estimation problem consists of estimating voltage phase angles at the buses given the power flow and injection measurement vector [2]. Typically, one (arbitrary) phase angle is selected as the reference angle, and its value is fixed to an arbitrary value, e.g., zero. We describe the phase angles to be estimated in region r by the state vector x_r, and we refer to a component of the vector x_r as a *state variable*. The state variables of the vector x_r correspond to the phase angles on buses that belong to region r, and to the phase angles on border buses in other regions that are needed to describe the measurements on the tie lines and to describe power injection measurements at border buses in region r. Consequently, the sets of state variables defined by vectors x_r, $\forall r \in \mathcal{R}$, are overlapping. We denote by $x_{r,r'}$ the vector of state variables estimated in region r that correspond to state variables shared between regions r and r'. Observe that all components in the vector $x_{r,r'}$ are also contained in the vector x_r. We say that region r and region r' are neighbors if the vector $x_{r,r'}$ is non empty, and we denote the set of all neighbors of region r by $\mathcal{N}(r)$ ($|\mathcal{N}(r)| = N(r)$). For convenience, we introduce the vector $x_{r,b}$ for all state variables that are shared with the neighboring regions $\mathcal{N}(r)$, i.e., all components in the vectors $x_{r,r'}, \forall r' \in \mathcal{N}(r)$ form the vector $x_{r,b}$. The vectors $x_{r',r}$ and $x_{b,r}$ can be defined in a similar way.

A. Distributed State Estimation (DSE)

In an inter-connected power system, the control center of each region obtains a state estimate of its part of the system using measurements from its region and a model of its region. A control center needs to estimate only those phase angles that are necessary to describe its measurements, but it cooperates with neighboring control centers, typically by exchanging the state variables of the border buses, to ensure that the power flows on the tie lines are correctly estimated. In most of the recently proposed DSE algorithms, e.g., [3], [4], [5], [6], state variables are exchanged at the beginning or at the end of every iteration, and are used as an input when calculating the next state vector update. For the purpose of our study, we consider a state of the art algorithm proposed in [6], which is highly robust and can acquire highly accurate estimates of the power flows on the tie lines. The algorithm works as follows.

The goal of the DSE is to estimate x_r in every region under the condition that the estimates of shared state variables match between neighboring regions. One (arbitrary) phase angle in the entire interconnected system is selected as the reference angle, and its value is fixed to zero. Each region estimates x_r by minimizing the squares of the weighted deviations of the estimated active power flows and injections (which are functions of x_r) from the measured values (comprehended in z_r). Therefore, the distributed state estimation problem can be formulated as,

$$\begin{aligned} \min_{x_r,\, \forall r \in \mathcal{R}} \sum_{\forall r \in \mathcal{R}} [z_r - f(x_r)]^T [W_r^{-1}] [z_r - f(x_r)] \\ s.t. \quad x_{r,r'} = x_{r',r} \quad \forall r \in \mathcal{R} \; and \, \forall r' \in \mathcal{N}(r), \end{aligned} \tag{1}$$

where $f_r(x)$ is the vector of non-linear functions describing the active power flows and power injections in region r as a function of the state vector x_r.

The constraints in (1) couple the estimation across regions. In order to have a fully distributed algorithm, auxiliary variables can be introduced so that the problem can be solved using the alternating direction method of multipliers (ADMM) [6]. The resulting iterative solution scheme is

$$\begin{aligned} x_r^{(k+1)} &= (H_r^{(k)T} W^{-1} H_r^{(k)})^{-1} (H_r^{(k)T} z_r + c D_r p_r^{(k)}),\ \forall r \\ s_r^{(k+1)} &= U_{x_r} \cdot \sum_{\forall r' \in \mathcal{N}(r)} Y_{r,r'} \cdot x_{r',r} \\ p_r^{(k+1)} &= p_r^{(k)} + s_r^{(k+1)} - \frac{1}{2}(Y_{r,b} \cdot Y_{r,b}^T \cdot x_r^{(k)} - s_r^{(k)}), \end{aligned}$$

where $c > 0$ is a predefined constant, the matrix $H_r^{(k)}$ is the Jacobian of vector $f_r(x^{(k)})$, and matrices D_r, U_{x_r}, $Y_{r,r'}$ are defined as follows. D_r is a diagonal matrix whose element $d_{i,i}$ equals to the number of regions sharing the ith component (state variable) of the vector x_r. U_{x_r} is a diagonal matrix whose elements are defined as: $u_{i,i}$ equals to the inverse of the number of regions (if greater than 0) sharing the ith component (state variable) of the vector x_r, and zero otherwise. Finally, $Y_{r,r'}$ is a matrix that determines the connection between vector x_r and vector $x_{r,r'}$, and its elements are: $y_{i,j} = 1$ if the ith element (state variable) in x_r corresponds to the jth element (state variable) in $x_{r,r'}$, and $y_{i,j} = 0$ otherwise. Consequently, we have

$$x_{r,r'} = Y_{r,r'}^T \cdot x_r \ . \tag{2}$$

Similar to (2), we introduce the matrix $Y_{r,b}$, which has a similar structure as $Y_{r,r'}$ so that we have

$$x_{r,b} = Y_{r,b}^T \cdot x_r \tag{3}$$

The matrix $Y_{b,r}$ can be defined in a similar way.

The DSE is said to converge when for some k^* the maximum change of the state variables in every region is smaller than the *convergence threshold* $\epsilon > 0$, i.e., $\forall r \in \mathcal{R}, \quad ||x_r^{(k^*+1)} - x_r^{k^*}||_\infty < \epsilon$, where $|| \cdot ||_\infty$ denotes the maximum norm of a vector. We refer to the number of iterations k^* required for convergence as the *convergence time*.

III. A DoS Attack on DSE

We consider an attacker whose goal is to disable the DSE. The attacker compromises the control center of a single region $r^a \in \mathcal{R}$ so that it can manipulate with the data exchanged between r^a and its neighbors $\mathcal{N}(r^a)$ that are used as an input to the DSE. The exchanged data are the state variables defined by the vectors $x_{r,r^a}^{(k)}$, $\forall r \in \mathcal{N}(r^a)$, and the vectors $x_{r^a,r}^{(k)}$, $\forall r \in \mathcal{N}(r^a)$. We describe the attack against the state variables sent from regions $r \in \mathcal{N}(r^a)$ to region r^a (from r^a to r) at the end of iteration k by the *attack vector* $a_{r,r^a}^{(k)}$ ($a_{r^a,r}^{(k)}$). We define the attack vector $a_{r,r^a}^{(k)}$ as the vector of phase angles whose elements correspond to the value that the attacker adds to that phase angle, or

$$\tilde{x}_{r,r^a}^{(k)} = x_{r,r^a}^{(k)} + a_{r,r^a}^{(k)}, \tag{4}$$

where $\tilde{x}_{r,r^a}^{(k)}$ is the resulting corrupted vector of state variables. The vector $\tilde{x}_{r,r^a}^{(k)}$ is used as input to the next iteration of DSE in region r^a, instead of the originally exchanged vector $x_{r,r^a}^{(k)}$. The attack vector $a_{r,r^a}^{(k)}$ can be defined in a similar way.

In the rest of this Section, we describe the attack against the state variables sent to region r^a from its neighbors $r \in \mathcal{N}(r^a)$. The attack against the state variables sent from region r^a to its neighbors can be described in a similar way, but we omit it for brevity. For convenience, we introduce the attack vector $a_{b,r^a}^{(k)}$ for the state variables that the region r^a receives from all its neighboring regions

$$a_{b,r^a}^{(k)} = [a_{r_{i_1},r^a}^{(k)T} \; a_{r_{i_2},r^a}^{(k)T} \; \dots]^T \;\; \forall r_{i_j} \in \mathcal{N}(r^a), \tag{5}$$

and the corresponding corrupted vector of state variables

$$\tilde{x}_{b,r^a}^{(k)} = x_{b,r^a}^{(k)} + a_{b,r^a}^{(k)}, \tag{6}$$

Fig. 1 illustrates an attack on a power system with three regions. Observe that $\tilde{x}_{b,r^a}^{(k)}$ is the input to iteration $k+1$ of DSE, and thus, the attack $a_{b,r^a}^{(k)}$ leads to a *corrupted* state vector $\tilde{x}_{r^a}^{(k+1)}$.

We define *the size of the attack* as the Euclidean norm of the attack vector, i.e., $||a_{b,r^a}^{(k)}||_2$. Intuitively, a smaller attack size implies smaller corruption added to the exchanged values, which could make the detection and the localization of the attack harder. Thus, it would be natural for the attacker to look for the smallest attack vector that prevents the DSE from converging ($k^* = \infty$), or formally

$$\min_{a_{b,r^a}^{(k)}, k=1,\dots} \beta \qquad \text{s.t. } k^* = \infty \text{ and } \beta = ||a_{b,r^a}^{(k)}||_2; \forall k. \tag{7}$$

Since the distributed state estimation problem is non-linear, solving (7) is non-trivial.

A. First Singular Vector Attack (FSV)

The FSV attack [16] is an approximation of (7) done by maximizing the introduced disturbances for a given attack size. Note that the attack vector $a_{b,r^a}^{(k)}$ results in corrupted vectors

$$\begin{aligned} \tilde{s}_{r^a}^{((k+1)} &= \tilde{s}_{r^a}^{((k+1)} + U_{x_r} \cdot Y_{b,r^a} \cdot a_{b,r^a}^{(k)} \\ \tilde{p}_{r^a}^{((k+1)} &= \tilde{p}_{r^a}^{((k+1)} + U_{x_r} \cdot Y_{b,r^a} \cdot a_{b,r^a}^{(k)}, \end{aligned} \tag{8}$$

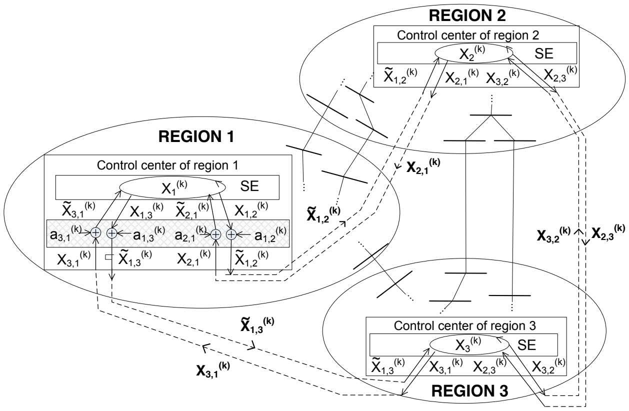

Fig. 1. Interconnected power system with three regions. The attacker corrupts the control center of Region 1, and tampers with the state variables $x_{1,2}^{(k)}$ and $x_{1,3}^{(k)}$ sent from Region 1, and the state variables $x_{2,1}^{(k)}$ and $x_{3,1}^{(k)}$ received by Region 1. The symbol (+) indicates that the components of the attack vector are added to the corresponding components (phase angles) of the vector of exchanged state variables. The attacker cannot tamper with the state variables exchanged between Regions 2 and 3.

which yield a corrupted state vector

$$\tilde{x}_{r^a}^{(k+1)} = x_{r^a}^{(k+1)} + K \cdot a_{b,r^a}^{(k)}, \tag{9}$$

where $K = (H_r^{(k)T} W^{-1} H_r^{(k)})^{-1} \cdot c D_r U_{x_r} Y_{b,r^a}$. Note that the addend in (9) is a vector with the same number of elements as the vector $x_{r^a}^{(k+1)}$, and we refer to it as the *addend vector*. The Euclidean norm of the addend vector is maximized if the attack vector $a_{b,r^a}^{(k)}$ is aligned with the first right singular vector of the matrix K, that is, with the singular vector with highest singular value. The complexity of singular vector decomposition is $O(mn^2)$ [17], low enough for the computation to be done on-line. Nevertheless, the computation of the Jacobian $H_r^{(k)}$ requires knowledge of the current system state $x_{r^a}^{(k)}$ for the attacked region r^a. Therefore, we approximate $H_r^{(k)}$ with the Jacobian calculated at the initial state $H_r^{(0)}$.

Observe that in (9) the size of the corrupted vector $\tilde{x}_{r^a}^{(k+1)}$ depends on the direction of the addend vector, and consequently, on the direction of the first singular vector. Since the attacker does not know the state vector $x_{r^a}^{(k)}$, finding the correct direction is not trivial. In order to estimate the direction, the attacker can assume that the estimates of the power flows on a tie line are closer to their actual values when using the most recent exchanged state variables. Then, the attacker applies the attack so that the introduced estimation errors take the estimates in the direction towards the previous iteration estimates, i.e., farther from the measured values.

B. Impact of FSV Attack on DSE

We show the impact of the FSV attack on the IEEE 118 bus power system, divided into six regions as shown in Fig. 2. We consider that the attacker corrupts the control center of one of the regions, and performs the attacks against the state variables sent from and to that region. Bus 69, located in region r_6, is used as the reference bus, as specified in the IEEE 118 bus power system. Measurements are taken at every power injection and power flow, and the convergence threshold is $\epsilon = 10^{-3}$. The phase angles, thereby the state variables and the attack vector, are in radians.

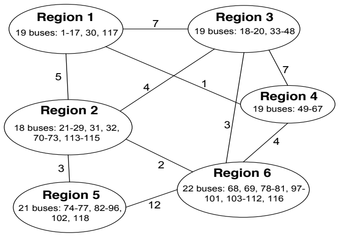

Fig. 2. IEEE 118 bus system divided into six regions. Neighboring regions are connected by a line and the number next to the line represents the number of shared state variables. Note that the reference bus (69) is not a state variable.

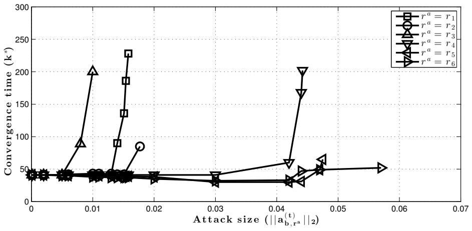

Fig. 3. Convergence time for cases when the DSE converges as a function of the attack size.

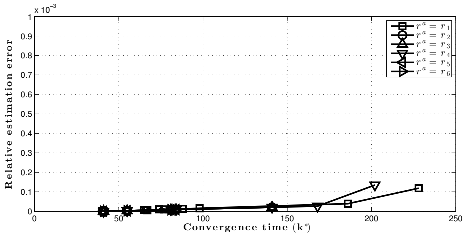

Fig. 4. Relative estimation error (50th percentile) for the upper 50% utilized power flows and injections vs. convergence time.

Fig. 3 shows the convergence time k^* (when the DSE converges) as a function of the FSV attack size considering individually each region as the attacked region. The convergence time increases with the attack size. For all considered cases, the FSV attack can prevent the DSE from converging, i.e., leads to denial of service. One might expect that the DSE is more sensitive when the region containing the reference bus is attacked, since it may be harder for other regions to synchronize with the reference bus. However, the results show that this is not the case: there is no significant difference when the region containing the reference bus is attacked (region r_6), and when some other region is attacked.

Observe that in Fig. 3 it does not take a big FSV attack to prevent the DSE from converging. For example, the FSV attack with the size $||a^{(k)}_{b,r^a}||_2 = 0.07$ prevents the DSE from converging regardless of which region is attacked. This size corresponds to the average value of the attack vector elements of 0.0265 radians (1.51 degrees) if the region r_1 is attacked, or 0.019 (1.07 degrees) if the region r_6 is attacked.

Although for small attacks the DSE converges, the estimated state and thus the estimated power flows could be erroneous. Fig. 4 shows the 50th percentile of the relative estimation error compared to the estimate with no attacks for the highest 50% the power flows as a function of convergence time (and thus the attack size). The relative estimation error increases monotonically with the convergence time, and thereby the attack size, but it is reasonably low. The DSE seems highly robust against the small attacks that admit convergence.

IV. Attack Detection and Localization

Given the potential of FSV to prevent the DSE from converging, a natural question is whether an attack can be detected and the compromised region localized. In the following, we show that both are possible. We focus on attacks that prevent the DSE from converging, since the ones that admit convergence introduce relatively low estimation errors.

Detection: Let us start by elaborating on the convergence of the DSE. Recall that in order to solve (1) in a fully distributed fashion, the right-hand side of the condition $x_{r,r'} = x_{r',r}$ is replaced with an auxiliary variable for each $r \in \mathcal{R}$ and $\forall r' \in \mathcal{N}(r)$. In iteration k and for regions r and r', the auxiliary variable equals to the average of the shared state variables between the regions, i.e., $(x^{(k)}_{r,r'} + x^{(k)}_{r',r})/2$ [6]. Consequently, the condition in (1) can be expressed as $x^{(k)}_{r,r'} = (x^{(k)}_{r,r'} + x^{(k)}_{r',r})/2$, or $(x^{(k)}_{r,r'} - x^{(k)}_{r',r})/2 = 0$. The resulting decomposed problem is solved with the ADMM, which guarantees convergence if the following criteria are satisfied (based on [18]).

Proposition 1. *If the iterative function $J_r(x_r)$ ($\forall r \in \mathcal{R}$) is closed, proper, and convex, and the augmented Lagrangian*

$$\mathcal{L} = \sum_{\forall r \in \mathcal{R}} J_r(x_r) + y^T \frac{x^{(k)}_{r,r'} - x^{(k)}_{r',r}}{2} + c || \frac{x^{(k)}_{r,r'} - x^{(k)}_{r',r}}{2} ||_2^2 \quad (10)$$

(y is Lagrange multiplier) has a saddle point, then the ADMM converges and $||(x^{(k)}_{r,r'} - x^{(k)}_{r',r})/2||_2^2 \to 0$ as $k \to \infty$ [18, Appendix A,p. 106–110].

Assuming that the conditions in Proposition 1 are satisfied, and therefore the DSE converges, it does not necessarily hold that $||(x^{(k)}_{r,r'} - x^{(k)}_{r',r})/2||_2^2$ monotonically decreases. However, for large k and when the DSE approaches a solution, we may expect that

$$||(x^{(k+1)}_{r,r'} - x^{(k+1)}_{r',r})/2||_2^2 < ||(x^{(k)}_{r,r'} - x^{(k)}_{r',r})/2||_2^2 \quad (11)$$

holds for all state variables exchanged between regions. In the following we investigate if (11) can indicate convergence problems due to an attack. Furthermore, we investigate if we can locate the region that causes the convergence problems by observing the evolution of $||(x^{(k+1)}_{r,r'} - x^{(k+1)}_{r',r})/2||_2^2$ in every region ($\forall r \in \mathcal{R}$) and for each of its neighbors ($\forall r' \in \mathcal{N}(r)$).

Definition. *Let the mean squared disagreement (MSD) between regions r and r' at iteration k be $d^{(k)}_{r,r'} = \frac{||(x^{(k)}_{r,r'} - x^{(k)}_{r',r})/2||_2^2}{|x^{(k)}_{r,r'}|}$, where $|x^{(k)}_{r,r'}|$ denotes the number of elements in vector $x^{(k)}_{r,r'}$. Observe that by definition $d^{(k)}_{r,r'} = d^{(k)}_{r',r}$.*

Figs. 5 and 6 show how the MSD $d^{(k)}_{r_6,r'}$ between region r_6 and its neighbors $r' \in \mathcal{N}(r_6)$ evolves without an attack, and with an attack in region r_2 that does not permit convergence, respectively. In the case of no attack, the MSD decreases for all $r' \in \mathcal{N}(r_6)$. In the case of an attack, the MSD for all regions is higher than in the case of no attack, i.e., all regions are affected by the attack. Moreover, the MSD oscillates for the attacked region (region r_2). Observe that not all MSDs decrease with the iterations, which is in contradiction with Proposition 1. This is the phenomenon we use to detect convergence problems as described in the following.

Proposition 2. *Let $\sup\{\cdot\}$ be the supremum of a set. If the conditions in Proposition 1 are satisfied, but for large k there*

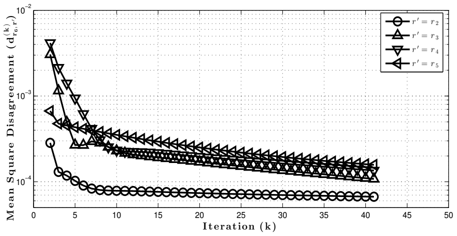

Fig. 5. Evolution of the MSDs observed in region r_6 $(d^{(k)}_{r_6,r'}, \forall r' \in \mathcal{N}(r_6))$. No attack

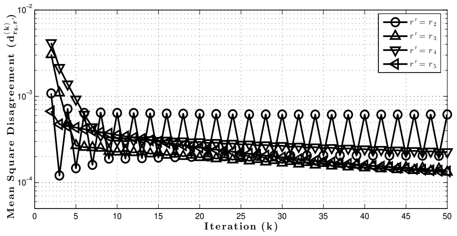

Fig. 6. Evolution of the MSDs observed in region r_6 $(d^{(k)}_{r_6,r'}, \forall r' \in \mathcal{N}(r_6))$ in presence of attacks in region $r^a = r_2$ $(r^a \in \mathcal{N}(r_6))$ with attack size 0.1.

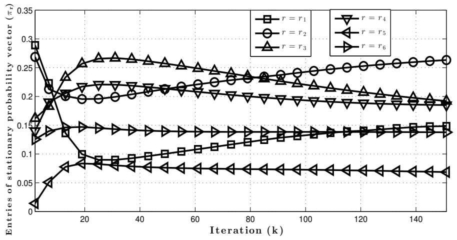

Fig. 7. Evolution of the elements of the stationary probability vector $\pi^{(k)}$ in presence of attacks in region $r^a = r_2$ with attack size 0.1. $\alpha = 0.2$.

are some r and $r' \in \mathcal{N}(r)$ such that $\sup\{d^{(k')}_{r,r'} : k' > k\} > 0$, $||x^{(k+1)}_r - x^{(k)}_r||_\infty > \epsilon$, *and* $\nexists t \in \mathbb{N}$ *so that*

$$\sup\{d^{(k')}_{r,r'} : k' > k\} > \sup\{d^{(k')}_{r,r'} : k' > k + t\} \tag{12}$$

then there is a convergence problem (an attack).

Proof: The proof follows from Proposition 1, where if the conditions hold, then $||(x^{(k)}_{r,r'} - x^{(k)}_{r',r})/2||^2_2 \to 0$ and $d^{(k)}_{r,r'} \to 0$ as $k \to \infty$, and, consequently, $\sup\{d^{(k')}_{r,r'} : k' > k\} \to 0$. ■

The regions can thus use Proposition 2 to detect an attack.

Localization: We now turn to the problem of localizing the attacked region. The localization scheme we propose consists of two steps. First, each region forms its own *belief of the attack location* (BAL) as a function of its MSDs. Second, regions exchange their BALs and use them as an input to a localization scheme at every region so that the attack can be correctly localized. The way BALs are formed and the localization scheme based on the exchanged BALs are explained as follows.

Step 1: We denote the BAL of region r that its neighbor $r' \in \mathcal{N}(r)$ is the attacked region at iteration k by $B^{(k)}_{r,r'}$. Since MSDs can oscillate and the difference between the MSD related to the attacked region ($d^{(k)}_{r,r^a}$) and others ($d^{(k)}_{r,r'}$ $r' \neq r^a$) typically increases with k (Fig. 6), we calculate the BAL $B^{(k)}_{r,r'}$ $\forall r' \in \mathcal{N}(r)$ as

$$B^{(k)}_{r,r'} = \frac{\tilde{d}^{(k)}_{r,r'}}{\sum\limits_{\forall r' \in \mathcal{N}(r)} \tilde{d}^{(k)}_{r,r'}} = \frac{\alpha d^{(k)}_{r,r'} + (1-\alpha)\tilde{d}^{(k-1)}_{r,r'}}{\sum\limits_{\forall r' \in \mathcal{N}(r)} (\alpha d^{(k)}_{r,r'} + (1-\alpha)\tilde{d}^{(k-1)}_{r,r'})}, \tag{13}$$

where $\tilde{d}^{(k)}_{r,r'}$ is the exponentially weighted moving average (EWMA) of the MSD $d^{(k)}_{r,r'}$ with the degree of weighting decrease $\alpha \in (0, 1)$. For $\forall r' \notin \mathcal{N}(r)$, $B^{(k)}_{r,r'}$ equals to 0. Observe that it is possible to have $B^{(k)}_{r,r'} \neq B^{(k)}_{r',r}$.

Step 2: Given the BALs of the regions, we want to find the probability that a particular region is attacked consistent with all BALs. This can be done by constructing the right stochastic $R \times R$ matrix $B^{(k)}$. Each row of $B^{(k)}$ corresponds to a r and contains $B^{(k)}_{r,r'}$ $\forall r' \in \mathcal{R}$. This matrix can be constructed if regions exchange their BAL vectors. The distribution that is consistent with all BALs is the stationary distribution $\pi^{(k)}$ of the Markov chain for which the matrix $B^{(k)}$ is the state transition matrix, thus $\pi^{(k)}B^{(k)} = \pi^{(k)}$ [19]. The following proposition establishes the existence and the uniqueness of the stationary distribution $\pi^{(k)}$.

Proposition 3. *Consider a system with $R > 2$. If (i) there exists a 3-clique in the graph $G = (\mathcal{R}, E)$ where $E = \{e_{r,r'} | r \in \mathcal{R}, r' \in \mathcal{N}(r)\}$, (ii) for finite k the DSE does not converge, and (iii) all regions obtain the correct matrix $B^{(k)}$, then the stationary probability vector $\pi^{(k)}$ exists, it is unique and it can be computed in every region.*

Proof: Observe that because $\alpha < 1$, it holds that $\forall r, r'$ s.t. $r \in \mathcal{N}(r')$, $B^{(k)}_{r,r'} > 0$ since from (13) it is clear that $B^{(k)}_{r,r'}$ will always take into account with non-zero weights the initial disagreements on the shared state variables and the mis-synchronization to the reference bus. Consequently, the state transition diagram of the Markov chain described by $B^{(k)}$ is a symmetric directed graph. Furthermore, it is clear that all states of such a Markov chain lie in a single communicating class, and consequently, such a chain is irreducible. Since the Markov chain is irreducible, it has a stationary distribution [19, Proposition 1.14] and it is unique [19, Corollary 1.17]. Since $B^{(k)}_{r,r} = 0$ $\forall r \in \mathcal{R}$, $R > 2$ and (i) ensure that the Markov chain is aperiodic. Finally, the aperiodicity was a sufficient condition that the (irreducible) Markov chain converges to its stationary distribution [19, Theorem 4.9]. Since all regions obtain the same matrix $B^{(k)}$, and the vector $\pi^{(k)}$ is unique, all regions obtain the same vector $\pi^{(k)}$. ■

We use the vector $\pi^{(k)}$ to locate the attacked region, as described in the following.

Conjecture 4. *Let us denote by $\pi^{(k)}_r$ the entry of vector $\pi^{(k)}$ that corresponds to region r. Let region $r^{*(k)} = \arg\max_r \pi^{(k)}_r$. Then the attacked region is $r^a = r^{*(k)}$ for large k, i.e., the stationary distribution converges.*

Observe that the localization is fully distributed: if the BALs are exchanged between all regions at iteration k, then each region can locally construct the matrix $\mathbf{B}^{(k)}$, find the vector $\pi^{(k)}$ and its maximal component.

Remark: One may argue that the attacker may corrupt the BALs sent from and to region r^a. Recall that $d^{(k)}_{r,r'} = d^{(k)}_{r',r}$, and consequently $\tilde{d}^{(k)}_{r,r'} = \tilde{d}^{(k)}_{r',r}$. Therefore, the nominator in (13) can be verified between neighbors. However, the denominator in (13) is only known to region r. Therefore, in order to prevent attacks against exchanged BALs, the regions could exchange $\tilde{d}^{(k)}_{r,r'}$ so that every region can compare if $\tilde{d}^{(k)}_{r,r'} = \tilde{d}^{(k)}_{r',r}$ holds for $\forall r \in \mathcal{R}$ and $r' \in \mathcal{N}(r)$. If there are multiple inequalities, then there will be a common region in all those inequalities

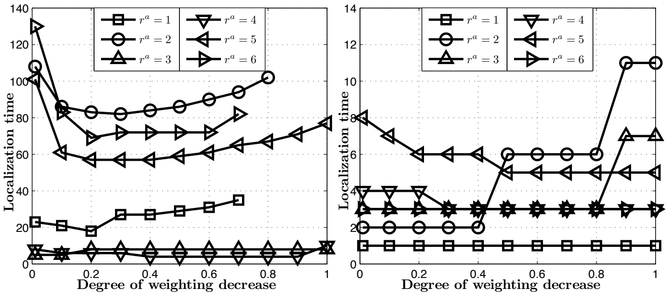

Fig. 8. The localization time (k^d) as a function of the degree of weighting decrease (α) considering the attacks in every region $\forall r \in \mathcal{R}$ individually with attack size 0.1 (left) and 0.5 (right)

from which the corrupted BALs originate, and that region is the attack location. If there is one inequality, then the attacked region is one of the two regions. In that case regions could construct two matrices $B^{(k)}$ (each considering one side of the inequality) and if the localization scheme based on $B^{(k)}$ cannot identify which of the two regions is attacked, then the mitigation scheme could consider both regions as attacked. Otherwise, the attack is localized.

Once the attacked region is localized, the DSE algorithm can be continued by non-attacked regions by discarding all data from the attacked region r^a.

Numerical results: Fig. 7 shows how the elements of the stationary probability vector $\pi^{(k)}$ evolve in the presence of attacks in region $r^a = r_2$ with attack size 0.1. The degree of weighting decrease (α) equals to 0.2. The elements of vector $\pi^{(k)}$ approach stable values as k increases, and the largest element corresponds to the attacked region r^a. We refer to the minimum iteration number k^d such that for all $k \geq k^d$ the largest element of $\pi^{(k)}$ does not change, and we call this for region r^a as the *localization time*. In the case showed in Fig. 7, the localization time is $k^d = 82$.

Fig. 8 shows how the localization time (k^d) changes with α considering the attacks in every region $r \in \mathcal{R}$ individually with attack size 0.1 (left) and 0.5 (right). The localization time significantly depends on the region that is attacked as well as on the attack size: for larger the attack size the localization is faster, which supports the formulation in (7). For most of the regions, the optimal α is in the range $\alpha \in (0.2, 0.3)$. In some cases for small attack sizes (≈ 0.1) very high $\alpha > 0.7$ prevents localization. This implies that one should use a low weighting factor to ensure successful localization.

V. Conclusion

We addressed the detection and localization of false data injection attacks against distributed state estimation. We considered an attacker that compromises a single control center, and tampers with the data that are exchanged between a control center and its neighbors. We showed that a denial of service attack can be launched against a state of the art state estimator this way. We proposed an attack detection scheme based on the evolution of the exchanged data and based on the convergence properties of the distributed algorithm. We proposed an attack localization scheme based on the steady state distribution of a Markov chain where the states are the control centers and the transition probabilities are beliefs about the attack location. We showed the efficiency of the attack detection and localization schemes via simulations on an IEEE benchmark power system. Our results show that the attacks are faster localized when most of the evolution of the exchanged data is considered, rather than just most recent values. Moreover, our results show that stronger attacks are faster localized.

Acknowledgment

This work is partially funded by the EIT ICTLabs through activity ASES 13030.

References

[1] A. Monticelli, “Electric power system state estimation,” *Proc. of the IEEE*, vol. 88, no. 2, pp. 262–282, 2000.

[2] A. Abur and A. G. Exposito, *Power System State Estimation: Theory and Implementation*. Marcel Dekker, Inc., 2004.

[3] M. Shahidehpour and Y. Wang, *Communication and Control in Electric Power Systems*. John Wiley and Sons, 2003.

[4] A. J. Conejo, S. d. Torre, and M. Canas, “An optimization approach to multiarea state estimation,” *IEEE Transactions on Power Systems*, vol. 22, no. 1, pp. 213–221, February 2007.

[5] L. Xie, D.-H. Choi, S. Kar, and H. V. Poor, “Fully distributed state estimation for wide-area monitoring systems,” *IEEE Transactions on Smart Grid*, vol. 3, no. 3, pp. 1154–1169, September 2012.

[6] V. Kekatos and G. B. Giannakis, “Distributed robust power system state estimation,” *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 1617–1626, 2013.

[7] Y. Liu, P. Ning, and M. Reiter, “False data injection attacks against state estimation in electric power grids,” in *Proc. of the 16th ACM conference on Computer and Communications Security (CCS)*, 2009, pp. 21–32.

[8] A. Teixeira, G. Dán, H. Sandberg, and K. H. Johansson, “A cyber security study of a SCADA energy management system: Stealthy deception attacks on the state estimator,” in *Proc. IFAC World Congress*, Aug. 2011.

[9] R. B. Bobba, K. M. Rogers, Q. Wang, H. Khurana, K. Nahrstedt, and T. J. Overbye, “Detecting false data injection attacks on dc state estimation,” in *Preprints of the First Workshop on Secure Control Systems, CPSWEEK*, Stockholm, Sweden, April 2010.

[10] O. Kosut, L. Jia, R. Thomas, and L. Tong, “Malicious data attacks on smart grid state estimation: Attack strategies and countermeasures,” in *Proc. of IEEE SmartGridComm*, Oct. 2010.

[11] G. Dán and H. Sandberg, “Stealth attacks and protection schemes for state estimators in power systems,” in *Proc. of IEEE SmartGridComm*, Oct. 2010.

[12] O. Vuković, K. C. Sou, G. Dán, and H. Sandberg, “Network-aware mitigation of data integrity attacks on power system state estimation,” *IEEE JSAC: Smart Grid Communications Series*, vol. 30, no. 6, pp. 176–183, July 2012.

[13] T. T. Kim and H. V. Poor, “Strategic protection against data injection attacks on power grids,” *IEEE Trans. on Smart Grid*, vol. 2, pp. 326–333, Jun. 2011.

[14] A. Giani, E. Bitar, M. Garcia, M. McQueen, P. Khargonekar, and K. Poolla, “Smart grid data integrity attacks: Characterizations and countermeasures,” in *Proc. of IEEE SmartGridComm*, Oct. 2011.

[15] A. Tajer, S. Kar, H. V. Poor, and S. Cui, “Distributed joint cyber attack detection and state recovery in smart grids,” in *Proc. of IEEE SmartGridComm*, October 2011, pp. 202–207.

[16] O. Vuković and G. Dán, “On the security of distributed power system state estimation under targeted attacks,” in *Proc. of ACM Symposium on Applied Computing (SAC)*, March 2013.

[17] R. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1990.

[18] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2010.

[19] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov Chains and Mixing Times*. American Mathematical Society, 2009.