# Malicious Data Attacks on Smart Grid State Estimation: Attack Strategies and Countermeasures

Oliver Kosut, Liyan Jia, Robert J. Thomas, and Lang Tong
School of Electrical and Computer Engineering
Cornell University, Ithaca, NY 14853
Email: {oek2,lj92,rjt1,lt35}@cornell.edu

***Abstract*— The problem of constructing malicious data attack of smart grid state estimation is considered together with countermeasures that detect the presence of such attacks. For the adversary, using a graph theoretic approach, an efficient algorithm with polynomial-time complexity is obtained for the design of unobservable malicious data attacks. When the unobservable attack does not exist due to restrictions of meter access, attacks are constructed to minimize the residue energy of attack while guaranteeing a certain level of increase of mean square error. For the control center, a computationally efficient algorithm is derived to detect and localize attacks using the generalized likelihood ratio test regularized by an $L_1$ norm penalty on the strength of attack.**

## I. Introduction

Future smart grid will likely to be more tightly integrated with the cyber infrastructure for sensing, control, scheduling, dispatch, and billing. Already utilities company are rely on computer networks to manage generation and facilitate two way communications between users and suppliers. While such integration are essential for the grid to be smart, it also makes this vital physical infrastructure more vulnerable to cyber-attacks by adversaries around the globe. It has already been widely reported that the United State electrical grid has been penetrated by cyber spies, and there has been report that an experimental cyber attack launched by researchers caused a generator self-destruct [1]

The nature of attacks on smart grids can be very different from that on communication networks such as the Internet. The objective of an adversary may not be just gaining unauthorized information; an adversary could in theory cripple the power grid by attacking the energy management system (EMS) which collects data from remote meters and produces estimates of system states at the intervals of roughly 15 minutes. If an adversary is able to hack into the power grid and generates fake meter data, the energy management system at the control center may be misled by the state estimator, potentially making erroneous decisions on contingency analysis, dispatch, or even billing.

We consider in this paper strategies of covert attack by adversaries on meters of the smart grid by injecting malicious data with the goal of biasing power system state estimation. If successful, such attacks may mislead the control center to take erroneous actions that are detrimental to the network, or at the minimum, make the control center distrust state estimation.

Also considered in this paper are counter measures to malicious data attack at the control center in the form of attack detection. The problem of detecting malicious data attack can be viewed as a form of classical bad data detection. It is however important to note that, because the adversary can choose the site of attack judiciously and design attack data carefully, it is far more difficult to detect malicious data attacks than to detect random errors in the power systems. We will examine attacks with different degrees of sophistication.

### A. Summary of Results and Contributions

It was first discovered in [2] that in some cases, it is possible for the adversary to arbitrarily perturb the state estimator without being detected by the any bad data detector. We will discuss in Sec. III the close relationship between these highly damaging attacks and the classical notion of power system observability [3]. As such, we refer to these attacks as *unobservable* attacks. There are two primary regimes in which malicious data attacks occur, depending on whether the adversary controls enough meters to execute this unobservable attack. The two regimes have quite different behavior, and we present results in both.

In the case that the adversary may perform an unobservable attack, it is important to know how susceptible a power system is to such an attack. In particular, we are interested in the smallest number of meters that must be compromised by the adversary in order to perform such an attack. In Sec. III, we present an efficient algorithm to find small sets of meters that, if controlled by the adversary, could cause an unobservable attack. The algorithm is based on the purely topological conditions for observability developed in [4]. As such, it is graph-theoretic in nature and uses techniques of submodular function minimization [5], [6], [7]. Our algorithm can help learn how vulnerable a power system is to such an attack, and where the attack might take place.

We also investigate the worst malicious data attacks in the regime that the adversary cannot perform an unobservable attack. We develop a heuristic that allows us to obtain attacks that minimize attack power leakage to the detector while increasing the mean square error at the state estimator beyond a predetermined objective. This heuristic reduces to an eigenvalue problem that can be solved off line.

It is obviously important to develop countermeasures to these malicious attacks. To that end, we study detectors that can identify these attacks if they take place. In Sec. IV, we

present a decision theoretic formulation of detecting malicious data injection by an adversary. Because the adversary can choose where to attack the network and design the injected data, the problem of detecting malicious data cannot be formulated as a simple hypothesis test, and the uniformly most powerful test does not exist in general. We study a detector based on the generalized likelihood ratio test (GLRT) that was originally introduced in our prior work [8]. GLRT is not optimal in general, but it is known to perform well in practice and it has well established asymptotic optimality [9], [10], [11]. In other words, if the detector has many data samples, the detection performance of GLRT is close to optimal.

When guarding against attacks with large numbers of meters, it is infeasible to use the GLRT itself, because it requires searching over an exponentially large number of possible attacks. Therefore, we extend our prior work in [8] to develop a more practical technique based on $l_1$ norm minimization. This is based on the well-known nature of the $l_1$ norm as a heuristic to find sparse solutions to optimization problems.

Finally, in Sec. V, we conduct numerical simulations on a small scale example using the IEEE 14 bus network. We compare the GLRT detector with two classical detection schemes: the $J(\hat{\mathbf{x}})$ detector and the (Bayesian) largest normalized residue (LNR) detector [12], [13].

### *B. Related Work*

The first paper that addresses cyber-attack on power system state estimation appears to be [2], which inspired the work presented here. Lin, Ning, and Reiter consider the problem of malicious data attack under a deterministic model of network state variables and arbitrary attack patterns. They obtain a simple condition under which the attack we refer to as the unobservable attack exists, in which case the attack may increase the state estimation error arbitrarily. The authors of [2] also found that for many standard networks, the undetectable conditions are easily met if the adversary can control only a limited number of meters.

Another relevant recent work is by Gorinevsky, Boyd, and Poll [15] where a quadratic programming formulation for estimating faults is presented. The main difference between the approach in [15] and ours is that the formulation in [15] has the interpretation that the attack vector has the Laplacian prior and the state variables deterministic whereas, in this paper, the state vector is Gaussian and attack vector deterministic.

Bad data detection is a classical problem that is part of the original formulation of state estimation [12]. See [16] for an earlier comparison study. Malicious data attack can be viewed as the *worst interacting bad data* injected by an adversary. To this end, very little is known about the worst case scenario although the detection of interacting bad data has been considered [13], [17], [18], [19].

## II. Problem Formulation

A power system is composed of a collection of busses, transmission lines, and power flow meters. We adopt a graph-theoretic model for such a system. Therefore the power system is modeled as an undirected graph $(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ represents the set of busses, and $\mathcal{E}$ is the set of transmission lines. Each line connects two meters, so each element $e \in \mathcal{E}$ is an unordered pair of busses in $\mathcal{V}$. Fig 1 shows the graph structure of the IEEE 14-bus test system, which we use in our simulations. The control center receives measurements from various meters deployed throughout the system, from which it performs state estimation. Meters come in two varieties: transmission line flow meters, which measure the power flow through a single transmission line, and bus injection meters, which measure the total outgoing flow on all transmission lines connected to a single bus. Therefore each meter is associated with either a bus in $\mathcal{V}$ or a line in $\mathcal{E}$. We allow for the possibility of multiple meters on the same bus or line. Indeed, in our simulations, we assume that a meter is placed in every bus, and two meters on every line, one in each direction.

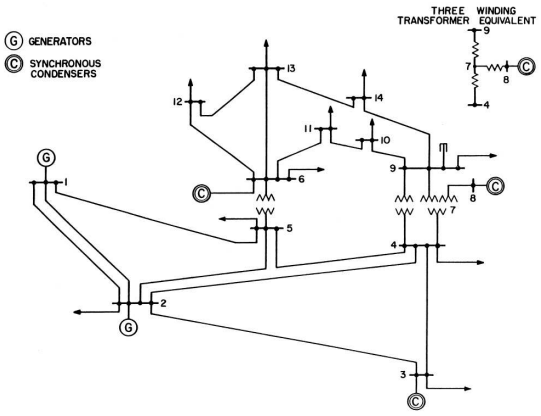


Fig. 1. IEEE 14 bus test system.

The graph-theoretic model for the power system yields the following DC power flow model, a linearized version of the AC power flow model:

$$\begin{aligned} \mathbf{z} &= \mathbf{H}\mathbf{x} + \mathbf{a} + \mathbf{e} \\ &\quad \mathbf{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_e), \\ &\quad \mathbf{a} \in \mathcal{A}_k = \{\mathbf{a} \in \mathbb{R}^m : \|\mathbf{a}\|_0 \leq k\} \end{aligned} \tag{1}$$

where $\mathbf{z} \in \mathbb{R}^m$ is the vector power flow measurements, $\mathbf{x} \in \mathbb{R}^n$ the system state, $\mathbf{e}$ the Gaussian measurement noise with zero mean and covariance matrix $\mathbf{\Sigma}_e$, and vector $\mathbf{a}$ is malicious data injected by an adversary. Here we assume that the adversary can at most control $k$ meters, *i.e.,* $\mathbf{a}$ is a vector with at most $k$ non-zero entries ($\|\mathbf{a}\|_0 \leq k$). A vector $\mathbf{a}$ is said to have sparsity $k$ if $\|\mathbf{a}\|_0 = k$. The $\mathbf{H}$ matrix in (1) arises from the graph theoretic model as follows. For each transmission line $(b_1, b_2) \in \mathcal{E}$, the DC power flow through this line from bus $b_1$ to bus $b_2$ is given by

$$\begin{bmatrix} 0 \cdots 0 & \underbrace{Y_{(b_1,b_2)}}_{b_1\text{th element}} & 0 \cdots 0 & \underbrace{-Y_{(b_1,b_2)}}_{b_2\text{th element}} & 0 \cdots 0 \end{bmatrix} \mathbf{x} \tag{2}$$

where $A_{(b_1,b_2)}$ is the susceptance of the transmission line $(b_1, b_2)$. Let $h_{(b_1,b_2)}$ be the row vector in (2). If a meter measures the flow through the transmission line connecting busses $b_1$ and $b_2$, then the associated row of $\mathbf{H}$ is given by $h_{(b_1,b_2)}$. If a meter measures the power injection for bus $b_1$,

then the associated row of $\mathbf{H}$ is given by

$$\sum_{b_2:(b_1,b_2)\in\mathcal{E}} h_{(b_1,b_2)}. \tag{3}$$

### A. A Bayesian Framework and MMSE Estimation

We consider in this paper a Bayesian framework where the state variables are random vectors with Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$. We assume that, in practice, the mean $\boldsymbol{\mu}_x$ and covariance $\boldsymbol{\Sigma}_x$ can be estimated from historical data. By subtracting the mean from the data, we can assume without loss of generality that $\boldsymbol{\mu}_x = \mathbf{0}$.

In the absence of an attack, *i.e.,* $\mathbf{a} = \mathbf{0}$ in (1), $(\mathbf{z}, \mathbf{x})$ are jointly Gaussian. The minimum mean square error (MMSE) estimator of the state vector $\mathbf{x}$ is a linear estimator given by

$$\hat{\mathbf{x}}(\mathbf{z}) = \underset{\hat{\mathbf{x}}}{\operatorname{argmin}}\, \mathbb{E}(\|\mathbf{x} - \hat{\mathbf{x}}(\mathbf{z})\|^2) = \mathbf{K}\mathbf{z} \tag{4}$$

where

$$\mathbf{K} = \boldsymbol{\Sigma}_x \mathbf{H}^{\mathrm{T}} (\mathbf{H}\boldsymbol{\Sigma}_x \mathbf{H}^{\mathrm{T}} + \boldsymbol{\Sigma}_e)^{-1}. \tag{5}$$

The minimum mean square error, in the absence of attack, is given by

$$\mathcal{E}_0 = \min_{\hat{\mathbf{x}}} \mathbb{E}(||\mathbf{x} - \hat{\mathbf{x}}(\mathbf{z})||^2) = \operatorname{Tr}\left(\boldsymbol{\Sigma}_x - \mathbf{K}_x \mathbf{H}\boldsymbol{\Sigma}_x\right).$$

If an adversary injects malicious data $\mathbf{a} \in \mathcal{A}_k$ but the control center is unaware of it, then the state estimator defined in (4) is no longer the true MMSE estimator (in the presence of attack); the estimator $\hat{\mathbf{x}} = \mathbf{K}\mathbf{z}$ is a "naive" MMSE estimator that ignores the possibility of attack, and it will incur a higher mean square error (MSE). In particular, it is not hard to see that the MSE in the presence of $\mathbf{a}$ is given by

$$\mathcal{E}_0 + \|\mathbf{K}\mathbf{a}\|_2^2. \tag{6}$$

The impact on the estimator from a particular attack $\mathbf{a}$ is given by the second term in (6). To increase the MSE at the state estimator, the adversary necessarily has to increase the "energy" of attack, which increases the probability of being detected at the control center.

## III. Strategies for Malicious Attacks

### A. Unobservable Attacks

Liu, Ning and Reiter observe in [2] that if there exists a nonzero $k$-sparse $\mathbf{a}$ for which $\mathbf{a} = \mathbf{H}\mathbf{c}$ for some $\mathbf{c}$, then

$$\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{a} + \mathbf{e} = \mathbf{H}(\mathbf{x} + \mathbf{c}) + \mathbf{e}.$$

Thus as a deterministic quantity, $\mathbf{x}$ is observationally equivalent to $\mathbf{x}+\mathbf{c}$. Therefore, if both $\mathbf{x}$ and $\mathbf{x}+\mathbf{c}$ are valid network states, the adversary's injection of data $\mathbf{a}$ when the true state is $\mathbf{x}$ will lead the control center to believe that the true network state is $\mathbf{x}+\mathbf{c}$, and vector $\mathbf{c}$ can be scaled arbitrarily. Since no detector can distinguish $\mathbf{x}$ from $\mathbf{x}+\mathbf{c}$, we call hereafter an attack vector $\mathbf{a}$ *unobservable* if it has the form $\mathbf{a} = \mathbf{H}\mathbf{c}$.

Note that it is unlikely that random bad data $\mathbf{a}$ will satisfy $\mathbf{a} = \mathbf{H}\mathbf{c}$. But an adversary can synthesize their attack vector to satisfy the unobservable condition. The following theorem provides insight into the adversary action by showing that an unobservable attack is closely related to the classical network observability conditions [3].

*Theorem 1:* A $k$-sparse attack vector $\mathbf{a}$ comprises an unobservable attack if and only if the network becomes unobservable when the $k$ meters associated with the nonzero entries of $\mathbf{a}$ are removed from the network; *i.e.,* the $(m-k)\times n$ submatrix of $\mathbf{H}$ taken from the rows of $\mathbf{H}$ corresponding to the zero entries of $\mathbf{a}$ does not have full column rank.

*Proof:* Without loss of generality, let $\mathbf{H}$ be partitioned into $\mathbf{H}^{\mathrm{T}} = [\mathbf{H}_1^{\mathrm{T}} \mid \mathbf{H}_2^{\mathrm{T}}]$, and submatrix $\mathbf{H}_1$ does not have full column rank, *i.e.,* there exists a vector $\mathbf{c} \neq \mathbf{0}$ such that $\mathbf{H}_1\mathbf{c}=\mathbf{0}$. We now have $\mathbf{a} = \mathbf{H}\mathbf{c} \in \mathcal{A}_k$, which is unobservable by definition. Conversely, consider an unobservable $\mathbf{a} = \mathbf{H}\mathbf{c} \in \mathcal{A}_k$. Without loss of generality, we can assume that the first $m-k$ entries of $\mathbf{a}$ are zero. We therefore have $\mathbf{H}_1\mathbf{c} = \mathbf{0}$ where $\mathbf{H}_1$ is the submatrix made of the first $m-k$ rows of $\mathbf{H}$. □

The implication from the above theorem is that the attack discovered in [2] is equivalent to removing $k$ meters from the network thus making the network not observable.

### B. Graph-Theoretic Approach to Minimum Size Unobservable Attacks

To know how susceptible a power system is to this highly damaging unobservable attack, it is important to know how few meters that must be controlled by the adversary before one can be performed. From Theorem 1, we know that there is an unobservable $k$-sparse attack vector $a$ if and only if it is possible to remove $k$ rows from $\mathbf{H}$ and cause $\mathbf{H}$ not to have full column rank. Finding the minimum such $k$ for an arbitrary $\mathbf{H}$ is a hard problem. However, it becomes easier given the extra structure on $\mathbf{H}$ imposed by the network topology.

We now give a simple method to find sets of meters whose removal make the system unobservable. Moreover, we show that it is possible to efficiently minimize the size of the set of meters produced by this method; thereby one may efficiently compute small sets of meters from which an adversary may execute an unobservable attack.

For a set of lines $\mathcal{A} \subseteq \mathcal{E}$, let $g(\mathcal{A})$ be the set of meters either on lines in $\mathcal{A}$ or on busses adjacent to lines in $\mathcal{A}$. Let $h(\mathcal{A})$ be the number of connected components in the graph $(\mathcal{V}, \mathcal{E} \setminus \mathcal{A})$; *i.e.,* the original graph after all lines in $\mathcal{A}$ have been removed. The following theorem gives a simple method for determining a number of meters in $g(\mathcal{A})$ to remove from the network to make it unobservable. The proof relies on [4], which gave an efficient method to determine the observability of a network based only on its topology.

*Theorem 2 (Sufficient condition for unobservable attacks):* For all $\mathcal{A} \subseteq \mathcal{E}$, removing an arbitrary subset of $g(\mathcal{A})$ of size $|g(\mathcal{A})| - h(\mathcal{A}) + 2$ makes the system unobservable.

*Proof:* Let $\bar{\mathcal{V}}$ and $\bar{\mathcal{E}}$ be the sets of busses and lines respectively with a meter placed on them. Theorem 5 in [4] states that the power system given by $(\mathcal{V}, \mathcal{E}, \bar{\mathcal{V}}, \bar{\mathcal{E}})$ is observable if and only if there exists a $\mathcal{F} \subseteq \mathcal{E}$ comprising a spanning tree of $\mathcal{V}$ and an assignment function

$$\phi : \mathcal{F} \to \bar{\mathcal{V}} \cup \bar{\mathcal{E}} \tag{7}$$

satisfying:

1) If $l \in \bar{\mathcal{E}}$, then $\phi(l) = l$.
2) If $\phi(l) \in \bar{\mathcal{V}}$, then line $l$ is incident to the bus $\phi(l)$.
3) If $l_1, l_2 \in \mathcal{F}$ are distinct, then $\phi(l_1) \neq \phi(l_2)$.

The principle behind this theorem is that a bus injection meter may "impersonate" a single line meter on a line incident to the bus. If a bus $b = \phi(l)$ for some line $l$, this represents the meter at $b$ impersonating a meter on line $l$. The system is observable if and only if a spanning tree $\mathcal{F}$ exists made up of transmission lines with either real meters or impersonated meters by bus meters.

Not including the lines in $\mathcal{A}$, the network splits into $h(\mathcal{A})$ separate pieces. Therefore, any spanning tree $\mathcal{F}$ must include at least $h(\mathcal{A}) - 1$ lines in $\mathcal{A}$. Any assignment $\phi$ satisfying the conditions above must therefore employ at least $h(\mathcal{A}) - 1$ meters in $g(\mathcal{A})$. Hence, if any $|g(\mathcal{A})| - h(\mathcal{A}) + 2$ of these meters are removed from the network, only $h(\mathcal{A}) - 2$ remain, which is not enough to create a full spanning tree, so the network becomes unobservable. □

*Example 1:* Consider the IEEE 14-bus test system, shown in Fig. 1. Take $\mathcal{A} = \{(7, 8)\}$. Since bus 8 is only connected to the system through bus 7, removing this line from the network cuts it into two pieces. Therefore $h(\mathcal{A}) = 2$. The set of meters $g(\mathcal{A})$ consists of meters on the line $(7, 8)$, and bus injection meters at bus 7 and 8. Theorem 2 states that if we remove $|g(\mathcal{A})|$ meters from this set—that is, all the meters in $g(\mathcal{A})$—the system becomes unobservable. In our simulation examples, we assume there are two meters on each line, therefore it takes 4 meters to execute an unobservable attack. Furthermore, it is not hard to employ Theorem 2 to find similar 4-sparse unobservable attacks on the 30-bus, 118-bus, and 300-bus test systems.

Theorem 2 provides a method to find unobservable attacks, but we would like to find attacks using as few meters as possible. We use the theory of submodular functions to show that the quantity $|g(\mathcal{A})| - h(\mathcal{A}) + 2$ can be efficiently minimized over all sets of edges $\mathcal{A}$. This significantly increases the usefulness of Theorem 2, because it means we can efficiently find small unobservable attacks for arbitrary power systems.

A submodular function is a real-valued function $f$ defined on the collection of subsets of a set $W$ such that for any $A, B \subseteq W$,

$$f(A \cup B) + f(A \cap B) \leq f(A) + f(B). \tag{8}$$

Moreover, a function $f$ is supermodular if $-f$ is submodular. There are several known techniques to find the set $A \subseteq W$ minimizing $f(A)$ in time polynomial in the size of $W$ [5], [6], [7]. It is not hard to see that $|g(\mathcal{A})|$ is submodular in $\mathcal{A}$, and $h(\mathcal{A})$ is supermodular. Therefore, their difference is submodular, so it can be efficiently minimized.

### C. Minimum Residue Energy Attack

We now consider the problem of finding the worst attack in the regime that the adversary cannot perform an unobservable attack. In this regime, it is not possible to select an $\mathbf{a}$ vector that will never be detected by the control center. The best choice for the adversary will be to select an attack vector that is particularly damaging to the control center's state estimate without being easily detectable. There will thus be a trade-off for the adversary between cause large errors in the state estimate and being detected with low probability. We introduce a method based on a residue energy heuristic to approach this trade-off.

Given the naive MMSE state estimator $\hat{\mathbf{x}} = \mathbf{K}\mathbf{z}$ (4-5), the estimation residue error is given by

$$\mathbf{r} = \mathbf{G}\mathbf{z}, \quad \mathbf{G} \triangleq \mathbf{I} - \mathbf{H}\mathbf{K} \tag{9}$$

Substituting the measurement model, we have

$$\mathbf{r} = \mathbf{G}\mathbf{H}\mathbf{x} + \mathbf{G}\mathbf{a} + \mathbf{G}\mathbf{e}.$$

where $\mathbf{G}\mathbf{a}$ is the only term from the attack. Therefore, an attack vector $\mathbf{a}$ will be more difficult to detect at the control center if $\mathbf{G}\mathbf{a}$ is small. Recall from (6), the damage in MSE done by injecting $\mathbf{a}$ is $\|\mathbf{K}\mathbf{a}\|_2^2$. We therefore consider the following equivalent problems:

$$\max_{\mathbf{a} \in \mathcal{A}_k} \|\mathbf{K}\mathbf{a}\|_2^2 \quad \text{subject to} \quad \|\mathbf{G}\mathbf{a}\|_2^2 \leq \eta, \tag{10}$$

or equivalently,

$$\min_{\mathbf{a} \in \mathcal{A}_k} \|\mathbf{G}\mathbf{a}\|_2^2 \quad \text{subject to} \quad \|\mathbf{K}\mathbf{a}\|_2^2 \geq C. \tag{11}$$

The above optimizations remain difficult due to the constraint $\mathbf{a} \in \mathcal{A}_k$. However, given a specific sparsity pattern $\mathcal{S} \subset \{1, \cdots, n\}$ for which $a_i = 0$ for all $i \notin \mathcal{S}$, solving the optimal attack vector $\mathbf{a}$ for the above two formulations is a standard generalized eignevalue problem.

In particular, for fixed sparsity pattern $\mathcal{S}$, let $\mathbf{a}_\mathcal{S}$ be the nonzero subvector of $\mathbf{a}$, $\mathbf{K}_\mathcal{S}$ the corresponding submatrix of $\mathbf{K}$, and $\mathbf{G}_\mathcal{S}$ similarly defined. The problem (11) becomes

$$\min_{\mathbf{u} \in \mathbb{R}^{n-k}} \|\mathbf{G}_\mathcal{S}\mathbf{u}\|_2^2 \quad \text{subject to} \quad \|\mathbf{K}_\mathcal{S}\mathbf{u}\|_2^2 \geq C. \tag{12}$$

Let $\mathbf{Q}_G \triangleq \mathbf{G}_\mathcal{S}^\mathrm{T}\mathbf{G}_\mathcal{S}$, $\mathbf{Q}_K \triangleq \mathbf{K}_\mathcal{S}^\mathrm{T}\mathbf{K}_\mathcal{S}$. It can be shown that the optimal attack pattern has the form

$$\mathbf{a}_\mathcal{S}^* = \sqrt{\frac{C}{\|\mathbf{K}_\mathcal{S}\mathbf{v}\|_2^2}}\mathbf{v} \tag{13}$$

where $\mathbf{v}$ is the generalized eigenvector corresponding to the smallest generalized eigenvalue $\lambda_{\min}$ of the following matrix pencil

$$\mathbf{Q}_G\mathbf{v} - \lambda_{\min}\mathbf{Q}_K\mathbf{v} = \mathbf{0}.$$

The $k$ dimensional symmetrical generalized eigenvalue problem can be solved the QZ algorithm [21].

## IV. Detection of Malicious Data Attack

### A. Statistical Model and Attack Hypotheses

We now present a formulation of the detection problem at the control center. We assume a Bayesian model where the state variables are random with a multivariate Gaussian distribution $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_x)$. Our detection model, on the other hand, is not Bayesian in the sense that we do not assume any prior probability of the attack nor do we assume any statistical model for the attack vector $\mathbf{a}$.

Under the observation model (1), we consider the following composite binary hypothesis:

$$\mathcal{H}_0 : \mathbf{a} = \mathbf{0} \quad \text{versus} \quad \mathcal{H}_1 : \mathbf{a} \in \mathcal{A}_k \setminus \{\mathbf{0}\}. \tag{14}$$

Given observation $\mathbf{z} \in \mathbb{R}^m$, we wish to design a detector $\delta : \mathbb{R}^m \to \{0, 1\}$ with $\delta(\mathbf{z}) = 1$ indicating a detection of attack ($\mathcal{H}_1$) and $\delta(\mathbf{z}) = 0$ the null hypothesis.

### *B. Generalized Likelihood Ratio Detector with $L_1$ Norm Regularization*

For the hypotheses test given in (14), the uniformly most powerful test does not exist. We propose a detector based on the generalized likelihood ratio test (GLRT). We note in particular that, if we have multiple measurements under the same $\mathbf{a}$, the GLRT proposed here is asymptotically optimal in the sense that it offers the fastest decay rate of miss detection probability [20].

The distribution of the measurement $\mathbf{z}$ under the two hypotheses differ only in their means

$$\begin{aligned} \mathcal{H}_0 &: \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_z) \\ \mathcal{H}_1 &: \mathbf{z} \sim \mathcal{N}(\mathbf{a}, \mathbf{\Sigma}_z), \mathbf{a} \in \mathcal{A}_k \setminus \{\mathbf{0}\} \end{aligned}$$

where $\mathbf{\Sigma}_z \triangleq \mathbf{H}\mathbf{\Sigma}_x\mathbf{H}^{\mathrm{T}} + \mathbf{\Sigma}_e$. The GLRT is given by

$$L(\mathbf{z}) \triangleq \frac{\max\limits_{\mathbf{a}\in\mathcal{A}_k} f(\mathbf{z}|\mathbf{a})}{f(\mathbf{z}|\mathbf{a}=\mathbf{0})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \tau, \tag{15}$$

where $f(\mathbf{z}|\mathbf{a})$ be the Gaussian density function with mean $\mathbf{a}$ and covariance $\mathbf{\Sigma}_z$, and the threshold $\tau$ is chosen from under null hypothesis for a certain false alarm rate. This is equivalent to

$$\min_{\mathbf{a}\in\mathcal{A}_k} \mathbf{a}^{\mathrm{T}}\mathbf{\Sigma}_z^{-1}\mathbf{a} - 2\mathbf{z}^{\mathrm{T}}\mathbf{\Sigma}_z^{-1}\mathbf{a} \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\gtrless}} \tau. \tag{16}$$

Thus the GLRT reduces to solving

$$\begin{aligned} \text{minimize} \quad & \mathbf{a}^{\mathrm{T}}\mathbf{\Sigma}_z^{-1}\mathbf{a} - 2\mathbf{z}^{\mathrm{T}}\mathbf{\Sigma}_z^{-1}\mathbf{a} \\ \text{subject to} \quad & \|\mathbf{a}\|_0 \le k. \end{aligned} \tag{17}$$

For a fixed sparsity pattern, *i.e.,* if we know the support but not necessarily the actual values of $\mathbf{a}$, the above optimization is easy to solve. In other words, if we know a small set of suspect meters from which malicious may be injected, the above test is easily computable. The sparsity condition on $\mathbf{a}$ makes the above optimization problem non-convex, but for small $k$ it can be solved exactly simply by exhaustively searching through all sparsity patterns. For larger $k$, this is not feasible. It is a well known technique that (17) can be approximated by a convex optimization:

$$\begin{aligned} \text{minimize} \quad & \mathbf{a}^{\mathrm{T}}\mathbf{\Sigma}_z^{-1}\mathbf{a} - 2\mathbf{z}^{\mathrm{T}}\mathbf{\Sigma}_z^{-1}\mathbf{a} \\ \text{subject to} \quad & \|\mathbf{a}\|_1 \le \nu \end{aligned} \tag{18}$$

where the $L_1$ norm constraint is a heuristic for the sparsity of $\mathbf{a}$. The constant $\nu$ needs to be adjusted until the solution involves an $\mathbf{a}$ with sparsity $k$. This requires solving (18) several times.

## V. Numerical Simulations

We present some simulation results on the IEEE 14 bus system shown in Fig. 1 to compare the performance of the GLRT with the $J(\hat{x})$ test and the LNR test [12], [13]. For various sparsity levels, we find the minimum energy residue attack as discussed in Sec. III-C. The adversary may then scale this attack vector depending on how much it wishes to influence the mean square error. We make two plots to show performance of the detectors. The first is the standard *Receiver Operating Characteristics* (ROC) that characterize the tradeoff between the probability of attack detection vs. the probability of false alarm, which we may plot for a single attack vector. In addition, we plot the *Attacker Operating Characteristic* (AOC), which was introduced in [8], and characterizes the tradeoff between the probability of being detected vs. resulting (extra) mean-square error at the state estimator. In particular, we fix a probability of false alarm and vary the length of the attack vector along the direction minimizing the energy residue. This plot illustrates the trade-off faced by the adversary between increasing the state estimation error and minimizing its probability of detection. In our simulations, we characterize the mean square error increase at the control center using the ratio between the resulting MSE from the attack and the MSE under no attack (*i.e.,* $\mathbf{a} = 0$) in dB.

Fig. 2 shows the ROC and AOC curves for the worst-case 2-sparse attack. We implement the GLRT using exhaustive search over all possible sparsity patterns. This is feasible because of the low sparsity level, so we need not resort to the $L_1$ minimization as in (18). Observe that the GLRT performs consistently better than the other two conventional detectors.

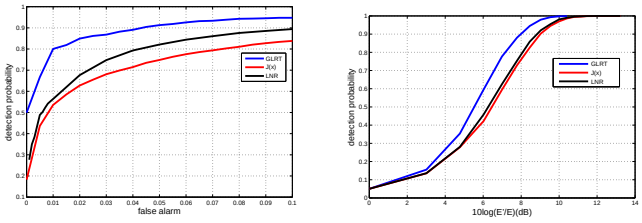


Fig. 2. Left: ROC Performance of GLRT for 2 sparsity case. MSE with attack is 8db. SNR=10db. Right: AOC Performance of GLRT for 2 sparsity case. False alarm rate is 0.05. SNR=10dB. The horizontal axis

Fig. 3 shows the ROC and AOC curves for the worst-case 3-sparse attack, again using exhaustive search for the GLRT. Interestingly, the LNR test outperforms the GLRT at this sparsity level. We believe the reason for this is that the GLRT has little recourse when there is significant uncertainty in the sparsity pattern of the attack. In particular, the meters being controlled by the adversary here are the bus injection meter at bus 1, and the two meters on the transmission line between bus 1 and 2. These constitute three of the seven meters that hold any information about the state at bus 1. Thus, it may be difficult for the detector to determine which of the several meters around bus 1 are the true adversarial meters. The GLRT does not react to this uncertainty: it can only choose the most likely sparsity pattern, which is often wrong. Indeed, in our simulations the GLRT identified the correct sparsity pattern

only 4.2% of the time.

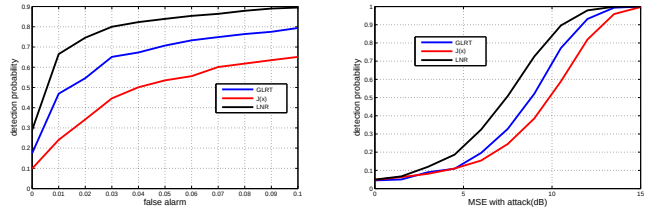


Fig. 3. Left: ROC Performance of GLRT for 3 sparsity case. MSE with attack is 10db. SNR=10db. Right: AOC Performance of GLRT for 3 sparsity case. False alarm rate is 0.05. SNR=10dB

Continuing our analysis of 3-sparsity attacks, we conduct simulations when the adversaries are placed randomly in the network, instead of at the worst-case meters. Once their random meters are chosen, we find the worst-case attack vector using the energy residual heuristic. This simulates the situation that the adversaries cannot choose their locations, but are intelligent and cooperative in their attack. The resulting performance of the three detectors is shown in Fig. 4. Observe that we have recovered the outperformance of the GLRT as compared to the conventional detectors, if only slightly. When the placement of the adversaries is random, they are not as capable of cooperating with one another, therefore their attack is easier to detect.

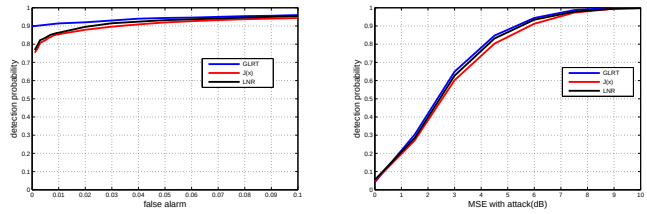


Fig. 4. Left: ROC Performance of GLRT under random attack for 3 sparsity case. MSE with attack is 6db. SNR=10db. Right: AOC Performance of GLRT under random attack for 3 sparsity case. False alarm rate is 0.05. SNR=10dB

Finally, we increase the sparsity level to 6, at which it is impossible to perform exhaustive search for the GLRT. At this sparsity level, it becomes possible to perform an unobservable attack, so it is not as illuminating to choose the worst-case sparsity pattern, as that would be very difficult to detect. Instead, we again choose the sparsity pattern randomly but optimize the attack within it. Fig. 5 compares the performance of the GLRT implemented via $l_1$ minimization as in (18) to the two conventional detectors. Note again that the GLRT outperforms the others.

## VI. Conclusions

We present in this paper adversarial strategies for malicious data attacks, as well as countermeasures for the control center. We present a polynomial-time algorithm to find small but highly damaging unobservable attacks, and, for the case that this is impossible, we discussed the minimum residue energy heuristic to find the worst attacks. We also studied the generalized likelihood ratio test as a detector for this problem; in particular, this detector was implement using convex optimization via $L_1$ norm regularization.

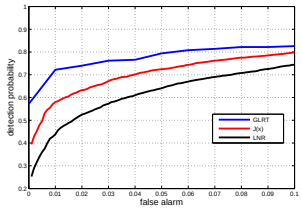


Fig. 5. ROC Performance of GLRT under random attack for 6 sparsity case. MSE with attack is 6db. SNR=10db.

## References


[1] J. Meserve, "Sources: Staged cyber attack reveals vulnerability in power grid," *CNN*, Sept 26, 2007.

[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*, pp. 21–32, 2009.

[3] A. Monticelli and F. Wu, "Network observability: Theory," *IEEE Trans. Power Apparatus and Systems*, vol. PAS-104, pp. 1042–1048, May 1985.

[4] G. R. Krumpholz, K. A. Clements, and P. W. Davis, "Power system observability: a practical algorithm using network topology," *IEEE Trans. Power Apparatus and Systems*, vol. PAS-99, no. 4, pp. 1534–1542, July 1980.

[5] M. Grötschel, L. Lovász, A. Schrijver, "The ellipsoid method and its consequences in combinatorial optimization", *Combinatorica*, vol. 1, no. 2, pp. 169–197, June 1981.

[6] W. H. Cunningham, "On submodular function minimization," *Combinatorica*, vol. 5, no. 3, pp. 185–192, Sep. 1985.

[7] A. Schrijver, A combinatorial algorithm minimizing submodular functions in strongly polynomial time," *Journal of Combinatorial Theory Series B*, vol .80, no .2, pp .346–355, Nov. 2000.

[8] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, "On false data attacks on power system state estimation," to appear in *Proc. 45th International Universities' Power Engineering Conference*, Aug/Sep 2010.

[9] O. Zeitouni, J. Ziv, and N. Merhav, "When si the generalized likelihood ratio test optimal," *IEEE Tran. Information Theory*, vol. 38, pp. 1597–1602, Mar. 1991.

[10] B. C. Levy, *Principles of Signal Detection and Parameter Estimation*. New York, NY: Springer, 2008.

[11] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications, 2nd ed.* NY: Springer, 1998.

[12] F. C. Schweppe, J. Wildes, and D. P. Rom, "Power system static state estimation, Parts I, II, III," *IEEE Tran. on Power Appar. & Syst.*, vol. PAS-89, pp. 120–135, 1970.

[13] E. Handschin, F. C. Schweppe, J. Kohlas, and A. Fiechter, "Bad data analysis for power system state estimation," *IEEE Trans. Power Apparatus and Systems*, vol. PAS-94, pp. 329–337, Mar/Apr 1975.

[14] F. F. Wu and W. E. Liu, "Detection of topology errors by state estimation," *IEEE Trans. Power Systems*, vol. 4, pp. 176–183, Feb 1989.

[15] D. Gorinevsky, S. Boyd, and S. Poll, "Estimation of faults in DC electrical power systems," in *Proc. 2009 American Control Conf.*, (St. Louis, MO.), pp. 4334–4339, June 2009.

[16] L. Mili, T. V. Cutsem, and M. Ribbens-Pavalla, "Bad data identification methods in power system state estimation—A comparative study," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 953–972, August 1998.

[17] A. Monticelli, F. Wu, and M. Yen, "Multiple bad data identification for state estimation by combinatorial optimization," *IEEE Trans. Power Systems*, vol. PWRD-1, pp. 361–369, July 1986.

[18] M. G. Cheniae, L. Mili, and P. Rousseeuw, "Identification of multiple interacting bad data via power system decomposition," *IEEE Trans. Power Systems*, vol. 11, pp. 1555–1563, August 1996.

[19] A. Abur and A. G. Expósito, *Power System State Estimation: Theory and Implementation*. CRC, 2000.

[20] S. Kourouklis, "A large deviation result for the likelihood ratio statistic in exponential families," *The Annals of Statistics*, vol. 12, no. 4, pp. 1510–1521, 1984.

[21] G. Golub and C. V. Loan, *Matrix Computations*. Baltimore, Maryland: The Johns Hopkins University Press, 1990.