Decentralized Maximum Distortion MMSE Attacks in Electricity Grids

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

▶ **To cite this version:**

Iñaki Esnaola, Samir M. Perlaza, H. Vincent Poor, Oliver Kosut. Decentralized Maximum Distortion MMSE Attacks in Electricity Grids. [Technical Report] RT-0466, Inria - Research Centre Grenoble – Rhône-Alpes. 2015, pp.26. hal-01194369

HAL Id: hal-01194369

https://inria.hal.science/hal-01194369v1

Submitted on 6 Sep 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

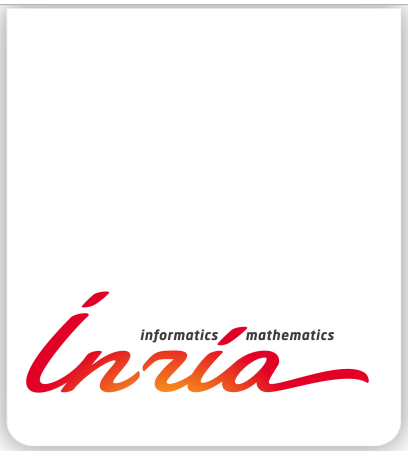
L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Decentralized Maximum Distortion MMSE Attacks in Electricity Grids

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

TECHNICAL REPORT
N° 466
September 2015
Project-Team Socrate

ISSN 0249-0803 ISRN INRIA/RT--466--FR+ENG



Decentralized Maximum Distortion MMSE Attacks in Electricity Grids

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

Project-Team Socrate

Technical Report n° 466 — September 2015 — 24 pages

Abstract: Multiple attacker data injection attack construction in electricity grids with minimum-mean-square-error (MMSE) 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. Within 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 (NE) in the game is analyzed. For the particular case of two attackers, numerical results based on IEEE test systems are presented. These results suggest that attackers perform better when they seize control of power flow measurements instead of power injection measurements.

Key-words: MMSE, Data Injection Attacks, Smart Grids, Decentralized Attacks, Nash Equilibrium, IEEE Test Systems.

Iñaki Esnaola is with the Department of Automatic Control and Systems Engineering, University of Sheffield, S1 3JD, UK. He is also with the Department of Electrical Engineering at Princeton University, Princeton, NJ 08544, USA. (esnaola@sheffield.ac.uk). Samir M. Perlaza is with the Institut National de Recherche en Informatique et Automatique (INRIA), Lyon, France. He is also with the Department of Electrical Engineering at Princeton University, Princeton, NJ 08544, USA. (samir.perlaza@inria.fr). H. Vincent Poor is with the Department of Electrical Engineering at Princeton University, Princeton, NJ 08544, USA. (poor@princeton.edu). Oliver Kosut is with the School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85287. (okosut@asu.edu)

This research was supported in part by the European Commission under Marie Sklodowska-Curie Individual Fellowship No. 659316 (CYBERNETS) and in part by the U.S. National Science Foundation under Grant CPS-1449080. This work has been submitted in part to the IEEE Transactions on Smart Grid - Special Issue on Theory of Complex Systems with Applications to Smart Grid Operations in September 2015.

RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Attaques Décentralisées de Distorsion MMSE Maximale dans les Réseaux Éléctriques

Résumé : Dans ce rapport, les constructions centralisée et décentralisée d'attaques d'injection de données vers les réseaux de distribution d'électricité sont étudiées sous la condition d'une estimation d'état basée sur la minimisation de l'erreur quadratique moyenne (MMSE pour minimum mean squared error). Une analyse du compromis entre la distorsion maximale induite par une attaque et sa probabilité de détection est présentée. A partir de cette analyse, les attaques optimales dans les cas centralisé et décentralisé sont caractérisés en utilisant des arguments de la théorie de matrices et la théorie de jeux. Dans le cas décentralisé, il est montré que l'interaction entre tous les attaquants peut être modélisée par un jeux de potentiel en forme normale. La cardinalité de l'ensemble d'équilibres de Nash est bornée en fonction des paramètres du réseau. Pour le cas particulier de deux attaquants dans un système de test IEEE, des résultas numériques suggèrent que les attaquants arrivent à induire une erreur quadratique moyenne plus importante lorsque les mesures de flux de puissance sont atteintes au lieu des mesures d'injection de puissance.

Mots-clés : MMSE, attaques centralisées et décentralisées d'injection de données, réseaux de distribution d'électricité, équilibres de Nash, système de test IEEE.

Contents

1 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 obtained by a network operator. This problem was first formulated in [2]. Therein, attacks are studied and construction procedures for attackers with access to a limited number of meters were presented. However, the analysis in [2] 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 where the statistical structure of the state variables is exploited. In [3], the state variables are modeled as a multivariate Gaussian process whose second order moments are available to the attacker and the operator. Therein, 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 detection probability in which the attacker incurs. A framework for analyzing the joint estimation and attack detection under structured data attacks is presented in [4]. Attack construction and detection with imperfect system model information are studied in [5, 6]. Alternatively, when the operator has access to training data, machine learning techniques are effective attack detection approaches [7].

Given the complexity and extension of most electricity grids, it is plausible to think of scenarios in which several attackers intrude 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 [8, 9, 10].

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 the smart grid services and applications that can be tackled with game theory is given in [11]. In [12], centralized data injection attacks are studied in a game theoretic setting in which the operator performs least squares estimation. 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 and different

degrees of communication/coordination capabilities among attackers are considered. Thus, tools from game theory are used to determine optimal individual behaviors and resulting distributed attacks. In particular, a novel utility function is proposed that models the features of the dynamic between the attackers and the operator. 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 (NE) 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 3. The decentralized case and the properties of the resulting game are analyzed in Section 4. Section 5 presents simulations of the attack strategies in IEEE Test Systems. The paper ends with some concluding remarks in Section 6.

2 System Model

Let $\mathbf{x} \in \mathbb{R}^N$ be a vector containing the voltages and angles at all generation and load buses, namely the state vector of a given power system. In general, these variables are observed through an acquisition function $F : \mathbb{R}^N \to \mathbb{R}^M$ determined by the components and topology of the network. The resulting measurements are corrupted by noise and might eventually be impaired by a data-injection attack vector $\mathbf{a}$. However, for simplicity a linearized observation model is considered yielding

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

where $\mathbf{H} \in \mathbb{R}^{M \times N}$ is the Jacobian of the acquisition function F 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.

2.1 State Estimation and Data-Injection Attacks

The aim of the network operator is to obtain an estimation $\hat{\mathbf{x}}$ of the state vector $\mathbf{x}$ using the observations $\mathbf{y}$. In general, linear estimators are privileged due to their simplicity and thus, $\hat{\mathbf{x}} = \mathbf{L}\mathbf{y}$, 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:

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

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 M-dimensional real Gaussian distribution with zero mean and covariance matrix $\mathbf{\Sigma_{xx}}$, the MMSE estimator is

$$\mathbf{M} = \mathbf{\Sigma_{xx}}\mathbf{H}^\mathsf{T}(\mathbf{H}\mathbf{\Sigma_{xx}}\mathbf{H}^\mathsf{T} + \sigma^2\mathbf{I})^{-1}, \tag{3}$$

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

$$\hat{\mathbf{x}}_{\mathsf{MMSE}} \triangleq \mathbf{M}\mathbf{y}. \tag{4}$$

The aim of an attacker is to choose an attack vector $\mathbf{a} \in \mathbb{R}^M$ in order to hinder the network operator's ability to estimate the state variables without being detected. Note that the impact of

the attack vector, $\mathbf{a}$, on the estimation $\hat{\mathbf{x}}_{\mathsf{MMSE}}$ is quantified by the second term in the right-hand side of the following equality

$$\hat{\mathbf{x}}_{\mathsf{MMSE}} \quad = \quad \mathbf{M}(\mathbf{H}\mathbf{x}+\mathbf{z})+\mathbf{M}\mathbf{a}. \tag{5}$$

The term $\mathbf{Ma}$ is referred to as the *excess distortion* induced by the data injection $\mathbf{a}$ and is denoted by

$$\mathbf{x}_a \triangleq \mathbf{M}\mathbf{a} = \mathbf{\Sigma}_{\mathbf{xx}}\mathbf{H}^{\mathsf{T}}(\mathbf{H}\mathbf{\Sigma}_{\mathbf{xx}}\mathbf{H}^{\mathsf{T}}+\sigma^2\mathbf{I})^{-1}\mathbf{a}. \tag{6}$$

2.2 Attack Detection

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

$$\mathcal{H}_0: \qquad \text{There is no attack} \tag{7}$$
$$\mathcal{H}_1: \qquad \text{Measurements are compromised.} \tag{8}$$

Assuming the operator knows that $\mathbf{x}\sim\mathcal{N}(0,\mathbf{\Sigma}_{\mathbf{xx}})$ it can obtain the joint density function of the measurements and the state variables. From (1) 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:

$$\mathbf{\Sigma}_{\mathbf{yy}} = \mathbf{H}\mathbf{\Sigma}_{\mathbf{xx}}\mathbf{H}^{\mathsf{T}}+\sigma^2\mathbf{I}, \tag{9}$$

and mean $\mathbf{a}$ when there is an attack or zero mean when there is no attack. Within this setting, the hypothesis testing described before adapted to the attack detection problem compares the following hypotheses:

$$\mathcal{H}_0: \qquad \mathbf{y}\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma}_{\mathbf{yy}}) \tag{10}$$
$$\mathcal{H}_1: \qquad \mathbf{y}\sim\mathcal{N}(\mathbf{a},\mathbf{\Sigma}_{\mathbf{yy}}). \tag{11}$$

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 or not the grid is under attack 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

$$L(\mathbf{y},\mathbf{a}) = \frac{f_{\mathcal{N}(\mathbf{0},\mathbf{\Sigma}_{\mathbf{yy}})}(\mathbf{y})}{f_{\mathcal{N}(\mathbf{a},\mathbf{\Sigma}_{\mathbf{yy}})}(\mathbf{y})} = \exp\left(\frac{1}{2}\mathbf{a}^{\mathsf{T}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}-\mathbf{a}^{\mathsf{T}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{y}\right). \tag{12}$$

Hence, either hypothesis is accepted by evaluating the inequalities:

$$L(\mathbf{y},\mathbf{a}) \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\gtrless}} \tau, \tag{13}$$

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

3 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 (1). 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 Sec. 4.

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 when 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:

$$\mathcal{A} = \left\{ \mathbf{a} : \mathbf{a}^{\mathsf{T}}\mathbf{a} \leqslant E \text{ and } \mathbf{a} \in \mathbb{R}^M \right\}. \tag{14}$$

3.1 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 (6) that it might induce 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

$$\mathsf{P}_{\mathsf{ND}}(\mathbf{a}) = \mathsf{E}\left(\mathbb{1}_{\{\mathsf{L}(\mathbf{y},\mathbf{a})>\tau\}}\right), \tag{15}$$

where the expectation is taken over the state variables $\mathbf{x}$ and the noise $\mathbf{z}$. Note that under these assumptions, $\mathbf{y}$ is a random variable with Gaussian distribution with mean $\mathbf{a}$ and covariance matrix $\mathbf{\Sigma_{yy}}$. The following Lemma provides the exact probability $\mathsf{P}_{\mathsf{ND}}(\mathbf{a})$ of a vector $\mathbf{a}$ being a successful attack, i.e., a non-detected attack.

Proposition 1 (Probability of Non-Detection)**.** *For all $\mathbf{a} \in \mathcal{A}$, it holds that*

$$\mathsf{P}_{\mathsf{ND}}(\mathbf{a}) = \frac{1}{2}\mathsf{erfc}\left(\frac{\frac{1}{2}\mathbf{a}^{\mathsf{T}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a} + \log\tau}{\sqrt{2\mathbf{a}^{\mathsf{T}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}}}\right). \tag{16}$$

Proof. Consider the set $\mathcal{S}(\mathbf{a})$ of all possible realizations of $\mathbf{y}$ such that even in the presence of a data-injection attack $\mathbf{a}$, the hypothesis $\mathcal{H}_0$ is chosen. That is,

$$\begin{aligned} \mathcal{S}(\mathbf{a}) &= \left\{\mathbf{y} \in \mathbb{R}^M : L(\mathbf{y}, \mathbf{a}) \geqslant \tau\right\} \\ &= \left\{\mathbf{y} \in \mathbb{R}^M : -\mathbf{a}^{\mathsf{T}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{y} \geqslant \log\tau - \frac{1}{2}\mathbf{a}^{\mathsf{T}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}\right\}. \end{aligned} \tag{17}$$

Then, the probability $\mathsf{P}_{\mathsf{ND}}(\mathbf{a})$ in (15) can be written as follows:

$$\mathsf{P}_{\mathsf{ND}}(\mathbf{a}) = \int_{\mathcal{S}(\mathbf{a})} f(\mathbf{y} - \mathbf{a})\mathrm{d}\mathbf{y}, \tag{18}$$

where f is the probability density function of $\mathcal{N}(\mathbf{a}, \mathbf{\Sigma_{yy}})$. Let also

$$\mathbf{b}^{\mathsf{T}} = -\mathbf{a}^{\mathsf{T}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1} \text{ and} \tag{19}$$

$$c = \log\tau - \frac{1}{2}\mathbf{a}^{\mathsf{T}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}. \tag{20}$$

From (18), the following holds:

$$\begin{aligned}
\mathsf{P}_{\mathsf{ND}}(\mathbf{a}) &\overset{(a)}{=} \int_{\mathcal{S}_1(\mathbf{a})} f(\mathbf{y}_1)\mathrm{d}\mathbf{y}_1, \qquad (21)\\
&\overset{(b)}{=} \frac{1}{\sqrt{(2\pi)^M \det\mathbf{\Sigma_{yy}}}} \int_{\mathcal{S}_2(\mathbf{a})} \exp\left(-\frac{1}{2}\mathbf{y}_2^{\mathsf{T}}\mathbf{\Lambda}_{\mathbf{yy}}^{-1}\mathbf{y}_2\right)\mathrm{d}\mathbf{y}_2,\\
&\overset{(c)}{=} \frac{1}{\sqrt{(2\pi)^M}} \int_{\mathcal{S}_3(\mathbf{a})} \exp\left(-\frac{1}{2}\mathbf{y}_3^{\mathsf{T}}\mathbf{y}_3\right)\mathrm{d}\mathbf{y}_3,\\
&\overset{(d)}{=} \frac{1}{\sqrt{(2\pi)^M}} \int_{\mathcal{S}_4(\mathbf{a})} \exp\left(-\frac{1}{2}\mathbf{y}_4^{\mathsf{T}}\mathbf{y}_4\right)\mathrm{d}\mathbf{y}_4,\\
&\overset{(e)}{=} \frac{1}{\sqrt{(2\pi)}} \int_{d}^{\infty} \exp\left(-\frac{1}{2}v^2\right)\mathrm{d}v,\\
&\overset{(f)}{=} \frac{1}{2}\mathsf{erfc}\left(\frac{\frac{1}{2}\mathbf{a}^{\mathsf{T}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a} + \log\tau}{\sqrt{2\mathbf{a}^{\mathsf{T}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}}}\right),
\end{aligned}$$

where (a) follows from a change of variable

$$\mathbf{y}_1 = \mathbf{y} - \mathbf{a} \tag{22}$$

and integration domain

$$\mathcal{S}_1(\mathbf{a}) = \left\{\mathbf{y}_1 \in \mathbb{R}^M : \mathbf{b}^{\mathsf{T}}\mathbf{y}_1 + (\mathbf{b}^{\mathsf{T}}\mathbf{a} - c) \geqslant 0\right\}; \tag{23}$$

(b) uses an SVD of $\mathbf{\Sigma_{yy}}$ of the form

$$\mathbf{\Sigma_{yy}} = \mathbf{U}_{\mathbf{yy}}^{\mathsf{T}}\mathbf{\Lambda_{yy}}\mathbf{U_{yy}}, \tag{24}$$

with $\mathbf{U_{yy}}$ a unitary matrix and Λ_{yy} a diagonal matrix with strictly positive diagonal entries, the change of variables

$$\mathbf{y}_2 = \mathbf{U_{yy}}\,\mathbf{y}_1 \tag{25}$$

and the integration domain

$$\mathcal{S}_2(\mathbf{a}) = \left\{\mathbf{y}_2 \in \mathbb{R}^M : \mathbf{b}_1^{\mathsf{T}}\mathbf{y}_2 + (\mathbf{b}^{\mathsf{T}}\mathbf{a} - c) \geqslant 0\right\}, \tag{26}$$

with $\mathbf{b}_1 = \mathbf{U_{yy}}\mathbf{b}$;
(c) follows from a change of variables

$$\mathbf{y}_3 = \mathbf{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}}\mathbf{y}_2 \tag{27}$$

and integration domain

$$\mathcal{S}_3(\mathbf{a}) = \left\{\mathbf{y}_3 \in \mathbb{R}^M : \mathbf{b}_2^{\mathsf{T}}\mathbf{y}_3 + (\mathbf{b}^{\mathsf{T}}\mathbf{a} - c) \geqslant 0\right\}, \tag{28}$$

with

$$\mathbf{b}_2 = \mathbf{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}} \mathbf{b}_1; \tag{29}$$

(d) uses the fact that it is always possible to build an orthogonal matrix with the n-th column given by the vector $\frac{\mathbf{b}_2}{||\mathbf{b}_2||_2}$, with $n \in \{1, \ldots, M\}$, and thus,

$$\mathbf{B}\, \mathbf{b}_2 = ||\mathbf{b}_2||_2 \mathbf{e}_n \tag{30}$$

and the change of variables

$$\mathbf{y}_4 = \mathbf{B}\, \mathbf{y}_3 \tag{31}$$

with integration domain

$$\begin{aligned} \mathcal{S}_4(\mathbf{a}) \quad = \quad & \{\mathbf{y}_4 = (y_{4,1}, \ldots, y_{4,M}) \in \mathbb{R}^M : & (32) \\ & y_{4,n}||\mathbf{b}_2||_2 + (\mathbf{b}^\mathsf{T}\mathbf{a} - c) \geqslant 0\}; & (33) \end{aligned}$$

(e) follows from solving the $M-1$ integrals of the single dimension density function of a zero mean and unitary variance of a Gaussian random variable over $[-\infty, \infty]$ and

$$d = y_{4,n} = -\frac{\mathbf{b}^\mathsf{T}\mathbf{a} - c}{||\mathbf{b}_2||_2} = \frac{\frac{1}{2}\mathbf{a}^\mathsf{T}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a} + \log\tau}{\sqrt{2\mathbf{a}^\mathsf{T}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}}}; \tag{34}$$

and (f) follows from the definition of the complementary error function

$$\operatorname{erfc}(d) = \frac{2}{\sqrt{\pi}} \int_d^\infty \exp\left(-t^2\right) \mathrm{d}t. \tag{35}$$

This completes the proof.

□

Often, the knowledge of the threshold τ in (13) is not available to the attacker and thus, it cannot determine the exact average probability of a given attack vector $\mathbf{a}$ to be successful. However, possessing the knowledge of whether $\tau \geqslant 1$ or $\tau \leqslant 1$ might induce different behaviors on the attacker. The following corollaries follow immediately from Propostion 1 and the properties of the complementary error function.

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

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

$$\mathbf{a} = \pm\sqrt{\lambda_{yy,k} 2\log\tau}\mathbf{u}_{yy,k}, \tag{36}$$

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

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

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

$$g(x) = \frac{\frac{1}{2}x + \log\tau}{\sqrt{2x}}. \tag{37}$$

The first derivative of $g(x)$ is

$$g'(x) = \frac{1}{2\sqrt{2x}} \left(\frac{1}{2} - \frac{\log \tau}{x} \right). \tag{38}$$

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 Corollary 2 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 $\frac{1}{2}\text{erfc}(g(x))$ occurs at the minimum of $g(x)$ given that erfc is monotonically decreasing with its argument. Hence, the minimum of $\mathsf{P}_{\mathsf{ND}}(\mathbf{a})$ occurs at any $\mathbf{a}$ satisfying the condition:

$$\mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{a} = 2 \log \tau. \tag{39}$$

Solving for $\mathbf{a}$ in (39) yields (36) and this completes the proof of Corollary 2. □

The relevance of Corollary 1 is that it states that when $\tau \geqslant 1$, 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., no-attack. Alternatively, when $\tau > 1$ it follows from Corollary 2 that there always exists a non-zero vector that possesses minimum probability of not being detected. However, in both cases, it is clear that the corresponding data-injection vectors which induce the lowest 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 minimizes the probability $\mathsf{P}_{\mathsf{ND}}(\mathbf{a})$ of being detected at the same time that it induces a given distortion $||\mathbf{x}_a|| \geqslant \mathsf{D}_0$ to the measurements. In the case in which $\tau \leqslant 1$, it follows from Corollary 1 and (6) that this problem can be formulated as the following optimization problem:

$$\min_{\mathbf{a} \in \mathcal{A}} \mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{a} \quad \text{s.t.} \quad \mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{H} \boldsymbol{\Sigma}_{\mathbf{xx}}^{2} \mathbf{H}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{a} \geq \mathsf{D}_0. \tag{40}$$

The solution to the optimization problem in (40) is given by the following proposition.

Proposition 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* $\mathbf{U} = (\mathbf{u}_{\mathbf{G},i}, \ldots, \mathbf{u}_{\mathbf{G},M})$ *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* $\tau \leqslant 1$, *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*

$$\mathbf{a} \quad = \quad \sqrt{\frac{\mathsf{D}_0}{\lambda_{\mathbf{G},1}}}\, \boldsymbol{\Sigma}_{\mathbf{yy}}^{\frac{1}{2}} \mathbf{u}_{\mathbf{G},1}. \tag{41}$$

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

Proof. Consider the Lagrangian

$$L(\mathbf{a}) = \mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{a} - \gamma \left(\mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{H} \boldsymbol{\Sigma}_{\mathbf{xx}}^{2} \mathbf{H}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{a} - \mathsf{D}_0 \right), \tag{42}$$

with $\gamma > 0$ a Lagrangian multiplier. Then, necessary conditions for $\mathbf{a}$ to be a solution to the optimization problem (40) are:

$$\nabla_{\mathbf{a}} L(\mathbf{a}) = 2\left(\mathbf{\Sigma}_{\mathbf{yy}}^{-1} - \gamma \mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{H}\mathbf{\Sigma}_{\mathbf{xx}}^{2}\mathbf{H}^{\mathsf{T}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\right)\mathbf{a} = 0 \tag{43}$$

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} L(\mathbf{a}) = \mathbf{a}^{\mathsf{T}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{H}\mathbf{\Sigma}_{\mathbf{xx}}^{2}\mathbf{H}^{\mathsf{T}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a} - \mathsf{D}_0 = 0. \tag{44}$$

Note that any

$$\mathbf{a}_i = \sqrt{\frac{\mathsf{D}_0}{\lambda_{\mathbf{G},i}}}\mathbf{\Sigma}_{\mathbf{yy}}^{\frac{1}{2}}\mathbf{u}_{\mathbf{G},i} \text{ and} \tag{45}$$

$$\gamma_i = \lambda_{\mathbf{G},i}, \tag{46}$$

with $1 \leqslant i \leqslant \mathrm{rank}\,(\mathbf{G})$, satisfy $\gamma_i > 0$ and conditions (43) and (44). Hence, the set of vectors that satisfy the necessary conditions to be a solution of (40) is

$$\left\{\mathbf{a}_i = \pm\sqrt{\frac{\mathsf{D}_0}{\lambda_{\mathbf{G},i}}}\mathbf{\Sigma}_{\mathbf{yy}}^{\frac{1}{2}}\mathbf{u}_{\mathbf{G},i} : 1 \leqslant i \leqslant \mathrm{rank}\,(\mathbf{G})\right\}. \tag{47}$$

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,

$$\mathbf{a}_i^{\mathsf{T}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}_i = \frac{\mathsf{D}_0}{\lambda_{\mathbf{G},i}} \geqslant \frac{\mathsf{D}_0}{\lambda_{\mathbf{G},1}}. \tag{48}$$

Therefore, $\mathbf{a} = \pm\sqrt{\frac{\mathsf{D}_0}{\lambda_{\mathbf{G},1}}}\mathbf{\Sigma}_{\mathbf{yy}}^{\frac{1}{2}}\mathbf{u}_{\mathbf{G},1}$ are the unique solutions to (40). This completes the proof. □

Note that the construction of the data-injection attack $\mathbf{a}$ in (62) does not require the exact knowledge of τ. That is, only by knowing that $\tau \leqslant 1$ is enough to build the data-injection attack that has the highest probability of not being detected and produces 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 produces 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.

$$\min_{\mathbf{a}\in\mathcal{A}} \frac{\frac{1}{2}\mathbf{a}^{\mathsf{T}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a} + \log\tau}{\sqrt{2\mathbf{a}^{\mathsf{T}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}}} \text{ s.t. } \mathbf{a}^{\mathsf{T}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{H}\mathbf{\Sigma}_{\mathbf{xx}}^{2}\mathbf{H}^{\mathsf{T}}\mathbf{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a} \geq \mathsf{D}_0. \tag{49}$$

The solution to the optimization problem in (49) is given by the following proposition.

Proposition 3. *Let* $\mathbf{G} = \mathbf{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}}\mathbf{H}\mathbf{\Sigma}_{\mathbf{xx}}^{2}\mathbf{H}^{\mathsf{T}}\mathbf{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}}$ *have a singular value decomposition* $\mathbf{G} = \mathbf{U}_{\mathbf{G}}\mathbf{\Sigma}_{\mathbf{G}}\mathbf{U}_{\mathbf{G}}^{\mathsf{T}}$, *with* $\mathbf{U}_{\mathbf{G}} = (\mathbf{u}_{\mathbf{G},i}, \ldots, \mathbf{u}_{\mathbf{G},M})$ *a unitary matrix and* $\mathbf{\Sigma}_{\mathbf{G}} = diag\,(\lambda_{\mathbf{G},1}, \ldots, \lambda_{\mathbf{G},M})$ *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*

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

with

$$k^* = \arg \min_{k\in\{1,\ldots,\mathrm{rank}\mathbf{G}\}:\frac{\mathsf{D}_0}{\lambda_{\mathbf{G},k}} > 2\log(\tau)} \frac{\mathsf{D}_0}{\lambda_{\mathbf{G},k}}. \tag{50}$$

The proof of Proposition 3 is presented hereunder.

Proof. Consider the following Lagrangian

$$L(\mathbf{a}) = \frac{\frac{1}{2}\mathbf{a}^\mathsf{T}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a} + \log\tau}{\sqrt{2\mathbf{a}^\mathsf{T}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}}} - \gamma\left(\mathbf{a}^\mathsf{T}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}}\mathbf{G}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}}\mathbf{a} - \mathsf{D}_0\right) \tag{51}$$

with $\gamma > 0$ a Lagrangian multiplier. Then, necessary conditions for $\mathbf{a}$ to be a solution of the optimization problem (49) are:

$$\begin{aligned}
\nabla_{\mathbf{a}} L(\mathbf{a}) &= \frac{1}{\sqrt{2\mathbf{a}^\mathsf{T}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}}}\left(\frac{1}{2} - \frac{\log\tau}{\mathbf{a}^\mathsf{T}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}}\right)\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a} \\
&\quad -2\gamma\boldsymbol{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}}\mathbf{G}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}}\mathbf{a} = 0, \text{ and} \qquad (52) \\
\frac{\mathrm{d}}{\mathrm{d}\gamma}L(\mathbf{a}) &= \mathbf{a}^\mathsf{T}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}}\mathbf{G}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}}\mathbf{a} - \mathsf{D}_0 = 0. \qquad (53)
\end{aligned}$$

Let

$$\alpha(\mathbf{a}) = \frac{1}{2\sqrt{2\mathbf{a}^\mathsf{T}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}}}\left(\frac{1}{2} - \frac{\log\tau}{\mathbf{a}^\mathsf{T}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}}\right) \tag{54}$$

and note that any

$$\begin{aligned}
\mathbf{a}_i &= \pm\sqrt{\frac{\mathsf{D}_0}{\lambda_{\mathbf{G},i}}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{\frac{1}{2}}\mathbf{u}_{\mathbf{G},i} \text{ and} \qquad (55) \\
\gamma_i &= \alpha(\mathbf{a}_i)\sqrt{\frac{\lambda_{\mathbf{G},i}}{\mathsf{D}_0}}, \qquad (56)
\end{aligned}$$

with $1 \leqslant i \leqslant \operatorname{rank}(\mathbf{G})$, satisfy conditions (52) and (53), when $\alpha(\mathbf{a}_i) > 0$, i.e., $\frac{\mathsf{D}_0}{\lambda_{\mathbf{G},i}} \geqslant 2\log\tau$. Hence, the set of vectors that satisfy the necessary conditions to be a solution of (49) is

$$\left\{\mathbf{a}_i = \pm\sqrt{\frac{\mathsf{D}_0}{\lambda_{\mathbf{G},i}}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{\frac{1}{2}}\mathbf{u}_{\mathbf{G},i} : i \in \{k^*, \ldots, \operatorname{rank}\mathbf{G}\}\right\}, \tag{57}$$

with k^* as in (50). More importantly, any vector $\mathbf{a} \neq \mathbf{a}_i$, with $i \in \{k^*, \ldots, \operatorname{rank}\mathbf{G}\}$, does not satisfy the necessary conditions. Moreover,

$$\frac{\frac{1}{2}\mathbf{a}^\mathsf{T}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a} + \log\tau}{\sqrt{2\mathbf{a}^\mathsf{T}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}}} \geqslant \frac{\frac{1}{2}\frac{\mathsf{D}_0}{\lambda_{\mathbf{G},k^*}} + \log\tau}{\sqrt{2\frac{\mathsf{D}_0}{\lambda_{\mathbf{G},k^*}}}}, \tag{58}$$

and thus, $\mathbf{a} = \pm\sqrt{\frac{\mathsf{D}_0}{\lambda_{\mathbf{G},k^*}}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{\frac{1}{2}}\mathbf{u}_{\mathbf{G},k^*}$ are the unique solutions to (49) when $\frac{\mathsf{D}_0}{\lambda_{\mathbf{G},i}} \geqslant 2\log\tau$.

When $\alpha(\mathbf{a}) < 0$, that is, $\frac{\mathsf{D}_0}{\lambda_{\mathbf{G},\operatorname{rank}\mathbf{G}}} < 2\log\tau$, it does not exists a $\lambda_{\mathbf{G},i} > 0$, with $i \in \{1, \ldots, \operatorname{rank}\mathbf{G}\}$, that satisfies $\frac{\mathsf{D}_0}{\lambda_{\mathbf{G},i}} \geqslant 2\log\tau$. It can be verified that the objective function is monotonically decreasing with $\mathbf{a}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\mathbf{a}$ in the interval $(0, 2\log\tau)$. Thus, the choice

$$\mathbf{a} = \sqrt{2\log\tau}\boldsymbol{\Sigma}_{\mathbf{yy}}^{\frac{1}{2}}\mathbf{u}_{\mathbf{G},i} \tag{59}$$

minimizes the objective function and then the constraint in (49) is always satisfied. The choice $i = 1$ is made given that it is the one that produces the highest distortion, i.e.,

$$\mathbf{a}_i^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{H} \boldsymbol{\Sigma}_{\mathbf{xx}}^{2} \mathbf{H}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{a}_i = 2\lambda_{\mathbf{G},i} \log \tau > \mathsf{D}_0. \tag{60}$$

This completes the proof. □

3.2 Attacks with Maximum Distortion

In the previous subsection, the attacker designed its data-injection vector $\mathbf{a}$ aiming to minimize the probability of non-detection $\mathsf{P}_{\mathsf{ND}}(\mathbf{a})$ while guaranteeing a minimum distortion. However, this problem has 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 smaller than a given threshold $L_0' \in [0, \frac{1}{2}]$. This problem can be formulated as the following optimization problem:

$$\max_{\mathbf{a} \in \mathcal{A}} \mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{H} \boldsymbol{\Sigma}_{\mathbf{xx}}^{2} \mathbf{H}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{a} \text{ s.t. } \frac{\frac{1}{2} \mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{a} + \log \tau}{\sqrt{2 \mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{a}}} \leq L_0, \tag{61}$$

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

The solution to the optimization problem in (61) is given by the following proposition.

Proposition 4. *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 exceed* $L_0 \in [0, \frac{1}{2}]$ *is*

$$\mathbf{a} = \pm \left(\sqrt{2} L_0 + \sqrt{2L_0^2 - 2 \log \tau} \right) \boldsymbol{\Sigma}_{\mathbf{yy}}^{\frac{1}{2}} \mathbf{u}_{\mathbf{G},1}, \tag{62}$$

when a solution exists.

Proof. Consider the Lagrangian of the optimization problem in (61):

$$L(\mathbf{a}) = -\mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}} \mathbf{G} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}} \mathbf{a} + \gamma \left(\frac{\frac{1}{2} \mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{a} + \log \tau}{\sqrt{2 \mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{a}}} - L_0 \right) \tag{63}$$

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

$$\nabla_{\mathbf{a}} L(\mathbf{a}) = \boldsymbol{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}} \left(\mathbf{G} - \gamma \alpha(\mathbf{a}) \boldsymbol{I} \right) \boldsymbol{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}} \mathbf{a} = 0, \tag{64}$$

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} L(\mathbf{a}) = \frac{\frac{1}{2} \mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{a} + \log \tau}{\sqrt{2 \mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{a}}} - L_0 = 0, \tag{65}$$

with $\alpha(\mathbf{a})$ as defined in (54).

Assume that $\tau \leqslant \exp(L_0^2)$ and denote by $\beta_1 > \beta_2 > 0$ the two unique positive real solutions to the following equation:

$$\frac{\frac{1}{2} \beta + \log \tau}{\sqrt{2\beta}} - L_0 = 0, \tag{66}$$

that is,

$$\beta_1 = \left(\sqrt{2}L_0 + \sqrt{2L_0^2 - 2\log\tau}\right)^2 \text{ and} \tag{67}$$

$$\beta_2 = \left(\sqrt{2}L_0 - \sqrt{2L_0^2 - 2\log\tau}\right)^2. \tag{68}$$

Note that any

$$\mathbf{a}_{i,j} = \pm\sqrt{\beta_j}\boldsymbol{\Sigma}_{\mathbf{yy}}^{\frac{1}{2}}\mathbf{u}_{\mathbf{G},i} \text{ and} \tag{69}$$

$$\gamma_{ij} = \frac{1}{\alpha(\mathbf{a}_{ij})}\lambda_{\mathbf{G},i}, \tag{70}$$

with $(i,j) \in \{1,\ldots,\text{rank}\,(\mathbf{G})\} \times \{1,2\}$, satisfy conditions (64) and (65). Hence, the set of vectors that satisfy the necessary conditions to be a solution of (61) is

$$\left\{\mathbf{a}_{ij} = \pm\beta_j\boldsymbol{\Sigma}_{\mathbf{yy}}^{\frac{1}{2}}\mathbf{u}_{\mathbf{G},i} : (i,j) \in \{1,\ldots,\text{rank}\,\mathbf{G}\} \times \{1,2\}\right\}. \tag{71}$$

More importantly, any vector $\mathbf{a} \neq \mathbf{a}_{ij}$, with $(i,j) \in \{1,\ldots,\text{rank}\mathbf{G}\} \times \{1,2\}$, does not satisfy the necessary conditions. Moreover,

$$\mathbf{a}_{ij}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}}\mathbf{G}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}}\mathbf{a}_{ij} = \beta_j\lambda_{\mathbf{G},i} \leqslant \beta_1\lambda_{\mathbf{G},1} \tag{72}$$

and thus, $\mathbf{a} = \pm\beta_1\boldsymbol{\Sigma}_{\mathbf{yy}}^{\frac{1}{2}}\mathbf{u}_{\mathbf{G},1}$ are the unique solutions to (61), when $\tau \leqslant \exp(L_0^2)$.

Conversely, when $\tau > \exp(L_0^2)$, the problem is not feasible. This completes the proof. □

4 Decentralized Attacks

Let $\mathcal{K} = \{1,\ldots,K\}$ be the set of attackers that can potentially perform a data injection to 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

$$\mathcal{A}_k = \{\mathbf{a}_k \in \mathbb{R}^M : (\mathbf{a}_k)_j = 0 \text{ for all } j \notin \mathcal{C}_k, \mathbf{a}_k^T\mathbf{a}_k \leq E_k\}, \tag{73}$$

where the $M \times M$ matrix $\mathbf{A}_k$ has all entries zero except the j-th diagonal entry, with $j \in \mathcal{C}_k$. The constant $E_k < \infty$ represents the energy budget of attacker k. Let the sum between any two sets $\mathcal{A}_i$ and $\mathcal{A}_j$ be denoted by the set $\mathcal{A}_i + \mathcal{A}_j$, which represents the set of all possible sums between the elements of $\mathcal{A}_i$ and $\mathcal{A}_j$. Using this notation, let the set of all possible data-injection attacks be denoted by

$$\mathcal{A} = \sum_{i\in\{1,\ldots,K\}} \mathcal{A}_i, \tag{74}$$

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

$$\mathcal{A}_{-k} = \sum_{i\in\{1,\ldots,K\}\setminus\{k\}} \mathcal{A}_i. \tag{75}$$

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

$$\mathbf{a} = \sum_{i=1}^{K} \mathbf{a}_k \ \in \mathcal{A}. \tag{76}$$

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, e.g., $\mathcal{C}_i \neq \mathcal{C}_k$, for a given pair $(i,k) \in \mathcal{K}^2$. For modeling this scenario, attackers use the utility function $\phi : \mathbb{R}^M \ \rightarrow \ \mathbb{R}$, 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

$$\mathbf{a}_{-k} = \sum_{i \in \{1,\ldots,K\} \setminus \{k\}} \mathbf{a}_i \ \in \mathcal{A}_{-k} \tag{77}$$

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. Otherwise, if the attack is detected no damage is induced into the network as the operator is able to neglect the measurements that are compromised. 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 $\mathbf{a}$:

$$\begin{aligned} \phi(\mathbf{a}) &= \mathsf{E}\left(\mathbb{1}_{\{\mathsf{L}(\mathbf{Hx}+\mathbf{z}+\mathbf{a},\mathbf{a})>\tau\}}\right) \mathbf{x}_{\mathbf{a}}^{\mathsf{T}} \mathbf{x}_{\mathbf{a}}, & (78)\\ &= \mathsf{P}_{\mathsf{ND}}(\mathbf{a})\ \mathbf{a}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{H} \boldsymbol{\Sigma}_{\mathbf{xx}}^{2} \mathbf{H}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{a}, & (79) \end{aligned}$$

where the expectation is taken over the state variables $\mathbf{x}$ and the noise $\mathbf{z}$.

4.1 Game Formulation

The benefit $\phi(\mathbf{a})$ obtained by attacker k not only depends on its own data-injection vector $\mathbf{a}_k$, but also on the data-injection vectors $\mathbf{a}_{-k}$ of all the other attackers. This becomes clear from the construction of the global data-injection vector $\mathbf{a}$ in (76), the excess distortion $\mathbf{x}_a$ in (6) and the probability of not being detected $\mathsf{P}_{\mathsf{ND}}(\mathbf{a})$ in (16). Therefore, the interaction of all attackers in the network can be described by a game in normal form

$$\mathcal{G} = \left(\mathcal{K}, \{\mathcal{A}_k\}_{k \in \mathcal{K}}, \phi\right). \tag{80}$$

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 (73). 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,

$$\mathbf{a}_k \in \mathrm{BR}_k\left(\mathbf{a}_{-k}\right), \tag{81}$$

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

$$\mathrm{BR}_k\left(\mathbf{a}_{-k}\right) = \arg\max_{\mathbf{a}_k \in \mathcal{A}_k} \phi\left(\mathbf{a}_k + \mathbf{a}_{-k}\right). \tag{82}$$

From this perspective, a game solution that is particularly relevant for this analysis is the Nash equilibrium [13].

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*

$$\mathbf{a} = \mathrm{BR}\left(\mathbf{a}\right), \tag{83}$$

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

$$\mathrm{BR}\left(\mathbf{a}\right) = \mathrm{BR}_1\left(\mathbf{a}_{-1}\right) + \ldots + \mathrm{BR}_K\left(\mathbf{a}_{-K}\right). \tag{84}$$

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 (80).

Proposition 5. *The game* $\mathcal{G}$ *in* (80) *is a potential game.*

Proof. The proof follows immediately from the observation that all the players have the same utility function ϕ [14]. Thus, the function ϕ is a potential of the game $\mathcal{G}$ in (80) and any maximum of the potential function is an NE of the game $\mathcal{G}$. □

In general, potential games [14] possess numerous properties that are inherited by the game $\mathcal{G}$ in (80). These properties are detailed by the following propositions

Proposition 6. *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 [14], it follows that such a maximum corresponds to an NE. □

4.2 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 $\mathbf{a}_k^{(t)} \in \mathcal{A}$ 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 a best response dynamics if there exists an round-robin order of the elements of* $\mathcal{K}$ *in which at each round* $t \in \mathbb{N}$*, the following holds*

$$\mathbf{a}_k^{(t)} \in \mathrm{BR}_k\left(\mathbf{a}_1^{(t)} + \ldots + \mathbf{a}_{k-1}^{(t)} + \mathbf{a}_{k+1}^{(t-1)} + \ldots + \mathbf{a}_K^{(t-1)}\right). \tag{85}$$

From the properties of potential games (Lemma 4.2 in [14]), 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 Proposition 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.

4.3 Cardinality of the set of NEs

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

Theorem 1. *The cardinality of the set $\mathcal{A}_{\text{NE}}$ of NE of the game $\mathcal{G}$ satisfies*

$$2 \leqslant |\mathcal{A}_{\text{NE}}| \leqslant C \cdot \text{rank}(\mathbf{H}) \tag{86}$$

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

Proof. The lower bound follows from the symmetry of the utility function given in (78), i.e. $\phi(\mathbf{a}) = \phi(-\mathbf{a})$, and the existence of at least one NE claimed in Proposition 6.

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

$$\mathcal{S} = \{\mathbf{a} \in \mathbb{R}^M : \nabla_{\mathbf{a}} \phi(\mathbf{a}) = \mathbf{0}\}, \tag{87}$$

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

$$\nabla_{\mathbf{a}} \phi(\mathbf{a}) = \left(\alpha(\mathbf{a}) \mathbf{M}^{\mathsf{T}} \mathbf{M} - \beta(\mathbf{a}) \mathbf{\Sigma}_{\mathbf{yy}}^{-1}\right) \mathbf{a}, \tag{88}$$

where

$$\alpha(\mathbf{a}) \triangleq \text{erfc}\left(\frac{1}{\sqrt{2}} \frac{\frac{1}{2} \mathbf{a}^{\mathsf{T}} \mathbf{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{a} + \log \tau}{\left(\mathbf{a}^{\mathsf{T}} \mathbf{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{a}\right)^{\frac{1}{2}}} \right) \tag{89}$$

and

$$\begin{aligned} \beta(\mathbf{a}) \quad \triangleq \quad & \frac{\mathbf{a}^{\mathsf{T}} \mathbf{M}^{\mathsf{T}} \mathbf{M} \mathbf{a}}{\sqrt{2\pi} \mathbf{a}^{\mathsf{T}} \mathbf{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{a}} \left(\frac{1}{2} - \frac{\log \tau}{\mathbf{a}^{\mathsf{T}} \mathbf{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{a}} \right) \\ & \times \exp\left(- \left(\frac{1}{\sqrt{2}} \frac{\frac{1}{2} \mathbf{a}^{\mathsf{T}} \mathbf{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{a} + \log \tau}{\left(\mathbf{a}^{\mathsf{T}} \mathbf{\Sigma}_{\mathbf{yy}}^{-1} \mathbf{a}\right)^{\frac{1}{2}}} \right)^2 \right). \end{aligned} \tag{90}$$

Define $\delta(\mathbf{a}) \triangleq \frac{\beta(\mathbf{a})}{\alpha(\mathbf{a})}$ and note that combining (3) with (88) gives the following condition for the stationary points:

$$\left(\mathbf{H} \mathbf{\Sigma}_{\mathbf{xx}}^2 \mathbf{H}^{\mathsf{T}} \mathbf{\Sigma}_{\mathbf{yy}}^{-1} - \delta(\mathbf{a}) \mathbf{I}\right) \mathbf{a} = \mathbf{0}. \tag{91}$$

Note that the number of linearly independent attack vectors that are a solution of the linear system in (91) is given by

$$\begin{aligned} R \quad &\triangleq \quad \text{rank}\left(\mathbf{H} \mathbf{\Sigma}_{\mathbf{xx}}^2 \mathbf{H}^{\mathsf{T}} \mathbf{\Sigma}_{\mathbf{yy}}^{-1}\right) && (92) \\ &= \quad \text{rank}\left(\mathbf{H}\right). && (93) \end{aligned}$$

where (93) follows from the fact that $\mathbf{\Sigma}_{\mathbf{xx}}$ and $\mathbf{\Sigma}_{\mathbf{yy}}$ are positive definite. Define the eigenvalue decomposition

$$\mathbf{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}} \mathbf{H} \mathbf{\Sigma}_{\mathbf{xx}}^2 \mathbf{H}^{\mathsf{T}} \mathbf{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^{\mathsf{T}} \tag{94}$$

where $\mathbf{\Lambda}$ is a diagonal matrix containing the ordered eigenvalues $\{\lambda_i\}_{i=1}^M$ matching the order of of the eigenvectors in $\mathbf{U}$. As a result of (92) there are R eigenvalues, λ_k, which are different from

zero and $M - R$ diagonal elements of $\mathbf{\Lambda}$ which are zero. Combining this decomposition with some algebraic manipulation, the condition for stationary points in (91) can be recast as

$$\mathbf{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}} \mathbf{U} \left(\mathbf{\Lambda} - \delta(\mathbf{a}) \mathbf{I} \right) \mathbf{U}^{\mathsf{T}} \mathbf{\Sigma}_{\mathbf{yy}}^{-\frac{1}{2}} \mathbf{a} = \mathbf{0}. \tag{95}$$

Let $w \in \mathbb{R}$ be a scaling parameter and observe that attack vectors that satisfy $\mathbf{a} = w \mathbf{\Sigma}_{\mathbf{yy}}^{\frac{1}{2}} \mathbf{U} \mathbf{e}_k$ and $\delta(\mathbf{a}) = \lambda_k$ for $k = 1, \ldots, R$ are solutions of (95). Note that the critical points associated to zero eigenvalues are not NE. In fact, 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 (87) can be rewritten based on the condition in (95) as

$$\mathcal{S} = \bigcup_{k=1}^{R} \mathcal{S}_k, \tag{96}$$

where

$$\mathcal{S}_k = \{ \mathbf{a} \in \mathbb{R}^M : \mathbf{a} = w \mathbf{\Sigma}_{\mathbf{yy}}^{\frac{1}{2}} \mathbf{U} \mathbf{e}_k \quad \text{and} \quad \delta(\mathbf{a}) = \lambda_k \}. \tag{97}$$

Indeed, there are R linearly independent solutions of (95) 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 $\delta'(w) \stackrel{\Delta}{=} \delta(w \mathbf{\Sigma}_{\mathbf{yy}}^{\frac{1}{2}} \mathbf{U} \mathbf{e}_k)$. 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

$$|\mathcal{S}| \leq \sum_{i=1}^{R} |\mathcal{S}_k| \leq \sum_{i=1}^{R} C_k \leq \max_k C_k R. \tag{98}$$

□

5 Numerical Results

In this section the properties of the game $\mathcal{G}$ described in Section 4 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 $\mathbf{\Sigma}_{\mathbf{xx}}$. 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 [15], an exponentially decaying Toeplitz model is chosen where the strength of the correlation is set by a parameter ρ, namely, $\mathbf{\Sigma}_{\mathbf{xx}} = \left[s_{ij} = \rho^{|i-j|}; i, j = 1, 2, \ldots, n \right]$. 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 (78) when two attackers are present in the IEEE 30 bus test system with. The NE equilibria have been numerically evaluated and are represented by red squares. In this example, the number of NE coincides with the lower bound in Proposition 1 and the attack vectors are antisymmetric as expected from the symmetry of the utility function.

The utility function evaluated in a 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 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 reduces the probability of detection and improves 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 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, 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.

6 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 the 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 we show that correlation can be exploited by the attackers 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. In fact, we show 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. Interestingly, we provide 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. 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 estimation of the network operator.

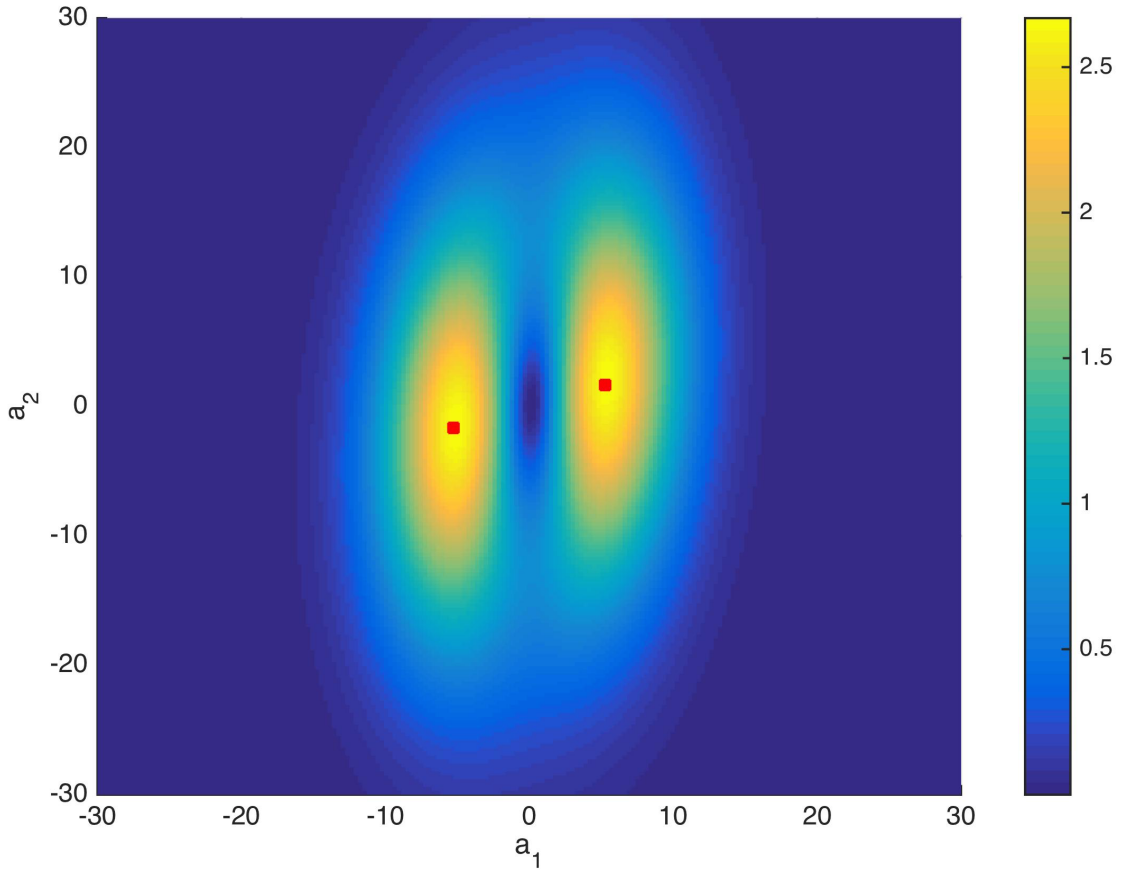


Figure 1: Utility function for 30 bus IEEE test system as a function of the attack vector where attacker 1 controls real power injection measurement 1 and attacker 2 controls real power injection measurement 2. The red squares show the location of the NE points.

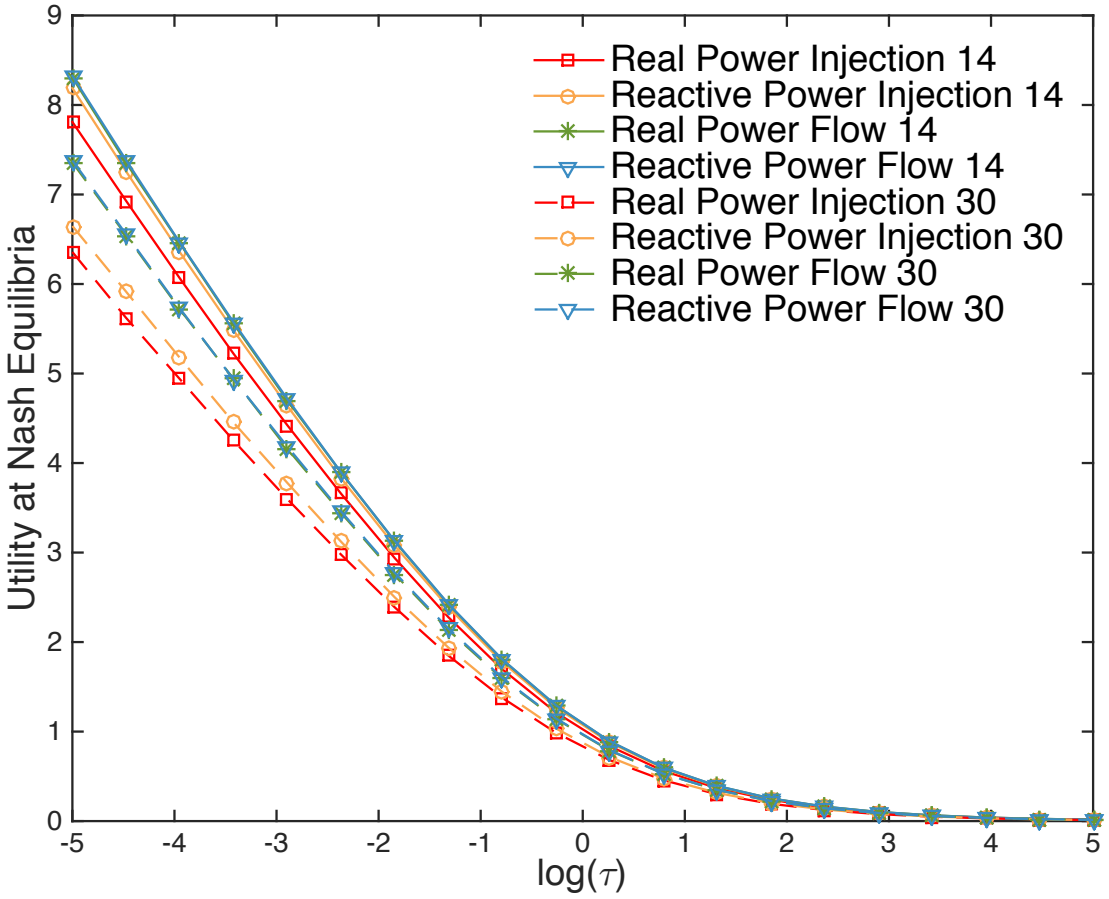


Figure 2: Utility at NE as a function of $\log \tau$ for different sets of measurement sensors. The solid lines correspond to the 14 bus IEEE test system and the dashed lines to the 30 bus IEEE test system.

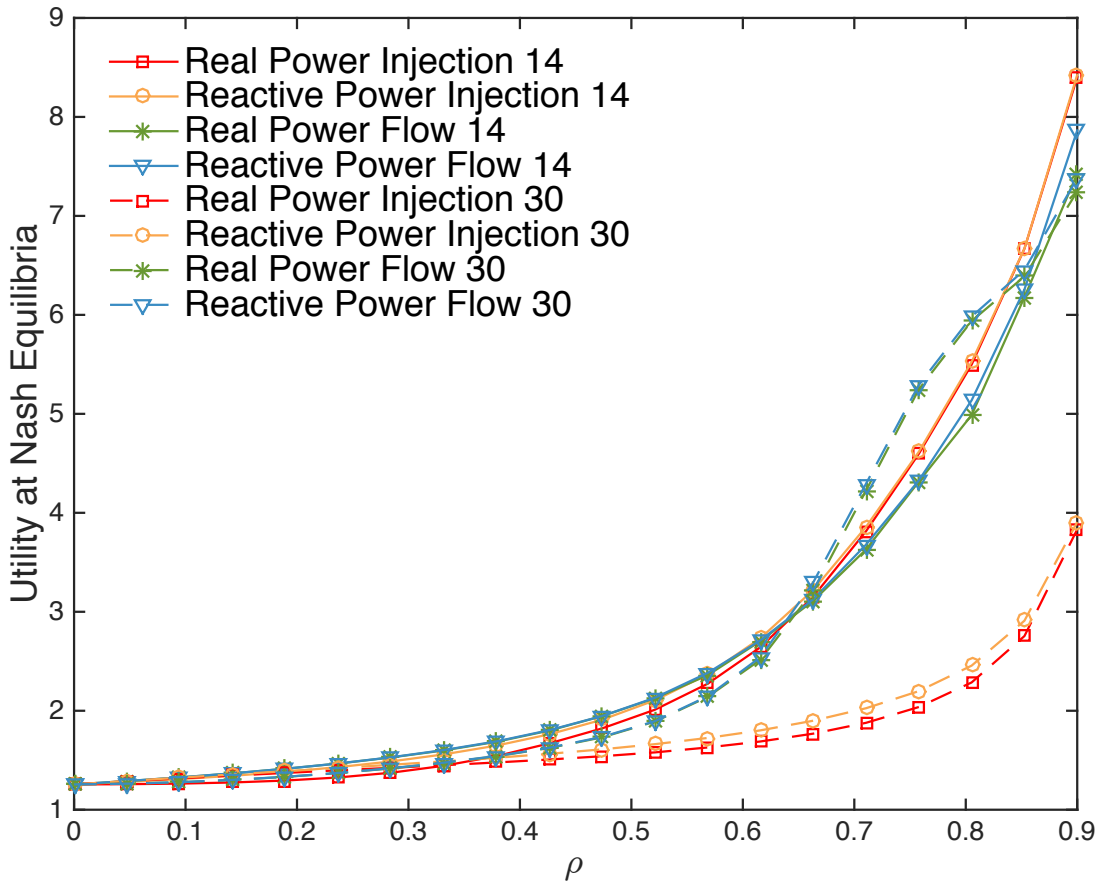


Figure 3: Utility at NE as a function of ρ for different sets of measurement sensors. The solid lines correspond to the 14 bus IEEE test system and the dashed lines to the 30 bus IEEE test system.

References

[1] E. Hossain, Z. Han, and H. Poor, *Smart grid communications and networking.* Cambridge University Press, 2012.

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

[3] 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.

[4] 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.

[5] A. Teixeira, S. Amin, H. Sandberg, K. Johansson, and S. Sastry, "Cyber security analysis of state estimators in electric power systems," in *Proc. IEEE Conference on Decision and Control*, Dec. 2010.

[6] 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.

[7] M. Ozay, I. Esnaola, F. Yarman Vural, S. Kulkarni, and H. Poor, "Machine learning methods for attack detection in the smart grid," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. PP, no. 99, pp. 1–1, 2015.

[8] 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 Processing Magazine*, vol. 29, no. 5, pp. 106–115, Sep. 2012.

[9] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, "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.

[10] M. Ozay, I. Esnaola, F. Vural, S. Kulkarni, and H. 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.

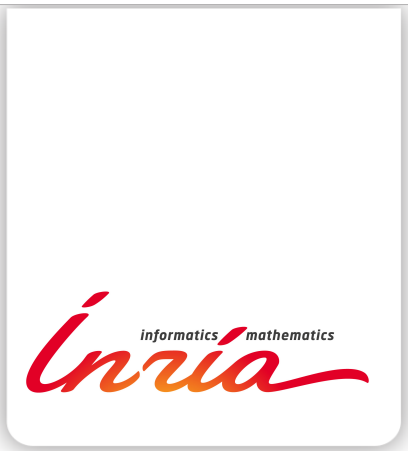
[11] W. Saad, Z. Han, H. Poor, and T. Basar, "Game-theoretic methods for the smart grid: An overview of microgrid systems, demand-side management, and smart grid communications," *IEEE Signal Process. Mag.*, vol. 29, no. 5, pp. 86–105, Sep. 2012.

[12] 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.

[13] 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.

[14] D. Monderer and L. Shapley, "Potential games," *Games and Economic Behavior*, vol. 14, no. 1, pp. 124–143, May 1996.

[15] R. M. Gray, "Toeplitz and circulant matrices: A review," *Foundations and TrendsÂ® in Communications and Information Theory*, vol. 2, no. 3, pp. 155–239, 2006. [Online]. Available: http://dx.doi.org/10.1561/0100000006



RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-0803