
Article:
Esnaola, I. orcid.org/0000-0001-5597-1718, Perlaza, S.M., Poor, H.V. et al. (1 more author) (2016) Maximum Distortion Attacks in Electricity Grids. IEEE Transactions on Smart Grid, 7 (4). pp. 2007-2015. ISSN 1949-3053

https://doi.org/10.1109/TSG.2016.2550420

Maximum Distortion Attacks in Electricity Grids

Iñaki Esnaola, Samir M. Perlaza, H. Vincent Poor, and Oliver Kosut.

***Abstract*—Multiple attacker data injection attack construction in electricity grids with minimum-mean-square-error state estimation is studied for centralized and decentralized scenarios. A performance analysis of the trade-off between the maximum distortion that an attack can introduce and the probability of the attack being detected by the network operator is considered. In this setting, optimal centralized attack construction strategies are studied. The decentralized case is examined in a game-theoretic setting. A novel utility function is proposed to model this trade-off and it is shown that the resulting game is a potential game. The existence and cardinality of the corresponding set of Nash Equilibria (NEs) of the game is analyzed. Interestingly, the attackers can exploit the correlation among the state variables to facilitate the attack construction. It is shown that attackers can agree on a data injection vector construction that achieves the best trade-off between distortion and detection probability by sharing only a limited number of bits offline. For the particular case of two attackers, numerical results based on IEEE test systems are presented.**

***Index Terms*—Data-injection attacks, MMSE estimation, decentralized attacks, game theory.**

I. Introduction

The smart grid paradigm is founded on the integration of existing power systems with advanced sensing and communication infrastructures. While the benefits provided by this setting are crucial for the development of future applications and services in electricity grids, it also paves the way for cyber-security threats [1].

In this paper, data injection attacks against electricity grids are studied. The fundamental assumption of this work is that malicious attackers have access to a subset of meters and thus, are able to tamper with their measurements to distort the global state estimate [2] obtained by a network operator. This problem is first formulated in [3]. Therein, attacks are studied and construction procedures for attackers with access to a limited number of meters are presented. However, the analysis in [3] relies on algebraic tools and assumes that the detector ignores the stochastic nature of the state variables. With growing data mining and analysis capabilities provided by modern computing, it is reasonable to assume that network operators can learn the statistical structure of the system and use attack detection strategies that incorporate the underlying stochastic process governing the network. Similarly, from the attacker's perspective, data injection attacks can be formulated within a Bayesian framework in which the statistical structure of the state variables is exploited. In [4], [5], [6] and [7], the state variables are modeled as a multivariate Gaussian process whose second order moments are available to the attacker and the operator. Admittedly, restricting the analysis to Gaussian processes limits the generality of the analysis. However, Gaussian processes have successfully been used to describe a broad set of spatial and temporal dynamics in real systems. In this work, it is assumed that the second order moments of the Gaussian process modelling the state variables are known to the attacker. The rationale for this is to consider a worst-case scenario setting for the operator. In [6] an attack construction that increases the mean square error inflicted to the network operator estimates is proposed. However, this construction does not take into account the probability with which the attacker is detected. In [8] it is shown that correlation among measurements can be exploited to identify bad data and remove its effect on a modified residual test. A bad data detection procedure based on partitioning is proposed in [9]. A framework for analyzing the joint estimation and attack detection under structured data attacks is presented in [10]. Attack construction and detection with imperfect system model information are studied in [11] and [12]. For the case in which the attacker has no access to the topology of the network a novel attack construction is proposed in [13]. Alternatively, if the operator has access to training data, machine learning techniques are effective for attack detection [14]. Mitigation strategies for attacks that compromise the communication network used to deliver measurement data are studied in [15].

Given the complexity and extent of most electricity grids, it is plausible to think of scenarios in which several attackers intrude upon the network at different locations. Similarly, it is common for network operators to interconnect their grids, which results in a larger and more complex system and which is often not managed in a centralized fashion. In this scenario in which multiple attackers are present and/or limited communication is available among different instantiations of the same attacker raises the notion of distributed attacks. Within the aforementioned algebraic framework, distributed attack and detection strategies are investigated in [4], [16].

The decentralized system with different actors operating over a large number of processes poses a suitable framework for the exploration of game theoretic techniques. A comprehensive account of smart grid services and applications that can be tackled with game theory is given in [17]. In [18], centralized data injection attacks are studied in a game

Iñaki Esnaola, Samir M. Perlaza, and H. Vincent Poor are with Department of Electrical Engineering at Princeton University, Princeton, NJ 08544, USA.

Iñaki Esnaola is also with the Department of Automatic Control and Systems Engineering, University of Sheffield, S1 3JD, UK.

Samir M. Perlaza is also with the CITI Lab of the Institut National de Recherche en Informatique et Automatique (INRIA), Université de Lyon and Institut National des Sciences Appliquées (INSA) de Lyon. 6 Av. des Arts 69621 Villeurbanne, France.

Oliver Kosut is with the School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85287. (esnaola@sheffield.ac.uk, samir.perlaza@inria.fr, poor@princeton.edu, okosut@asu.edu).

This work was supported in part by the European Commission under Individual Fellowship Marie Skłodowska-Curie Action (CYBERNETS) through Grant 659316, and in part by the U.S. National Science Foundation under Grants CMMI-1435778 and ECCS-1549881.

theoretic setting in which the operator performs least squares estimation. Attack constructions that aim to manipulate market prices are modelled as a zero-sum game in [19]. However, the case in which several attackers disrupt the state estimation process in an uncoordinated way is still not well understood. Furthermore, the impact of making the statistical structure of the state variables available to attackers in decentralized settings has not been studied either.

The main results of this paper are inscribed in the context of both centralized and distributed attack construction problems. The setting assumes that the state variables are described by a multivariate Gaussian process and that the operator performs minimum-mean-square-error (MMSE) estimation over the measurements. The trade-off between the damage to the network, e.g., the excess distortion term, and the ability to remain hidden to the network operator, e.g, to keep the probability of attack detection under a given threshold is studied in both scenarios. In the former, all attackers are sufficiently coordinated to be considered as a single entity and thus, classical tools from matrix theory and optimization theory are used to determine the optimal attack. The distributed scenario considers that attackers are fully distributed. That being the case, tools from game theory are used to determine optimal individual behaviors and the resulting distributed attack construction strategies. A novel utility function that models the features of the dynamic between the attackers and the operator is proposed. The game resulting from the implementation of this utility function is studied analytically and numerically. Specifically, existence results and bounds on the number of Nash Equilibria (NEs) of the game are provided.

The next section describes the system model, including the estimation and detection procedures. Centralized attack construction strategies are discussed in Section III. The decentralized case and the properties of the resulting game are analyzed in Section IV. Section V presents simulations of the attack strategies in IEEE Test Systems. The paper ends with concluding remarks in Section VI.

II. System Model

Let $\mathbf{x} \in \mathbb{R}^N$ be a vector containing the system state variables of a given power system with N buses. Assuming linearized system dynamics with M measurements corrupted by additive white Gaussian noise, the measurement vector $\mathbf{y}_o \in \mathbb{R}^M$ is given by

$$\mathbf{y}_o = \mathbf{H}\mathbf{x} + \mathbf{z}, \tag{1}$$

where $\mathbf{H} \in \mathbb{R}^{M\times N}$ is the Jacobian of the linearized system dynamics around a given operating point and $\mathbf{z} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_M)$ is thermal white noise with power spectral density σ^2. The data-injection attack $\mathbf{a}$ is an M-dimensional deterministic vector introduced by an external attacker. The attacker interferes with the measurements and modifies the observation model to

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{z} + \mathbf{a}, \tag{2}$$

where $\mathbf{y} \in \mathbb{R}^M$ are the measurements that have been corrupted by the data-injection attack.

A. State Estimation and Data-Injection Attacks

The aim of the network operator is to obtain an estimate $\hat{\mathbf{x}}$ of the state vector $\mathbf{x}$ using the observations $\mathbf{y}$. In general, linear estimators are privileged due to their simplicity and thus, the estimate can be obtained as $\hat{\mathbf{x}} = \mathbf{L}\mathbf{y}$, given a linear estimator matrix $\mathbf{L}$. In the case in which the operator knows the underlying random process governing the state of the network, the estimation can be performed aiming to minimize the mean square error (MSE). That is, the network operator uses an estimator $\mathbf{M}$ that is the unique solution to the following optimization problem:

$$\mathbf{M} = \arg \min_{\mathbf{L}\in\mathbb{R}^{M\times M}} \mathsf{E}\left[\frac{1}{N}\|\mathbf{x} - \mathbf{L}\mathbf{y}\|_2^2\right], \tag{3}$$

where the expectation is taken with respect to $\mathbf{x}$ and $\mathbf{z}$. Under the assumption that the network state vector $\mathbf{x}$ follows an N-dimensional real Gaussian distribution with zero mean and covariance matrix $\mathbf{\Sigma_{xx}}$, the MMSE estimation matrix is

$$\mathbf{M} = \mathbf{\Sigma_{xx}}\mathbf{H}^\mathsf{T}(\mathbf{H}\mathbf{\Sigma_{xx}}\mathbf{H}^\mathsf{T} + \sigma^2\mathbf{I})^{-1}, \tag{4}$$

and the MMSE estimate of the state vector $\mathbf{x}$ is

$$\hat{\mathbf{x}}_{\mathsf{MMSE}} \triangleq \mathbf{M}\mathbf{y}. \tag{5}$$

The aim of an attacker is to choose a data-injection vector $\mathbf{a} \in \mathbb{R}^M$ to curtail the ability of the network operator to estimate the state variables without being detected. Note that the impact of the data-injection vector $\mathbf{a}$ on the estimate $\hat{\mathbf{x}}_{\mathsf{MMSE}}$ is quantified by the second term on the right-hand side of the following equality:

$$\hat{\mathbf{x}}_{\mathsf{MMSE}} = \mathbf{M}(\mathbf{H}\mathbf{x} + \mathbf{z}) + \mathbf{M}\mathbf{a}. \tag{6}$$

The term $\mathbf{M}\mathbf{a}$ is referred to as the *excess distortion* induced by the attack vector $\mathbf{a}$ and is denoted by

$$\mathbf{x}_a \triangleq \mathbf{M}\mathbf{a} = \mathbf{\Sigma_{xx}}\mathbf{H}^\mathsf{T}(\mathbf{H}\mathbf{\Sigma_{xx}}\mathbf{H}^\mathsf{T} + \sigma^2\mathbf{I})^{-1}\mathbf{a}. \tag{7}$$

B. Attack Detection

As a part of the grid management, a network operator systematically attempts to identify the measurements that have been corrupted. This operation can be cast as a hypothesis testing problem with hypotheses

$$\mathcal{H}_0: \quad \text{There is no attack} \tag{8}$$

$$\mathcal{H}_1: \quad \text{Measurements are compromised.} \tag{9}$$

Assuming the operator knows that $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma_{xx}})$, it can obtain the joint density function of the measurements, $\mathbf{y}$, and the state variables $\mathbf{x}$. From (2) and the assumptions of the problem, it follows that the observations $\mathbf{y}$ are realizations of an M-dimensional real Gaussian random variable with covariance matrix

$$\mathbf{\Sigma_{yy}} = \mathbf{H}\mathbf{\Sigma_{xx}}\mathbf{H}^\mathsf{T} + \sigma^2\mathbf{I}, \tag{10}$$

and mean $\mathbf{a}$ when there is an attack; or zero mean when there is no attack. Within this setting, the hypothesis testing problem

described before is adapted to the attack detection problem by comparing the following hypotheses:

$$\mathcal{H}_0: \quad \mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma_{yy}}) \tag{11}$$

$$\mathcal{H}_1: \quad \mathbf{y} \sim \mathcal{N}(\mathbf{a}, \mathbf{\Sigma_{yy}}). \tag{12}$$

A worst case scenario approach is assumed for the attackers, namely, the operator knows the attack vector, $\mathbf{a}$, used in the attack. However, the operator does not know a priori whether the grid is under attack or not, which accounts for the need of an attack detection strategy. That being the case, the optimal detection strategy for the operator is to perform a likelihood ratio test $L(\mathbf{y}, \mathbf{a})$ with respect to the observations $\mathbf{y}$. Under the assumption that state variables follow a multivariate Gaussian distribution, the likelihood ratio can be calculated as

$$L(\mathbf{y}, \mathbf{a}) = \frac{f_{\mathcal{N}(\mathbf{0}, \mathbf{\Sigma_{yy}})}(\mathbf{y})}{f_{\mathcal{N}(\mathbf{a}, \mathbf{\Sigma_{yy}})}(\mathbf{y})} = \exp\left(\frac{1}{2}\mathbf{a}^{\mathsf{T}}\mathbf{\Sigma_{yy}^{-1}}\mathbf{a} - \mathbf{a}^{\mathsf{T}}\mathbf{\Sigma_{yy}^{-1}}\mathbf{y}\right), \tag{13}$$

where $f_{\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})}$ is the probability density function of a multivariate Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$. Therefore, either hypothesis is accepted by evaluating the inequalities

$$L(\mathbf{y}, \mathbf{a}) \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\gtrless}} \tau, \tag{14}$$

where $\tau \in [0, \infty)$ is tuned to set the trade-off between the probability of detection and the probability of false alarm.

III. Centralized Attacks

This section describes the construction of data-injection attacks in the special case in which there exists a unique attacker. This scenario is referred to as *centralized attacks* in order to highlight that there exists a unique entity deciding the data-injection vector $\mathbf{a} \in \mathbb{R}^M$ in (2). The difference between the scenario in which there exists a unique attacker or several (competing or cooperating) attackers is subtle and it is treated in Section IV.

Let $\mathcal{M} = \{1, \ldots, M\}$ denote the set of all M sensors available to the network operator. A sensor is said to be compromised if an attacker is able to arbitrarily modify its output. Given a total energy budget $E > 0$ at the attacker, the set of all possible attacks that can be injected to the network can be explicitly described:

$$\mathcal{A} = \left\{\mathbf{a} \in \mathbb{R}^M : \mathbf{a}^{\mathsf{T}}\mathbf{a} \leqslant E\right\}. \tag{15}$$

A. Attacks with Minimum Detection Probability

An attacker chooses a vector $\mathbf{a} \in \mathcal{A}$ taking into account the trade-off between the probability of being detected and the distortion (7) that it induces into the measurements. However, the choice of a particular data-injection vector is a task that is far from trivial as an attacker does not possess any information about the exact realization of the vector of state variables $\mathbf{x}$ and the noise $\mathbf{z}$. A reasonable assumption on the knowledge of the attacker is to consider that it knows the topology of the network and thus, it knows the matrix $\mathbf{H}$. It is also reasonable to consider that it knows the first and second moments of the state variables $\mathbf{x}$ and noise $\mathbf{z}$.

Under these knowledge assumptions, the average probability that the network operator is unable to detect the attack vector $\mathbf{a}$ is

$$\mathsf{P_{ND}}(\mathbf{a}) = \mathsf{E}\left[\mathbb{1}_{\{\mathsf{L}(\mathbf{y}, \mathbf{a}) > \tau\}}\right], \tag{16}$$

where the expectation is taken over the joint probability distribution of state variables $\mathbf{x}$ and the noise $\mathbf{z}$, and $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function. Note that under these assumptions, $\mathbf{y}$ is a random variable with Gaussian distribution with mean $\mathbf{a}$ and covariance matrix $\mathbf{\Sigma_{yy}}$. Thus, the probability $\mathsf{P_{ND}}(\mathbf{a})$ of a vector $\mathbf{a}$ being a successful attack, i.e., a non-detected attack is given by [20]

$$\mathsf{P_{ND}}(\mathbf{a}) = \frac{1}{2}\mathsf{erfc}\left(\frac{\frac{1}{2}\mathbf{a}^{\mathsf{T}}\mathbf{\Sigma_{yy}^{-1}}\mathbf{a} + \log\tau}{\sqrt{2\mathbf{a}^{\mathsf{T}}\mathbf{\Sigma_{yy}^{-1}}\mathbf{a}}}\right). \tag{17}$$

Often, the knowledge of the threshold τ in (14) is not available to the attacker and thus, it cannot determine the exact average probability of not being detected of a given attack vector $\mathbf{a}$. However, the knowledge of whether $\tau > 1$ or $\tau \leqslant 1$ induces different behaviors on the attacker. The following propositions follow immediately from (17) and the properties of the complementary error function.

Proposition 1 (Case $\tau \leqslant 1$) *Let $\tau \leqslant 1$. Then, for all $\mathbf{a} \in \mathcal{A}$, $\mathsf{P_{ND}}(\mathbf{a}) < \mathsf{P_{ND}}((0, \ldots, 0))$ and the probability $\mathsf{P_{ND}}(\mathbf{a})$ is monotonically decreasing with $\mathbf{a}^{\mathsf{T}}\mathbf{\Sigma_{yy}^{-1}}\mathbf{a}$.*

Proposition 2 (Case $\tau > 1$) *Let $\tau > 1$ and let also $\mathbf{\Sigma_{yy}} = \mathbf{U_{yy}}\mathbf{\Lambda_{yy}}\mathbf{U_{yy}^{\mathsf{T}}}$ be an SVD decomposition of $\mathbf{\Sigma_{yy}}$, with $\mathbf{U_{yy}^{\mathsf{T}}} = (\mathbf{u}_{yy,1}, \ldots, \mathbf{u}_{yy,M})$ and $\mathbf{\Lambda_{yy}} = diag\left(\lambda_{yy,1}, \ldots, \lambda_{yy,M}\right)$ and $\lambda_{yy,1} \geqslant \lambda_{yy,2} \geqslant \ldots, \geqslant \lambda_{yy,M}$. Then, any vector of the form*

$$\mathbf{a} = \pm\sqrt{\lambda_{yy,k} 2\log\tau}\,\mathbf{u}_{yy,k}, \tag{18}$$

with $k \in \{1, \ldots, M\}$, is a data-injection attack that satisfies for all $\mathbf{a}' \in \mathbb{R}^M$, $\mathsf{P_{ND}}(\mathbf{a}') \leqslant \mathsf{P_{ND}}(\mathbf{a})$.

The proof of Proposition 1 and Proposition 2 is as follows.

Proof: Let $x = \mathbf{a}^{\mathsf{T}}\mathbf{\Sigma_{yy}^{-1}}\mathbf{a}$ and note that $x > 0$ due to the positive definiteness of $\mathbf{\Sigma_{yy}}$. Let also the function $g : \mathbb{R} \to \mathbb{R}$ be

$$g(x) = \frac{\frac{1}{2}x + \log\tau}{\sqrt{2x}}. \tag{19}$$

The first derivative of $g(x)$ is

$$g'(x) = \frac{1}{2\sqrt{2x}}\left(\frac{1}{2} - \frac{\log\tau}{x}\right). \tag{20}$$

Note that in the case in which $\log\tau \leqslant 0$ (or $\tau \leqslant 1$), then $\forall x \in \mathbb{R}^+$, $g'(x) > 0$ and thus, g is monotonically increasing with x. Since the complementary error function erfc is monotonically decreasing with its argument, the statement of Proposition 1 follows and completes its proof. In the case in which $\log\tau \geqslant 0$ (or $\tau > 1$), the solution to $g'(x) = 0$ is $x = 2\log\tau$ and it corresponds to a minimum of the function g. The maximum of $\frac{1}{2}\mathrm{erfc}(g(x))$ occurs at the minimum of $g(x)$ given that erfc is monotonically decreasing with its argument. Hence, the maximum of $\mathsf{P_{ND}}(\mathbf{a})$ occurs at any $\mathbf{a}$ satisfying the condition:

$$\mathbf{a}^{\mathsf{T}}\mathbf{\Sigma_{yy}^{-1}}\mathbf{a} = 2\log\tau. \tag{21}$$

Solving for $\mathbf{a}$ in (21) yields (18) and this completes the proof of Proposition 2. ∎

The relevance of Proposition 1 is that it states that when $\tau \leqslant 1$, any non-zero data-injection attack vector possesses a non zero probability of being detected. Indeed, the highest probability $\mathsf{P}_{\mathsf{ND}}(\mathbf{a})$ of not being detected is guaranteed by the null vector $\mathbf{a} = (0, \ldots, 0)$, i.e., there is no attack. Alternatively, when $\tau > 1$ it follows from Proposition 2 that there always exists a non-zero vector that possesses maximum probability of not being detected. However, in both cases, it is clear that the corresponding data-injection vectors which induce the highest probability of not being detected are not necessarily the same that inflige the largest damage to the network, i.e., maximize the excess distortion.

From this point of view, an attacker faces the trade-off between maximizing the excess distortion and minimizing the probability of being detected. Thus, the attack construction can be formulated as an optimization problem in which the solution $\mathbf{a}$ is a data-injection vector that maximizes the probability $\mathsf{P}_{\mathsf{ND}}(\mathbf{a})$ of not being detected at the same time that it induces a given distortion $\|\mathbf{x}_a\|_2^2 \geqslant \mathsf{D}_0$ into the estimate. In the case in which $\tau \leqslant 1$, it follows from Proposition 1 and (7) that this problem can be formulated as the following optimization problem:

$$\min_{\mathbf{a}\in\mathcal{A}} \mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a} \quad \text{s.t.} \quad \mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{H}\boldsymbol{\Sigma}_{\mathbf{xx}}^{2}\mathbf{H}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a} \geq \mathsf{D}_0. \quad (22)$$

The solution to the optimization problem in (22) is given by the following theorem.

Theorem 1 *Let* $\mathbf{G} = \boldsymbol{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}}\mathbf{H}\boldsymbol{\Sigma}_{\mathbf{xx}}^{2}\mathbf{H}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}}$ *have a singular value decomposition* $\mathbf{G} = \mathbf{U}_{\mathbf{G}}\boldsymbol{\Sigma}_{\mathbf{G}}\mathbf{U}_{\mathbf{G}}^{\mathsf{T}}$, *with* $\mathbf{U} = (\mathbf{u}_{\mathbf{G},i}, \ldots, \mathbf{u}_{\mathbf{G},M})$ *a unitary matrix and* $\boldsymbol{\Sigma}_{\mathbf{G}} = diag\,(\lambda_{\mathbf{G},1}, \ldots, \lambda_{\mathbf{G},M})$ *a diagonal matrix with* $\lambda_{\mathbf{G},1} \geqslant \ldots \geqslant \lambda_{\mathbf{G},M}$. *Then, if* $\tau \leqslant 1$, *the attack vector* $\mathbf{a}$ *that maximizes the probability of not being detected* $\mathsf{P}_{\mathsf{ND}}(\mathbf{a})$ *while inducing an excess distortion not less than* D_0 *is*

$$\mathbf{a} = \pm\sqrt{\frac{\mathsf{D}_0}{\lambda_{\mathbf{G},1}}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{\frac{1}{2}}\mathbf{u}_{\mathbf{G},1}. \quad (23)$$

Moreover, $\mathsf{P}_{\mathsf{ND}}(\mathbf{a}) = \frac{1}{2}\mathrm{erfc}\left(\frac{\frac{\mathsf{D}_0}{2\lambda_{\mathbf{G},1}} + \log\tau}{\sqrt{\frac{2\mathsf{D}_0}{\lambda_{\mathbf{G},1}}}}\right)$.

Proof: Consider the Lagrangian

$$L(\mathbf{a}) = \mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a} - \gamma\left(\mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{H}\boldsymbol{\Sigma}_{\mathbf{xx}}^{2}\mathbf{H}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a} - \mathsf{D}_0\right), \quad (24)$$

with $\gamma > 0$ a Lagrangian multiplier. Then, the necessary conditions for $\mathbf{a}$ to be a solution to the optimization problem (22) are:

$$\nabla_{\mathbf{a}} L(\mathbf{a}) = 2\left(\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} - \gamma\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{H}\boldsymbol{\Sigma}_{\mathbf{xx}}^{2}\mathbf{H}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\right)\mathbf{a} = 0 \quad (25)$$

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} L(\mathbf{a}) = \mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{H}\boldsymbol{\Sigma}_{\mathbf{xx}}^{2}\mathbf{H}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a} - \mathsf{D}_0 = 0. \quad (26)$$

Note that any

$$\mathbf{a}_i = \pm\sqrt{\frac{\mathsf{D}_0}{\lambda_{\mathbf{G},i}}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{\frac{1}{2}}\mathbf{u}_{\mathbf{G},i} \text{ and} \quad (27)$$

$$\gamma_i = \lambda_{\mathbf{G},i}, \text{ with } 1 \leqslant i \leqslant \mathrm{rank}\,(\mathbf{G}), \quad (28)$$

satisfy $\gamma_i > 0$ and conditions (25) and (26). Hence, the set of vectors that satisfy the necessary conditions to be a solution of (22) is

$$\left\{\mathbf{a}_i = \pm\sqrt{\frac{\mathsf{D}_0}{\lambda_{\mathbf{G},i}}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{\frac{1}{2}}\mathbf{u}_{\mathbf{G},i} : 1 \leqslant i \leqslant \mathrm{rank}\,(\mathbf{G})\right\}. \quad (29)$$

More importantly, any vector $\mathbf{a} \neq \mathbf{a}_i$, with $1 \leqslant i \leqslant \mathrm{rank}\,(\mathbf{G})$, does not satisfy the necessary conditions. Moreover,

$$\mathbf{a}_i^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}_i = \frac{\mathsf{D}_0}{\lambda_{\mathbf{G},i}} \geqslant \frac{\mathsf{D}_0}{\lambda_{\mathbf{G},1}}. \quad (30)$$

Therefore, $\mathbf{a} = \pm\sqrt{\frac{\mathsf{D}_0}{\lambda_{\mathbf{G},1}}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{\frac{1}{2}}\mathbf{u}_{\mathbf{G},1}$ are the unique solutions to (22). This completes the proof. ∎

Note that the construction of the data-injection attack $\mathbf{a}$ in (23) does not require the exact knowledge of τ. That is, only knowing that $\tau \leqslant 1$ is enough to build the data-injection attack that has the highest probability of not being detected and induces a distortion of at least D_0.

In the case in which $\tau > 1$, it is also possible to find the data-injection attack vector that induces a distortion not less than D_0 and the maximum probability of not being detected. Such a vector is the solution to the following optimization problem.

$$\min_{\mathbf{a}\in\mathcal{A}} \frac{\frac{1}{2}\mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a} + \log\tau}{\sqrt{2\mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}}} \text{ s.t. } \mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{H}\boldsymbol{\Sigma}_{\mathbf{xx}}^{2}\mathbf{H}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a} \geq \mathsf{D}_0. \quad (31)$$

The solution to the optimization problem in (31) is given by the following theorem.

Theorem 2 *Let* $\mathbf{G} = \boldsymbol{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}}\mathbf{H}\boldsymbol{\Sigma}_{\mathbf{xx}}^{2}\mathbf{H}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}}$ *have a singular value decomposition* $\mathbf{G} = \mathbf{U}_{\mathbf{G}}\boldsymbol{\Sigma}_{\mathbf{G}}\mathbf{U}_{\mathbf{G}}^{\mathsf{T}}$, *with* $\mathbf{U}_{\mathbf{G}} = (\mathbf{u}_{\mathbf{G},i}, \ldots, \mathbf{u}_{\mathbf{G},M})$ *a unitary matrix and* $\boldsymbol{\Sigma}_{\mathbf{G}} = diag\,(\lambda_{\mathbf{G},1}, \ldots, \lambda_{\mathbf{G},M})$ *a diagonal matrix with* $\lambda_{\mathbf{G},1} \geqslant \ldots \geqslant \lambda_{\mathbf{G},M}$. *Then, when* $\tau > 1$, *the attack vector* $\mathbf{a}$ *that maximizes the probability of not being detected* $\mathsf{P}_{\mathsf{ND}}(\mathbf{a})$ *while producing an excess distortion not less than* D_0 *is*

$$\mathbf{a} = \begin{cases} \pm\sqrt{\frac{\mathsf{D}_0}{\lambda_{\mathbf{G},k^*}}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{\frac{1}{2}}\mathbf{u}_{\mathbf{G},k^*} & \text{if } \frac{\mathsf{D}_0}{2\log\tau\lambda_{\mathbf{G},\mathrm{rank}\,\mathbf{G}}} \geqslant 1, \\ \pm\sqrt{2\log\tau}\boldsymbol{\Sigma}_{\mathbf{yy}}^{\frac{1}{2}}\mathbf{u}_{\mathbf{G},1} & \text{if } \frac{\mathsf{D}_0}{2\log\tau\lambda_{\mathbf{G},\mathrm{rank}\,\mathbf{G}}} < 1 \end{cases}$$

with

$$k^* = \arg \min_{k\in\{1,\ldots,\mathrm{rank}\mathbf{G}\}:\frac{\mathsf{D}_0}{\lambda_{\mathbf{G},k}} > 2\log(\tau)} \frac{\mathsf{D}_0}{\lambda_{\mathbf{G},k}}. \quad (32)$$

Proof: The structure of the proof of Theorem 2 is similar to the proof of Theorem 1 and is omitted in this paper. A complete proof can be found in [21]. ∎

B. Attacks with Maximum Distortion

In the previous subsection, the attacker constructs its data-injection vector $\mathbf{a}$ aiming to maximize the probability of non-detection $\mathsf{P}_{\mathsf{ND}}(\mathbf{a})$ while guaranteeing a minimum distortion. However, this problem has a dual in which the objective is to maximize the distortion $\mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{H}\boldsymbol{\Sigma}_{\mathbf{xx}}^{2}\mathbf{H}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}$ while guaranteeing that the probability of not being detected remains

always larger than a given threshold $L_0' \in [0, \frac{1}{2}]$. This problem can be formulated as the following optimization problem:

$$\max_{\mathbf{a}\in\mathcal{A}} \mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{H}\boldsymbol{\Sigma}_{\mathbf{xx}}^{2}\mathbf{H}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a} \text{ s.t. } \frac{\frac{1}{2}\mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}+\log\tau}{\sqrt{2\mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}}} \leq L_0, \tag{33}$$

with $L_0 = \mathrm{erfc}^{-1}(2L_0') \in [0,\infty)$.

The solution to the optimization problem in (33) is given by the following theorem.

Theorem 3 *Let the matrix* $\mathbf{G} = \boldsymbol{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}}\mathbf{H}\boldsymbol{\Sigma}_{\mathbf{xx}}^{2}\mathbf{H}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}}$ *have a singular value decomposition* $\mathbf{U}_{\mathbf{G}}\boldsymbol{\Sigma}_{\mathbf{G}}\mathbf{U}_{\mathbf{G}}^{\mathsf{T}}$*, with* $\mathbf{U} = (\mathbf{u}_{\mathbf{G},i},\ldots,\mathbf{u}_{\mathbf{G},M})$ *a unitary matrix and* $\boldsymbol{\Sigma}_{\mathbf{G}} = diag(\lambda_{\mathbf{G},1},\ldots,\lambda_{\mathbf{G},M})$ *a diagonal matrix with* $\lambda_{\mathbf{G},1} \geqslant \ldots \geqslant \lambda_{\mathbf{G},M}$*. Then, the attack vector* $\mathbf{a}$ *that maximizes the excess distortion* $\mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}}\mathbf{G}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}}\mathbf{a}$ *with a probability of not being detected that does not go below* $L_0 \in [0,\frac{1}{2}]$ *is*

$$\mathbf{a} = \pm\left(\sqrt{2}L_0 + \sqrt{2L_0^2 - 2\log\tau}\right)\boldsymbol{\Sigma}_{\mathbf{yy}}^{\frac{1}{2}}\mathbf{u}_{\mathbf{G},1}, \tag{34}$$

when a solution to (33) *exists.*

Proof: The structure of the proof of Theorem 3 is similar to the proof of Theorem 1 and is omitted in this paper. A complete proof can be found in [21]. ∎

IV. Decentralized Attacks

Let $\mathcal{K} = \{1,\ldots,K\}$ be the set of attackers that can potentially perform a data injection attack on the network. Let also $\mathcal{C}_i$ be the set of sensors that attacker i can control. Assume that $\mathcal{C}_1,\ldots,\mathcal{C}_K$ are proper sets and form a partition of the set $\mathcal{M}$ of all sensors. The set $\mathcal{A}_k$ of data attack vectors $\mathbf{a}_k$ that can be injected into the network by attacker $k \in \mathcal{K}$ is of the form

$$\mathcal{A}_k = \{\mathbf{a}_k \in \mathbb{R}^M : (\mathbf{a}_k)_j = 0 \text{ for all } j \notin \mathcal{C}_k, \mathbf{a}_k^T\mathbf{a}_k \leq E_k\}. \tag{35}$$

The constant $E_k < \infty$ represents the energy budget of attacker k. Let the set of all possible sums of the elements of $\mathcal{A}_i$ and $\mathcal{A}_j$ be denoted by $\mathcal{A}_i \oplus \mathcal{A}_j$. That is, for all $\mathbf{a} \in \mathcal{A}_i \oplus \mathcal{A}_j$, there exists a pair of vectors $(\mathbf{a}_i,\mathbf{a}_j) \in \mathcal{A}_i \times \mathcal{A}_j$ such that $\mathbf{a} = \mathbf{a}_i + \mathbf{a}_j$. Using this notation, let the set of all possible data-injection attacks be denoted by

$$\mathcal{A} = \mathcal{A}_1 \oplus \mathcal{A}_2 \oplus \ldots \oplus \mathcal{A}_K, \tag{36}$$

and the set of complementary data-injection attacks with respect to attacker k be denoted by

$$\mathcal{A}_{-k} = \mathcal{A}_1 \oplus \ldots \oplus \mathcal{A}_{k-1} \oplus \mathcal{A}_{k+1} \oplus \ldots \oplus \mathcal{A}_K. \tag{37}$$

Given the individual data injection vectors $\mathbf{a}_i \in \mathcal{A}_i$, with $i \in \{1,\ldots,K\}$, the global attack vector $\mathbf{a}$ is

$$\mathbf{a} = \sum_{i=1}^{K}\mathbf{a}_k \in \mathcal{A}. \tag{38}$$

The aim of attacker k is to corrupt the measurements obtained by the set of meters $\mathcal{C}_k$ by injecting an error vector $\mathbf{a}_k \in \mathcal{A}_k$ that maximizes the damage to the network, e.g., the excess distortion, while avoiding the detection of the global data-injection vector $\mathbf{a}$. Clearly, all attackers have the same interest but they control different sets of measurements, i.e., $\mathcal{C}_i \neq \mathcal{C}_k$, for a any pair $(i,k) \in \mathcal{K}^2$. For modeling this behavior, attackers use the utility function $\phi : \mathbb{R}^M \to \mathbb{R}$, to determine whether a data-injection vector $\mathbf{a}_k \in \mathcal{A}_k$ is more beneficial than another $\mathbf{a}_k' \in \mathcal{A}_k$ given the complementary attack vector

$$\mathbf{a}_{-k} = \sum_{i\in\{1,\ldots,K\}\setminus\{k\}} \mathbf{a}_i \in \mathcal{A}_{-k} \tag{39}$$

adopted by all the other attackers. The function ϕ is chosen considering the fact that an attack is said to be successful if it induces a non-zero distortion and it is not detected. Alternatively, if the attack is detected no damage is induced into the network as the operator discards the measurements and no estimation is performed. Hence, given a global attack $\mathbf{a}$, the distortion induced into the measurements is $\mathbb{1}_{\{L(\mathbf{Hx}+\mathbf{z}+\mathbf{a},\mathbf{a})>\tau\}}\mathbf{x}_a^{\mathsf{T}}\mathbf{x}_a$. However, attackers are not able to know the exact state of the network $\mathbf{x}$ and the realization of the noise $\mathbf{z}$ before launching the attack. Thus, it appears natural to exploit the knowledge of the first and second moments of both the state variables $\mathbf{x}$ and noise $\mathbf{z}$ and consider as a metric the expected distortion $\phi(\mathbf{a})$ that can be induced by the attack vector $\mathbf{a}$:

$$\phi(\mathbf{a}) = \mathsf{E}\left[\left(\mathbb{1}_{\{\mathsf{L}(\mathbf{Hx}+\mathbf{z}+\mathbf{a},\mathbf{a})>\tau\}}\right)\mathbf{x}_{\mathsf{a}}^{\mathsf{T}}\mathbf{x}_{\mathsf{a}}\right], \tag{40}$$

$$= \mathsf{P}_{\mathsf{ND}}(\mathbf{a})\ \mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{H}\boldsymbol{\Sigma}_{\mathbf{xx}}^{2}\mathbf{H}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}, \tag{41}$$

where the expectation is taken over the distribution of state variables $\mathbf{x}$ and the noise $\mathbf{z}$. Note that under this assumptions of global knowledge, this model considers the worst case scenario for the network operator. Indeed, the result presented in this section corresponds to the case in which the attackers inflict the most harm.

A. Game Formulation

The benefit $\phi(\mathbf{a})$ obtained by attacker k does not only depend on its own data-injection vector $\mathbf{a}_k$, but also on the data-injection vector $\mathbf{a}_{-k}$ induced by the other attackers. This becomes clear from the construction of the global data-injection vector $\mathbf{a}$ in (38), the excess distortion $\mathbf{x}_a$ in (7) and the probability of not being detected $\mathsf{P}_{\mathsf{ND}}(\mathbf{a})$ in (17). Therefore, the interaction of all attackers in the network can be described by a game in normal form

$$\mathcal{G} = \left(\mathcal{K}, \{\mathcal{A}_k\}_{k\in\mathcal{K}}, \phi\right). \tag{42}$$

Each attacker is a player in the game $\mathcal{G}$ and it is identified by an index from the set $\mathcal{K}$. The actions player k might adopt are data-injection vectors $\mathbf{a}_k$ in the set $\mathcal{A}_k$ in (35). The underlying assumption in the following of this section is that, given a vector of data-injection attacks $\mathbf{a}_{-k}$, player k aims to adopt a data-injection vector $\mathbf{a}_k$ such that the expected excess distortion $\phi(\mathbf{a}_k + \mathbf{a}_{-k})$ is maximized. That is,

$$\mathbf{a}_k \in \mathrm{BR}_k(\mathbf{a}_{-k}), \tag{43}$$

where the correspondence $\mathrm{BR}_k : \mathcal{A}_{-k} \to 2^{\mathcal{A}_k}$ is the best response correspondence, i.e.,

$$\mathrm{BR}_k(\mathbf{a}_{-k}) = \arg\max_{\mathbf{a}_k\in\mathcal{A}_k} \phi(\mathbf{a}_k + \mathbf{a}_{-k}). \tag{44}$$

The notation $2^{\mathcal{A}_k}$ represents the set of all possible subsets of $\mathcal{A}_k$. Note that $\mathrm{BR}_k(\mathbf{a}_{-k}) \subseteq \mathcal{A}_k$ is the set of data-injection attack vectors that are optimal given that the other attackers have adopted the data-injection vector $\mathbf{a}_{-k}$. In this setting, each attacker tampers with a subset $\mathcal{C}_k$ of all sensors $\mathcal{C}$, as opposed to the centralized case in which there exists a single attacker able to tampers with all sensors in $\mathcal{C}$.

A game solution that is particularly relevant for this analysis is the NE [22].

Definition 1 (Nash Equilibrium) *The data-injection vector* $\mathbf{a}$ *is an NE of the game* $\mathcal{G}$ *if and only if it is a solution of the fix point equation*

$$\mathbf{a} = \mathrm{BR}(\mathbf{a}), \tag{45}$$

with $\mathrm{BR} : \mathcal{A} \to 2^{\mathcal{A}}$ *being the global best-response correspondence, i.e.,*

$$\mathrm{BR}(\mathbf{a}) = \mathrm{BR}_1(\mathbf{a}_{-1}) \oplus \ldots \oplus \mathrm{BR}_K(\mathbf{a}_{-K}). \tag{46}$$

Essentially, at an NE, attackers obtain the maximum benefit given the data-injection vector adopted by all the other attackers. This implies that an NE is an operating point at which attackers achieve the highest expected distortion induced over the measurements. More importantly, any unilateral deviation from an equilibrium data-injection vector $\mathbf{a}$ does not lead to an improvement of the average excess distortion. Note that this formulation does not say anything about the exact distortion induced by an attack but the average distortion. This is mainly because the attack is chosen under the uncertainty of the state vector $\mathbf{x}$ and the noise term $\mathbf{z}$.

The following proposition highlights an important property of the game $\mathcal{G}$ in (42).

Proposition 3 *The game* $\mathcal{G}$ *in* (42) *is a potential game.*

Proof: The proof follows immediately from the observation that all the players have the same utility function ϕ [23]. Thus, the function ϕ is a potential of the game $\mathcal{G}$ in (42) and any maximum of the potential function is an NE of the game $\mathcal{G}$. ∎

In general, potential games [23] possess numerous properties that are inherited by the game $\mathcal{G}$ in (42). These properties are detailed by the following propositions

Proposition 4 *The game* $\mathcal{G}$ *possesses at least one NE.*

Proof: Note that ϕ is continuous in $\mathcal{A}$ and $\mathcal{A}$ is a convex and closed set; therefore, there always exists a maximum of the potential function ϕ in $\mathcal{A}$. Finally from Lemma 4.3 in [23], it follows that such a maximum corresponds to an NE. ∎

B. Achievability of an NE

The attackers are said to play a sequential best response dynamic (BRD) if the attackers can sequentially decide their own data-injection vector $\mathbf{a}_k$ from their sets of best responses following a round-robin (increasing) order. Denote by $\mathbf{a}_k^{(t)} \in \mathcal{A}_k$ the choice of attacker k during round $t \in \mathbb{N}$ and assume that attackers are able to observe all the other attackers' data-injection vectors. Under these assumptions, the BRD can be defined as follows.

Definition 2 (Best Response Dynamics) *The players of the game* $\mathcal{G}$ *are said to play best response dynamics if there exists a round-robin order of the elements of* $\mathcal{K}$ *in which at each round* $t \in \mathbb{N}$*, the following holds:*

$$\mathbf{a}_k^{(t)} \in \mathrm{BR}_k\left(\mathbf{a}_1^{(t)} + \ldots + \mathbf{a}_{k-1}^{(t)} + \mathbf{a}_{k+1}^{(t-1)} + \ldots + \mathbf{a}_K^{(t-1)}\right). \tag{47}$$

From the properties of potential games (Lemma 4.2 in [23]), the following proposition follows.

Lemma 1 (Achievability of NE attacks) *Any BRD in the game* $\mathcal{G}$ *converges to a data-injection attack vector that is an NE.*

The relevance of Lemma 1 is that it establishes that if attackers can communicate in at least a round-robin fashion, they are always able to attack the network with a data-injection vector that maximizes the average excess distortion. Note that there might exists several NEs (local maxima of ϕ) and there is no guarantee that attackers will converge to the best NE, i.e., a global maximum of ϕ. It is important to note that under the assumption that there exists a unique maximum, which is not the case for the game $\mathcal{G}$ (see Theorem 4), all attackers are able to calculate such a global maximum and no communications is required among the attackers. Nonetheless, the game $\mathcal{G}$ always possesses at least two NEs, which enforces the use of a sequential BRD to converge to an NE.

C. Cardinality of the set of NEs

Let $\mathcal{A}_{\mathrm{NE}}$ be the set of all data-injection attacks that form NEs. The following theorem bounds the number of NEs in the game.

Theorem 4 *The cardinality of the set* $\mathcal{A}_{\mathrm{NE}}$ *of NE of the game* $\mathcal{G}$ *satisfies*

$$2 \leqslant |\mathcal{A}_{\mathrm{NE}}| \leqslant C \cdot \mathrm{rank}(\mathbf{H}) \tag{48}$$

where $C < \infty$ *is a constant that depends on* τ*.*

Proof: The lower bound follows from the symmetry of the utility function given in (40), i.e. $\phi(\mathbf{a}) = \phi(-\mathbf{a})$, and the existence of at least one NE claimed in Proposition 4.

To prove the upper bound the number of stationary points of the utility function is evaluated. This is equivalent to the cardinality of the set

$$\mathcal{S} = \{\mathbf{a} \in \mathbb{R}^M : \nabla_{\mathbf{a}}\phi(\mathbf{a}) = \mathbf{0}\}, \tag{49}$$

which satisfies $\mathcal{A}_{NE} \subseteq \mathcal{S}$. Calculating the gradient with respect to the attack vector yields

$$\nabla_{\mathbf{a}}\phi(\mathbf{a}) = \left(\alpha(\mathbf{a})\mathbf{M}^{\mathsf{T}}\mathbf{M} - \beta(\mathbf{a})\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\right)\mathbf{a}, \tag{50}$$

where

$$\alpha(\mathbf{a}) \triangleq \mathrm{erfc}\left(\frac{1}{\sqrt{2}}\frac{\frac{1}{2}\mathbf{a}^{\mathsf{T}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a} + \log\tau}{\left(\mathbf{a}^{\mathsf{T}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}\right)^{\frac{1}{2}}}\right) \tag{51}$$

and

$$\beta(\mathbf{a}) \triangleq \frac{\mathbf{a}^{\mathsf{T}}\mathbf{M}^{\mathsf{T}}\mathbf{M}\mathbf{a}}{\sqrt{2\pi}\mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}}\left(\frac{1}{2}-\frac{\log\tau}{\mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}}\right) \times \exp\left(-\left(\frac{1}{\sqrt{2}}\frac{\frac{1}{2}\mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}+\log\tau}{\left(\mathbf{a}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}\right)^{\frac{1}{2}}}\right)^{2}\right). \quad (52)$$

Define $\delta(\mathbf{a}) \triangleq \frac{\beta(\mathbf{a})}{\alpha(\mathbf{a})}$ and note that combining (4) with (50) gives the following condition for the stationary points:

$$\left(\mathbf{H}\boldsymbol{\Sigma}_{\mathbf{xx}}^{2}\mathbf{H}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}-\delta(\mathbf{a})\mathbf{I}\right)\mathbf{a}=\mathbf{0}. \quad (53)$$

Note that the number of linearly independent attack vectors that are a solution of the linear system in (53) is given by

$$R \triangleq \operatorname{rank}\left(\mathbf{H}\boldsymbol{\Sigma}_{\mathbf{xx}}^{2}\mathbf{H}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\right) \quad (54)$$

$$= \operatorname{rank}\left(\mathbf{H}\right). \quad (55)$$

where (55) follows from the fact that $\boldsymbol{\Sigma}_{\mathbf{xx}}$ and $\boldsymbol{\Sigma}_{\mathbf{yy}}$ are positive definite. Define the eigenvalue decomposition

$$\boldsymbol{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}}\mathbf{H}\boldsymbol{\Sigma}_{\mathbf{xx}}^{2}\mathbf{H}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}}=\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathsf{T}} \quad (56)$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix containing the ordered eigenvalues $\{\lambda_i\}_{i=1}^{M}$ matching the order of the eigenvectors in $\mathbf{U}$. As a result of (54) there are R eigenvalues, λ_k, which are different from zero and $M-R$ diagonal elements of $\boldsymbol{\Lambda}$ which are zero. Combining this decomposition with some algebraic manipulation, the condition for stationary points in (53) can be recast as

$$\boldsymbol{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}}\mathbf{U}\left(\boldsymbol{\Lambda}-\delta(\mathbf{a})\mathbf{I}\right)\mathbf{U}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}}\mathbf{a}=\mathbf{0}. \quad (57)$$

Let $w \in \mathbb{R}$ be a scaling parameter and observe that the attack vectors that satisfy $\mathbf{a} = w\boldsymbol{\Sigma}_{\mathbf{yy}}^{\frac{1}{2}}\mathbf{U}\mathbf{e}_k$ and $\delta(\mathbf{a}) = \lambda_k$ for $k = 1, \ldots, R$ are solutions of (57). Note that the critical points associated to zero eigenvalues are not NE. Indeed, the eigenvectors associated to zero eigenvalues yield zero utility. Since the utility function is strictly positive, these critical points are minima of the utility function and can be discarded when counting the number of NE. Therefore, the set in (49) can be rewritten based on the condition in (57) as

$$\mathcal{S} = \bigcup_{k=1}^{R}\mathcal{S}_k, \quad (58)$$

where

$$\mathcal{S}_k = \{\mathbf{a} \in \mathbb{R}^M : \mathbf{a} = w\boldsymbol{\Sigma}_{\mathbf{yy}}^{\frac{1}{2}}\mathbf{U}\mathbf{e}_k \text{ and } \delta(\mathbf{a}) = \lambda_k\}. \quad (59)$$

There are R linearly independent solutions of (57) but for each linearly independent solution there can be several scaling parameters, w, which satisfy $\delta(\mathbf{a}) = \lambda_k$. For that reason, $|\mathcal{S}_k|$ is determined by the number of scaling parameters that satisfy $\delta(\mathbf{a}) = \lambda_k$. To that end, define $\delta' : \mathbb{R} \to \mathbb{R}$ as $\delta'(w) \triangleq \delta(w\boldsymbol{\Sigma}_{\mathbf{yy}}^{\frac{1}{2}}\mathbf{U}\mathbf{e}_k)$. It is easy to check that $\delta'(w) = \lambda_k$ has a finite number of solutions for $k = 1, \ldots, R$. Hence, for all k there exists a constant C_k such that $|\mathcal{S}_k| \leq C_k$ which yields the upper bound

$$|\mathcal{S}| \leq \sum_{i=1}^{R}|\mathcal{S}_k| \leq \sum_{i=1}^{R}C_k \leq \max_k C_k R. \quad (60)$$

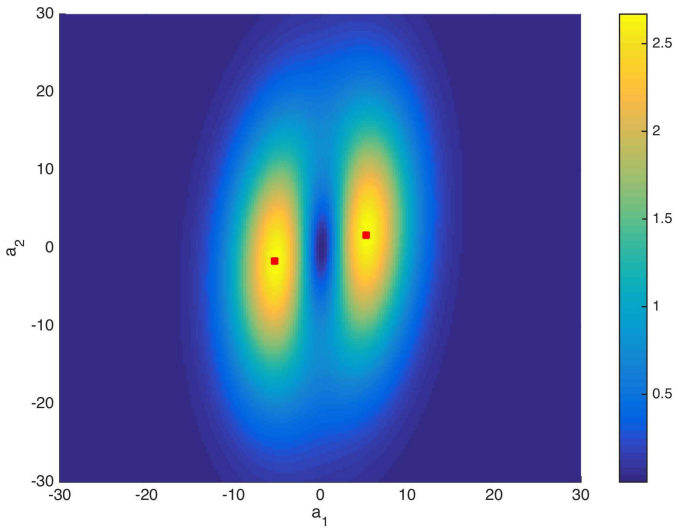

Fig. 1. Utility function for 30 bus IEEE test system as a function of the attack vector where attacker 1 controls real power injection measurement 1 and attacker 2 controls real power injection measurement 2. The red squares show the location of the NE points.

Noticing that the there is a finite number of solutions of $\delta'(w) = \lambda_k$ and that they depend only on τ yields the upper bound.

■

V. Numerical Results

In this section the properties of the game $\mathcal{G}$ described in Section IV are numerically evaluated for the 14 and 30 bus IEEE test systems. All numerical results are obtained for the case in which there are two attackers in the system where attacker one controls measurement sensor one, i.e. $\mathcal{C}_1 = \{1\}$, and attacker two controls measurement sensor two, i.e. $\mathcal{C}_2 = \{2\}$.

The results presented in this paper apply to any positive definite covariance matrix $\boldsymbol{\Sigma}_{\mathbf{xx}}$. However, for the sake of discussion and in order to illustrate the analytical results presented above, a particular covariance matrix model is chosen for the simulations. Since covariance matrices of weakly stationary random processes are Toeplitz [24], an exponentially decaying Toeplitz model is chosen where the strength of the correlation is set by the correlation strength parameter $\rho \in (0, 1]$, namely, $\boldsymbol{\Sigma}_{\mathbf{xx}} = \left[(\boldsymbol{\Sigma}_{\mathbf{xx}})_{i,j} = \rho^{|i-j|}; i, j = 1, 2, \ldots, n\right]$. Similarly, the standard deviation of the additive noise term, $\mathbf{z}$, is set to $\sigma = 0.1$ for all simulations which yields a signal to noise ratio of $10\log_{10}\left(\frac{1}{\sigma^2}\right) = 20$ dB.

Figure 1 depicts the utility function described by (40) when two attackers are present in the IEEE 30 bus test system and each controls one measurement sensor. The NEs have been numerically evaluated and are represented by red squares. In this example, the number of NE coincides with the lower bound in Theorem 4 and the attack vectors are antisymmetric as it is expected given the symmetry of the utility function.

The utility function evaluated in an NE as a function of the likelihood ratio threshold, τ, is shown in Figure 2 for different types of measurement sensors. For both IEEE test systems

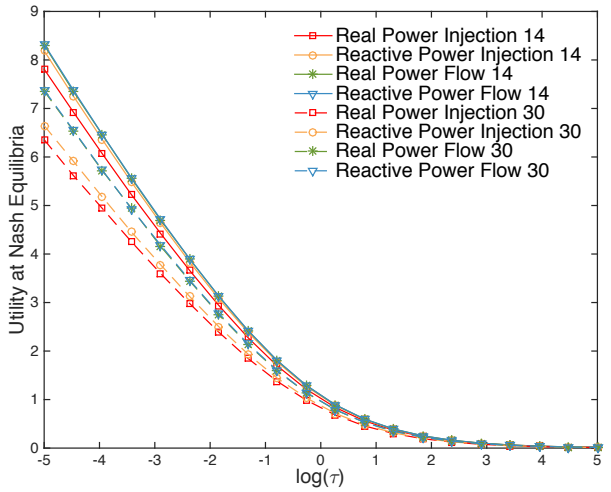

Fig. 2. Utility at NE as a function of $\log \tau$ for different sets of measurement sensors. The solid lines correspond to the 14 bus IEEE test system and the dashed lines to the 30 bus IEEE test system.

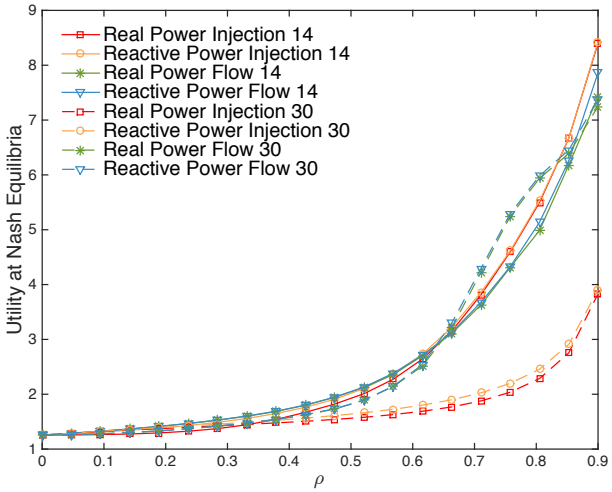

Fig. 3. Utility at NE as a function of ρ for different sets of measurement sensors. The solid lines correspond to the 14 bus IEEE test system and the dashed lines to the 30 bus IEEE test system.

considered, the 14 bus and 30 bus cases, power flow sensors consistently provide a higher utility to the attackers than the power injection counterparts. Interestingly, the difference with power injection measurements decreases when τ increases, i.e., the operator increases the probability of detection but also increases the probability of false alarm in the system. It is also worth noticing that the performance of the attackers is lower in the larger 30 bus system which suggests that large scale networks pose a more challenging scenario for decentralized attack strategies.

A main observation in this paper is that attackers can exploit the correlation between state variables to improve the performance of decentralized attack strategies. Figure 3 shows the utility function evaluated in an NE as a function of the correlation strength parameter ρ, governing the strength of the correlation between state variables. Remarkably, the utility in the NE increases monotonically as a function of the correlation strength parameter ρ, which suggests that increasing the dependency between state variables facilitates the coordination of decentralized attack strategies. That being said, it is assumed that attackers know the underlying statistical structure of the state variables, i.e. $\mathbf{\Sigma_{xx}}$, which demands a significant learning effort from the attackers.

VI. Conclusion

In this paper, we have considered the design of data injection vectors in state estimation for electricity grids. In particular, we have studied the case in which the operator acquires the state of the grid through MMSE estimation and the attack detection is based on a likelihood ratio test. Within this setting, the trade-off between achievable distortion and probability of detection has been characterised by deriving optimal centralized attack constructions for a given distortion and probability of detection pair. It is worth noting that the optimal attack strategy considers the statistical structure of the state variables and that correlation can be exploited by the attacker to construct more efficient attacks.

We have then extended the investigation to decentralized scenarios in which several attackers construct their respective attack without coordination. In this setting, we have posed the interaction between the attackers in a game theoretic setting. Central to this study is the derivation of a new utility function that captures the most important aspects of decentralized attack construction in electricity grids. We have shown that the proposed utility function results in a setting that can be described as a potential game which allows us to claim the existence of a NE and the convergence of BRD to a NE. We have then provided bounds on the number of NE and prove that there is always a finite number of NE and that there are always at least two NE. Interestingly, this implies that attackers cannot guarantee a strategy that will lead to an NE without coordination. In the numerical results section we evaluate the analytical results in IEEE Test systems with 14 and 30 buses. The numerical results corroborate that there is no single NE and that the statistical structure of the state variables can be exploited by the attackers to maximize the distortion that they induce in the state estimate of the network operator.

References

[1] E. Hossain, Z. Han, and H. V. Poor, *Smart Grid Communications and Networking*, Cambridge University Press, 2012.

[2] A. Abur and A. Gomez Exposito, *Power System State Estimation: Theory and Implementation*, CRC Press, 2004.

[3] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," in *Proc. ACM Conference on Computer and Communications Security*, Chicago, IL, USA, Nov. 2009, pp. 21–32.

[4] S. Cui, Z. Han, S. Kar, T. T. Kim, H. V. Poor, and A. Tajer, "Coordinated data-injection attack and detection in the smart grid: A detailed look at enriching detection solutions," *IEEE Signal Process. Mag.*, vol. 29, no. 5, pp. 106–115, Sep. 2012.

[5] L. Sankar, S. Kar, R. Tandon, and H. V. Poor, "Competitive privacy in the smart grid: An information-theoretic approach," in *Proc. IEEE International Conference on Smart Grid Communications*, Brussels, Belgium, 2011, pp. 220–225.

[6] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, "Malicious data attacks on the smart grid," *IEEE Trans. Smart Grid*, vol. 2, no. 4, pp. 645–658, Oct. 2011.

[7] Y. Huang, H. Li, K. Campbell, and Z. Han, "Defending false data injection attack on smart grid network using adaptive CUSUM test," in *Proc. Annual Conference on Information Sciences and Systems*, Princeton, NJ, USA, 2011, pp. 1–6.
[8] E. Caro, A.J. Conejo, R. Minguez, M. Zima, and G. Andersson, "Multiple bad data identification considering measurement dependencies," *IEEE Trans. Power Syst.*, vol. 26, no. 4, pp. 1953–1961, Nov. 2011.
[9] T. Liu, Y. Gu, D. Wang, Y. Gui, and X. Guan, "A novel method to detect bad data injection attack in smart grid," in *Proc. IEEE Conference on Computer Communications Workshops*, Turin, Italy, Apr. 2013, pp. 49–54.
[10] A. Tajer, "Energy grid state estimation under random and structured bad data," in *Proc. IEEE Sensor Array and Multichannel Signal Processing Workshop*, Coruna, Spain, Jun. 2014, pp. 65–68.
[11] A. Teixeira, S. Amin, H. Sandberg, K. H. Johansson, and S.S. Sastry, "Cyber security analysis of state estimators in electric power systems," in *Proc. IEEE Conference on Decision and Control*, Atlanta, GA, USA, Dec. 2010, pp. 5991–5998.
[12] X. Liu, Z. Bao, D. Lu, and Z. Li, "Modeling of local false data injection attacks with reduced network information," *IEEE Trans. Smart Grid*, vol. 6, no. 4, pp. 1686–1696, Jul. 2015.
[13] Yi Huang, M. Esmalifalak, H. Nguyen, R. Zheng, Z. Han, H. Li, and L. Song, "Bad data injection in smart grid: Attack and defense mechanisms," *IEEE Commun. Mag.*, vol. 51, no. 1, pp. 27–33, Jan. 2013.
[14] M. Ozay, I. Esnaola, F. T. Yarman Vural, S. R. Kulkarni, and H. V. Poor, "Machine learning methods for attack detection in the smart grid," *IEEE Trans. Neural Netw. Learn. Syst.*, to appear.
[15] O. Vukovic, K. Cheong Sou, G. Dan, and H. Sandberg, "Network-aware mitigation of data integrity attacks on power system state estimation," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 6, pp. 1108–1118, Jul. 2012.
[16] M. Ozay, I. Esnaola, F. T. Yarman Vural, S. R. Kulkarni, and H. V. Poor, "Sparse attack construction and state estimation in the smart grid: Centralized and distributed models," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 7, pp. 1306–1318, Jul. 2013.
[17] W. Saad, Zhu Han, H. V. Poor, and T. Basar, "Game-theoretic methods for the smart grid: An overview of microgrid systems, demand-side management, and smart grid communications," *IEEE Signal Process. Mag.*, vol. 29, no. 5, pp. 86–105, Sep. 2012.
[18] I. Esnaola, S. M. Perlaza, and H. V. Poor, "Equilibria in data injection attacks," in *Proc. IEEE Global Conference on Signal and Information Processing*, Atlanta, GA, USA, Dec. 2014, pp. 779–783.
[19] M. Esmalifalak, G. Shi, Z. Han, and L. Song, "Bad data injection attack and defense in electricity market using game theory study," *IEEE Trans. Smart Grid*, vol. 4, no. 1, pp. 160–169, Mar. 2013.
[20] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed. New York: Springer-Verlag, 1994.
[21] I. Esnaola, S. M. Perlaza, H. V. Poor, and O. Kosut, "Decentralized maximum distortion MMSE attacks in electricity grids," *INRIA, Lyon, Tech. Rep. 466*, Sep. 2015.
[22] J. F. Nash, "Equilibrium points in n-person games," *Proc. National Academy of Sciences of the United States of America*, vol. 36, no. 1, pp. 48–49, Jan. 1950.
[23] D. Monderer and L. S. Shapley, "Potential games," *Games and Economic Behavior*, vol. 14, no. 1, pp. 124–143, May 1996.
[24] R. M. Gray, "Toeplitz and circulant matrices: A review," *Foundations and Trends in Communications and Information Theory*, vol. 2, no. 3, pp. 155–239, 2006.

Iñaki Esnaola (S'08–M'11) received the combined BSc and MSc degree in Electrical Engineering from the University of Navarra, Spain in 2006 and the PhD in Electrical Engineering from the University of Delaware, Newark, DE in 2011. He is currently a Lecturer (Assistant Professor) in the Department of Automatic Control and Systems Engineering at The University of Sheffield, UK, and a Visiting Research Collaborator in the Department of Electrical Engineering at Princeton University, Princeton, NJ. In 2010-2011 he was a Research Intern at Bell Laboratories, Alcatel-Lucent, Holmdel, NJ, and in 2011-2013 he was a Postdoctoral Research Associate at Princeton University. His research interests include information theory and communication theory with an emphasis on the application to electricity grid problems.

Samir Perlaza (S'07–M'11–SM'15) is a research scientist with the Institut National de Recherche en Informatique et en Automatique (INRIA), France, and a visiting research collaborator at the School of Applied Science at Princeton University (NJ, USA). He received the M.Sc. and Ph.D. degrees from École Nationale Supérieure des Télécommunications (Telecom ParisTech), Paris, France, in 2008 and 2011, respectively. Previously, from 2008 to 2011, he was a Research Engineer at France Télécom - Orange Labs (Paris, France). He has held long-term academic appointments at the Alcatel-Lucent Chair in Flexible Radio at Supélec (Gif-sur-Yvette, France); at Princeton University (Princeton, NJ) and at the University of Houston (Houston, TX). His research interests lie in the overlap of signal processing, information theory, game theory and wireless communications. Dr. Perlaza has been distinguished by the European Commission with an Alban Fellowship in 2006 and a Marie Skłodowska-Curie Fellowship in 2015. He was also one of the recipients of the the Best Student Paper Award at Crowncom in 2009.

H. Vincent Poor (S72– M77–SM82–F87) received the Ph.D. degree in EECS from Princeton University in 1977. From 1977 until 1990, he was on the faculty of the University of Illinois at Urbana-Champaign. Since 1990 he has been on the faculty at Princeton, where he is the Michael Henry Strater University Professor of Electrical Engineering and Dean of the School of Engineering and Applied Science. Dr. Poor's research interests are in the areas of stochastic analysis, statistical signal processing, and information theory, and their applications in wireless networks and related fields including social networks and smart grid. Among his publications in these areas is the recent book *Mechanisms and Games for Dynamic Spectrum Allocation* (Cambridge University Press, 2014).

Dr. Poor is a Member of the National Academy of Engineering and the National Academy of Sciences, and a Foreign Member of Academia Europaea and the Royal Society. He is also a Fellow of the American Academy of Arts and Sciences and of other national and international academies. In 1990, he served as President of the IEEE Information Theory Society, and in 2004-07 he served as the Editor-in-Chief of the *IEEE Transactions on Information Theory*. He received a Guggenheim Fellowship in 2002 and the IEEE Education Medal in 2005. Recent recognition of his work includes the 2014 URSI Booker Gold Medal, the 2015 EURASIP Athanasios Papoulis Award, the 2016 John Fritz Medal, and honorary doctorates from Aalborg University, Aalto University, HKUST and the University of Edinburgh.

Oliver Kosut (S'06–M'10) received B.S. degrees in electrical engineering and mathematics from the Massachusetts Institute of Technology, Cambridge, MA in 2004 and the Ph.D. degree in electrical and computer engineering from Cornell, Ithaca, NY in 2010. Since 2012, he has been an Assistant Professor at Arizona State University, Tempe, AZ. Previously, he was a Postdoctoral Research Associate in the Laboratory for Information and Decision Systems at MIT from 2010 to 2012, and a visiting student at the University of California, Berkeley, CA in 2008–9. His research interests include information theory, cyber-security, and power systems. Prof. Kosut received the NSF CAREER award in 2015.