On Malicious Data Attacks on Power System State Estimation

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 detecting and characterizing impacts of malicious attacks against smart grid state estimation is considered. Different from the classical bad data detection for state estimation, the detection of malicious data injected by an adversary must take into account carefully designed attacks capable of evading conventional bad data detection. A Bayesian framework is presented for the characterization of fundamental tradeoffs at the control center and for the adversary. For the control center, a detector based on the generalized likelihood ratio test (GRLT) is introduced and compared with conventional bad detection detection schemes. For the adversary, the tradeoff between increasing the mean square error (MSE) of the state estimation vs. the probability of being detected by the control center is characterized. A heuristic is presented for the design of worst attack.**

***Index Terms*—Energy management systems, State estimation, smart grid security, false data attack.**

I. Introduction

The electric grid in the United States has evolved over the past century from a series of small independent community-based systems to perhaps the largest and most complex cyber-physical System in the world. The increasing reliance on cyber-infrastructure to manage highly complex smart grids comes with the risk of cyber-attacks by adversaries around the globe. It has been widely reported recently that "cyberspies have already penetrated the United States electrical grid and left software programs that can be used to disrupt the system " [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 the impact of malicious data attack on smart grid state estimation and counter measures against such attacks. We make a distinction between the conventional "bad data" due to natural causes (meter malfunction, communication outage, and topological errors) from malicious data injected by an adversary. The latter may be carefully designed to maximize the impact of attack and evade detection.

While the problem of bad data detection has been well studied for decades, the potential damage of malicious data attack has only be investigated recently. In a recent paper by Liu, Ning, and Reiter [2], the authors obtain conditions under which the adversary can arbitrarily perturb the state estimator without being detected by the conventional bad data detection. The argument presented in [2] can in fact be made even stronger: if an adversary can control enough meters and if there is no prior distributions on network states, no detector will ever be able to detect a carefully designed malicious data attack.

A. Summary of Results and Contributions

We aim to quantify the impact of a malicious data attack on power system state estimation and develop counter measures. To this end, we consider two aspects of the overall problem: (i) attack detection and localization strategies at the control center; (ii) attack strategies by the adversary.

We first 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 power test does not exist in general. We proposed a detector based on the generalized likelihood ratio test (GLRT). GLRT is not optimal in general, but it is known to perform well in practice and it has well established asymptotic optimality [3], [4], [5]. In other words, if the detector has many data samples, the detection performance of GLRT is close to optimal.

We note that the proposed detector has a different structure from those used in conventional bad data detectors which usually employ a test on the state estimator residues errors [6], [7], [8]. The proposed the GLRT detector does not compute explicitly the residue error. We show, however, that when there is at most one attacked meter (a single attacked data), the GLRT is identical to the classical largest normalized residue (LNR) test using the residue error from the minimum mean square error (MMSE) state estimator, which is the Bayesian counter part of the LNR test. The asymptotic optimality of GLRT lends a stronger theoretic basis for the LNR test for the single bad data test. The proposed detector also estimates

(identifies) the specific meters from which the attack may be originated.

Next we investigate malicious data attack from the perspective of an adversary who must make a tradeoff between inflicting the maximum damage on state estimation and being detected by the EMS at the control center. We define the notion of *Attacker Operating Characteristic* (AOC) that characterizes the tradeoff between the probability of being detected vs. resulting (extra) mean-square error at the state estimator. We therefore formulate the problem of optimal attack as minimizing the probability of being detected subject to causing the mean square error (MSE) to increase beyond a predetermined level. Finding the attack with the optimal AOC is intractable, unfortunately. We present a heuristic that allows us to obtain attacks that with minimum 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.

Finally, we conduct numerical simulations on a small scale example using the IEEE 14 bus network. For the control center, we present simulation results that compare different detection schemes based on the *Receiver operating Characteristics* (ROC) that characterize the tradeoff between the probability of attack detection vs. the probability of false alarm. We show that there is a substantial difference between the problem of detecting randomly appearing bad data from detecting malicious data injected by an adversary. For example, at the detection error probability of 0.9, the malicious data attack results in close to 5dB increase of MSE. See Fig 3. Next we compare the GLRT detector with two classical detection schemes: the $J(\hat{\mathbf{x}})$ detector and the (Bayesian) largest normalized residue (LNR) detector [6], [7]. Our test shows consistent improvement over the two well established detection schemes. See Figure 2. From the adversary perspective, we compare the *Attacker Operating Characteristics* (AOC). Our result shows again that the GLRT detector gives higher probability of detection than that those of conventional detectors for the same amount MSE increase at the state estimator. See Figure 2.

B. Related Work

The first paper that addresses cyber-attack on power system state estimation appears to be [2], which inspires 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 that malicious data are "undetectable" and the attack may increase the state estimation error arbitrarily. We show in Sec II that the undetectable condition obtained in [2] is equivalent to the classical network observability condition [9], [10]. 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.

The "undetectability" of certain malicious data injection shown in [2] motivates a Bayesian formulation first considered in [11] and further elaborated in this paper. Since state estimation is performed every few minutes, there is a wealth of historical data that can be used to characterize state distributions under normal operation conditions. The knowledge of such "prior" therefore limits the ability of adversary to perturb the network state arbitrarily. Under such a formulation, attacks considered in [2] are no longer "undetectable". The detector presented in [11] uses L_∞ norm on residue errors from the state estimator and can be considered as the Bayesian modification of the LRN test. Note that the idea of exploiting state variable distributions has been considered earlier by Lourenço, Costa, and Clements in [12] in the context of topology error identification.

There are several differences between the work in [2] and that in this paper. First, the work in [2] focuses on the conditions of "undetectable attacks", treating network states as deterministic and unknown quantities. When the undetectability condition is not satisfied, very little is known about how to detect attacks and the effect of adversary on state estimation. We present in this paper a detector that does not depend on the undetectability condition of [2], thanks to a Bayesian formulation of the state estimation [11]. The present paper extends the results in [11] in several directions. We provide a more general formulation of the problem, a new detector for attacks, and a characterization of the tradeoff for the adversary between probability of being detected and the extra mean square error introduced at the state estimator.

Another relevant recent work is by Gorinevsky, Boyd, and Poll [13] where a quadratic programming formulation for estimating faults is presented. The main difference between the approach in [13] and ours is that the formulation in [13] 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 [6]. See [14] 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 [7], [15], [16], [17].

II. Problem Formulation

Consider a DC power flow state estimation problem based on a linearized 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 $\|\mathbf{a}\|_0 = k$.

A. Detectability and Observability

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 the insight into the adversary action by connecting unobservable attack with the classical network observability conditions [9].

Theorem 1: There exists an unobservable k-sparse attack vector $\mathbf{a}$ if and only if the network becomes unobservable when some k measurements are removed, *i.e.,* there exists an $(m-k)\times n$ submatrix of $\mathbf{H}$ that 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. Indeed, with Theorem 1, it is not too difficult to select a set of meters to perturb the state estimator arbitrarily in the subspace of the unobservable states.

B. A Bayesian Framework and MMSE Estimation

If an attack $\mathbf{a}$ is unobservable, can it be detected? Merrill and Schweppe wrote in [18] that "If the system is controlled by a human being, watching, perhaps, a pen recorder, he will automatically ignore a large random spike that is obviously incorrect." In other words, if there is a prior distribution on the states under normal conditions, the perturbation $\mathbf{c}$ introduced by the adversary through $\mathbf{a}=\mathbf{H}\mathbf{c}$ will deviate the distribution from its prior, thus making the attack detectable. Mathematically, this calls for a Bayesian formulation that captures the statistical behavior of network states.

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

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

where

$$\mathbf{K}=\boldsymbol{\Sigma}_x\mathbf{H}^{\mathrm{T}}(\mathbf{H}\boldsymbol{\Sigma}_x\mathbf{H}^{\mathrm{T}}+\boldsymbol{\Sigma}_e)^{-1}. \tag{3}$$

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

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

Note that in the high signal to noise ratio (SNR) regime, the MMSE estimator is closely approximated by the conventional weighted least squares (WLS) estimator.

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 (2) 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, the MSE in the presence of $\mathbf{a}$ is given by

$$\begin{aligned}\mathbb{E}\|\mathbf{x}-\mathbf{K}\mathbf{z}\|_2^2 &= \mathbb{E}\|\mathbf{x}-\mathbf{K}(\mathbf{H}\mathbf{z}+\mathbf{e}+\mathbf{a})\|_2^2\\ &= \mathrm{Tr}\big[(\mathbf{I}-\mathbf{K}\mathbf{H})\boldsymbol{\Sigma}_x(\mathbf{I}-\mathbf{K}\mathbf{H})^{\mathrm{T}}\\ &\quad +\mathbf{K}\boldsymbol{\Sigma}_e\mathbf{K}^{\mathrm{T}}+\mathbf{K}\mathbf{a}\mathbf{a}^{\mathrm{T}}\mathbf{K}^{\mathrm{T}}\big]\\ &= \mathcal{E}_0+\|\mathbf{K}\mathbf{a}\|_2^2.\end{aligned} \tag{4}$$

The impact on the estimator from a particular attack $\mathbf{a}$ is given by the second term in (4). 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. 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 $\mathbf{x}\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}_x)$. 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:

$$\mathcal{H}_0:\mathbf{a}=\mathbf{0}\quad\text{versus}\quad\mathcal{H}_1:\mathbf{a}\in\mathcal{A}_k\setminus\{\mathbf{0}\}. \tag{5}$$

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.

An alternative formulation, one we will not pursue here, is based on the extra MSE $\|\mathbf{K}\mathbf{a}\|_2^2$ at the state estimator. See (4). In particular, we may want to distinguish, for $\|\mathbf{a}\|_0\leq k$,

$$\mathcal{H}_0':\|\mathbf{K}\mathbf{a}\|_2^2\leq C,\quad\text{versus}\quad\mathcal{H}_1':\|\mathbf{K}\mathbf{a}\|_2^2>C.$$

Here both null and alternative hypotheses are composite and the problem is more complicated. The operational interpretation, however, is significant because one may not care in practice about small attacks that only marginally increase the MSE of the state estimator.

B. Generalized Likelihood Ratio Detector

For the hypotheses test given in (5), 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 [19].

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 $\mathbf{\Sigma}_z \triangleq \mathbf{H}\mathbf{\Sigma}_x\mathbf{H}^{\mathrm{T}} + \mathbf{\Sigma}_e$. Let $f(\mathbf{z}|\mathbf{a})$ be the Gaussian density function with mean $\mathbf{a}$ and covariance $\mathbf{\Sigma}_z$,

$$f(\mathbf{z}|\mathbf{a}) = \frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}_z|}\exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{a})^{\mathrm{T}}\mathbf{\Sigma}_z^{-1}(\mathbf{z}-\mathbf{a})\right\}. \tag{6}$$

The GLRT is given by

$$L(\mathbf{z}) \triangleq \frac{\max\limits_{\mathbf{a}\in\mathcal{A}_k} f(\mathbf{z}|\mathbf{a})}{f(\mathbf{z}|\mathbf{a}=\mathbf{0})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \tau, \tag{7}$$

which is equivalent to

$$\min_{\mathbf{a}\in\mathcal{A}_k} \mathbf{a}^{\mathrm{T}}\mathbf{\Sigma}_z^{-1}\mathbf{a} - 2\mathbf{z}^{\mathrm{T}}\mathbf{\Sigma}_z^{-1}\mathbf{a} \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\gtrless}} \tau, \tag{8}$$

where the threshold τ is chosen from under null hypothesis for a certain false alarm rate. Thus the GLRT reduces to solving

$$\begin{array}{ll}\text{minimize} & \mathbf{a}^{\mathrm{T}}\mathbf{\Sigma}_z^{-1}\mathbf{a} - 2\mathbf{z}^{\mathrm{T}}\mathbf{\Sigma}_z^{-1}\mathbf{a}\\ \text{subject to} & \|\mathbf{a}\|_0 \le k.\end{array} \tag{9}$$

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. In general, the sparsity condition on $\mathbf{a}$ makes this optimization problem non-convex and difficult to solve. It is a well known technique that the above problem can be approximated by a convex optimization:

$$\begin{array}{ll}\text{minimize} & \mathbf{a}^{\mathrm{T}}\mathbf{\Sigma}_z^{-1}\mathbf{a} - 2\mathbf{z}^{\mathrm{T}}\mathbf{\Sigma}_z^{-1}\mathbf{a}\\ \text{subject to} & \|\mathbf{a}\|_1 \le \nu\end{array} \tag{10}$$

where the L_1 norm constraint is a heuristic for the sparsity of $\mathbf{a}$. The constant ν needs to be adjusted until the solution involves an $\mathbf{a}$ with sparsity k. This requires solving (10) several times.

C. Classical Detectors with MMSE State Estimation

We will compare the performance of the GLRT detector with two classical bad data detectors [6], [7], both based on the residual error $\mathbf{r} = \mathbf{z} - \mathbf{H}\hat{\mathbf{x}}$ resulted from the MMSE state estimator.

The first is the $J(\hat{\mathbf{x}})$ detector, given by

$$\mathbf{r}^{\mathrm{T}}\mathbf{\Sigma}_e^{-1}\mathbf{r} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \tau. \tag{11}$$

The second is the LRN test given by

$$\max_i \frac{|r_i|}{\sigma_{r_i}} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\lesseqgtr}} \tau, \tag{12}$$

where σ_{r_i} is the standard deviation of the ith residual error r_i. We may regard this is a test on the l_∞-norm of the measurement residual, normalized so that each element has unit variance.

The asymptotic optimality of the GLRT detector implies a better performance of GLRT over the above two detectors when the sample size is large. For the finite sample case (one shot in particular), numerical simulations shown in Sec V confirm that the GLRT detector improves the performance of the $J(\hat{\mathbf{x}})$ and LNR detectors. The interesting exception is the case when only one meter is under attack, *i.e.,* $\|\mathbf{a}\|_0 = 1$ and $\mathbf{\Sigma}_e = \sigma_e^2\mathbf{I}$. In this case, the GLRT turns out to be identical to the LNR detector. Therefore, the GLRT can be viewed as a generalization of the LNR detector, in that it can be tuned to any sparsity level. Moreover, this provides some theoretical justification for the LNR detector. The proof that establishes the equivalence between the GLRT and LNR for the 1-sparsity attack is an algebraic exercise omitted here due to space limitations. See Appendix for a proof.

IV. Attack Operating Characteristics and Optimal Attacks

We now study the impact of malicious data attack from the perspective of an attacker. We assume that the attacker knows the (MMSE) state estimator and the (GLRT) detector used by the control center. We also assume that the attacker can choose k meters arbitrarily in which to inject malicious data. In practice, however, the attacker may be much more limitted. Thus our results here are perhaps more pessimistic than in reality.

A. AOC and Optimal Attack Formulations

The attacker faces two conflicting objectives: maximizing the MSE by choosing the best data injection $\mathbf{a}$ vs. avoiding being detected by the control center. The tradeoff between increasing MSE of the state estimator and lower the probability of detection is characterized by *attacker operating characteristics* (AOC), analogous to the receiver operating characteristics (ROC) at the control center. Specifically, AOC is the probability of detection of the detector $\Pr(\delta(\mathbf{z}) = 1 \mid \mathbf{a})$ as a function of the extra MSE $\mathcal{E}(\mathbf{a}) = \mathcal{E}_0 + \|\mathbf{K}\mathbf{a}\|_2^2$ (4) at the state estimator, where $\mathcal{E}_0$ is the MMSE in the absence of attack.

The optimal attack in the sense of maximizing the MSE while limiting the probability of detection can be formulated as the following constrained optimization

$$\max_{\mathbf{a}\in\mathcal{A}_k} \|\mathbf{K}\mathbf{a}\|_2^2 \quad \text{subject to} \quad \Pr(\delta(\mathbf{z}) = 1|\mathbf{a}) \le \beta, \tag{13}$$

or equivalently,

$$\min_{\mathbf{a}\in\mathcal{A}_k} \Pr(\delta(\mathbf{z}) = 1|\mathbf{a}) \quad \text{subject to} \quad \|\mathbf{K}\mathbf{a}\|_2^2 \le C. \tag{14}$$

The design of optimal attack for the above is difficult in general. Here we propose a heuristic for $\Pr(\delta(\mathbf{z}) = 1|\mathbf{a})$, which will allow us to rewrite the above optimization in a way that is easier to solve.

B. The Minimum Residue Energy Attack

The difficulty of obtaining optimal attack defined in (13,14) is the lack of analytical expressions for the detection error probability $\Pr(\delta(\mathbf{z}) = 1|\mathbf{a})$. We present here an alternative approach using a residue energy heuristic.

Given the naive MMSE state estimator $\hat{\mathbf{x}} = \mathbf{Kz}$ (2-3), the estimation residue error is given by

$$\mathbf{r} = \mathbf{Gz}, \quad \mathbf{G} \triangleq \mathbf{I} - \mathbf{HK} \tag{15}$$

Substituting the measurement model, we have

$$\mathbf{r} = \mathbf{GHx} + \mathbf{Ga} + \mathbf{Ge}.$$

where $\mathbf{Ga}$ is the only term from the attack. Therefore, instead of working with the probability of attack detection, we can cast the problem of optimal attack as maximizing MSE while minimizing the residue energy of the attack. Specifically, we consider the following equivalent problems:

$$\max_{\mathbf{a}\in\mathcal{A}_k} \|\mathbf{Ka}\|_2^2 \quad \text{subject to} \quad \|\mathbf{Ga}\|_2^2 \leq \eta, \tag{16}$$

or equivalently,

$$\min_{\mathbf{a}\in\mathcal{A}_k} \|\mathbf{Ga}\|_2^2 \quad \text{subject to} \quad \|\mathbf{Ka}\|_2^2 \geq C. \tag{17}$$

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 (17) becomes

$$\min_{\mathbf{u}\in\mathbb{R}^{n-k}} \|\mathbf{G}_\mathcal{S}\mathbf{u}\|_2^2 \quad \text{subject to} \quad \|\mathbf{K}_\mathcal{S}\mathbf{u}\|_2^2 \geq C. \tag{18}$$

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

$$\mathbf{a}_\mathcal{S}^* = \sqrt{\frac{C}{\|\mathbf{K}_\mathcal{S}\mathbf{v}\|_2^2}}\mathbf{v} \tag{19}$$

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 [20].

V. Numerical Simulations

We present some simulation results on the IEEE 14 bus system as illustrated in Figure 1 where we have used Bus 1 as the reference. For the linearized DC model, there are 13 state variables and 54 measurements. In our simulation, the state variables are Gaussian with covariance $\mathbf{\Sigma}_x = \sigma_x^2 I$, and the measurement errors are given by $\mathbf{\Sigma}_e = \sigma_e^2$. The SNR in dB is defined as $\text{SNR} \triangleq 10\log\frac{\sigma_x^2}{\sigma_e^2}$.

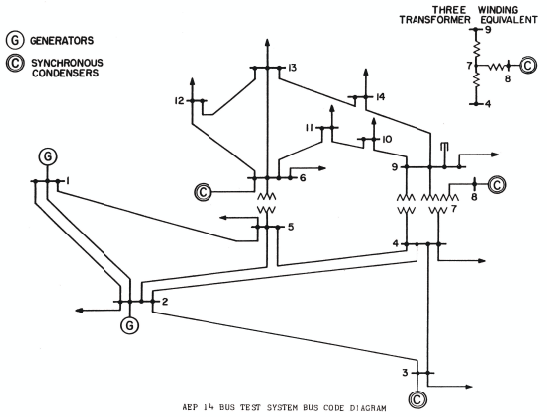

Fig. 1. IEEE 14 bus system.

A. Receiver Operating Characteristic

Figure 2 (left) shows ROC curves for the GLRT, $J(\hat{\mathbf{x}})$ detector, and the LNR test. In our simulations, under $\mathcal{H}_1$, we assumed that the attacker applied the minimum energy residue attack that increases the MSE at the state estimator to different levels. We observed that GLRT performed consistently better than the other two conventional detectors, and the gain is the largest for the more practically significant false alarm rate regions (0.001-0.1).

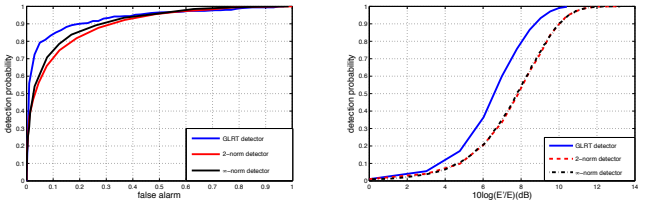

Fig. 2. Left: ROC Performance of GLRT. MSE increase is [**?**]db. SNR=10db. Right: AOC Performance of GLRT. False alarm rate is X. SNR=10dB

B. Attacker Operating Characteristic

Figure 2 (right) shows the result of AOC for the three detectors under the minimum residue energy attack. The results were consistent with the ROC at the state estimator. Again GLRT performed better than the other two detectors across the region of all MSE increments. It is interesting to note that, with detection probability 90%, the MSE increase by the attack is 9dB, which means the MSE is almost 10 times as much as the MSE without attack when the adversary only has access to 2 meters (optimized for the attack).

C. Random bad data vs. malicious data

We now demonstrate the difference between (conventional) bad data and malicious data. Here we assume the best detector (GLRT) at EMS and compare the AOC curves when two bad data are placed randomly in the network. The values of the bad data, however, are optimized to minimize the detection probability. In other words, the injected bad data have the most damaging value, but their locations are chosen randomly.

Figure 3 shows the AOC curves for the two scenarios. We observed pronounced decrease in detection probability

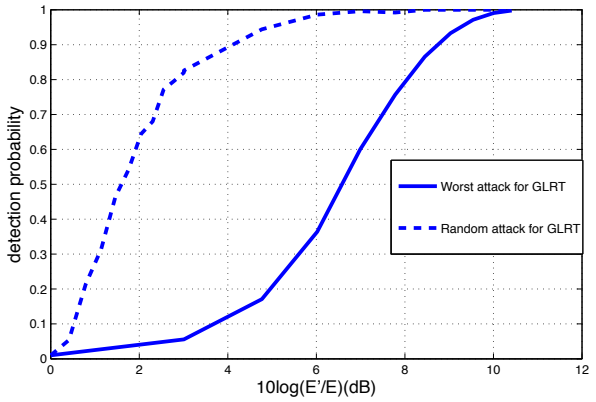

Fig. 3. Random bad data vs. Malicious data for GLRT. False alarm rate is 0.01. SNR = [**?**].

when the data change from bad to malicious. When viewed at the same detection probability level, the malicious data can increase the MSE significantly.

VI. Conclusions

We present in this paper an analytical framework to evaluate the impact and develop counter measures of malicious data attack by an adversary capable of selecting a set of meters to fabricate data. Our results show that there is a significant difference between malicious data attack and the conventional bad data problem in power system state estimation. We propose a new detector based on the principle of generalized likelihood ratio test. We introduce the attacker operating characteristic as a measure of optimality of adversary attacks and propose a heuristic based on the minimization of residue energy of the attack subject to a level guaranteed increase of mean square error at the control center.

There are a number of issues not addressed in this paper are being investigated currently. For example, there is a need to develop computationally efficient algorithms for the GLRT detector and the design of adaptive optimal minimum residue energy attack. The incorporation of PMU is another new component that is being studied and will be reported in the future.

Appendix

We establish here that the equivalence of GLRT and LNR when $\|\mathbf{a}\|_0 = 1$. If $k = 1$, then the left hand side of (8) becomes

$$\min_i \min_{a_i} (\mathbf{\Sigma}_z^{-1})_{ii} a_i^2 - 2\mathbf{z}^{\mathrm{T}} (\mathbf{\Sigma}_z^{-1})_i a_i \tag{20}$$

where $(\mathbf{\Sigma}_z^{-1})_{ii}$ is the ith diagonal element of $\mathbf{\Sigma}_z^{-1}$, and $(\Sigma_z^{-1})_i$ is the ith row of $\mathbf{\Sigma}_z^{-1}$. The second minimization can be solved in closed form, so (20) becomes

$$-\max_i \frac{[\mathbf{z}^{\mathrm{T}} (\Sigma_z^{-1})_i]^2}{(\Sigma_z^{-1})_{ii}}. \tag{21}$$

We may therefore write the GLRT as

$$\max_i \frac{|\mathbf{z}^{\mathrm{T}} (\mathbf{\Sigma}_z^{-1})_i|}{\sqrt{(\mathbf{\Sigma}_z^{-1})_{ii}}} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \tau. \tag{22}$$

The vector of numerators in (22) is given by $\mathbf{r}' = \mathbf{\Sigma}_z^{-1}\mathbf{z}$. Note that the covariance matrix of $\mathbf{r}'$ is simply Σ_z^{-1}. Therefore we may regard (22) as a test on the maximum element of the $\mathbf{r}'$ after each element is normalized to unit variance.

We now show that $\mathbf{r}'$ is just a constant multiple of $\mathbf{r}$, meaning that (22) is identical to (12), saving a constant factor. Recall that $\mathbf{r} = (\mathbf{I} - \mathbf{HK})\mathbf{z}$, where

$$\begin{aligned} \mathbf{I} - \mathbf{HK} &= I - \mathbf{H\Sigma}_x\mathbf{H}^{\mathrm{T}}(\mathbf{H\Sigma}_x\mathbf{H}^{\mathrm{T}} + \mathbf{\Sigma}_e)^{-1} \\ &= (\mathbf{H\Sigma}_x\mathbf{H}^{\mathrm{T}} + \mathbf{\Sigma}_e - \mathbf{H\Sigma}_x\mathbf{H}^{\mathrm{T}})(\mathbf{H\Sigma}_x\mathbf{H}^{\mathrm{T}} + \mathbf{\Sigma}_e)^{-1} \\ &= \mathbf{\Sigma}_e\mathbf{\Sigma}_z^{-1} = \sigma_e^2\mathbf{\Sigma}_z^{-1}. \end{aligned}$$

Thus $\mathbf{r} = \sigma_e^2\mathbf{r}'$; the two detectors are identical.

References

[1] S. Gorman, "Electricity grid in U.S. penetrated by spies," April 8 2009. The Wall Street Journal.

[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] 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.

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

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

[6] 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.

[7] 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.

[8] A. Monticelli, "Electric power system state estimation," *Proceedings of the IEEE*, vol. 88, no. 2, pp. 262–282, 2000.

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

[10] 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.

[11] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, "Limiting false data attacks on power system state estimation," in *Proc. 2010 Conference on Information Sciences and Systems*, Mar 2010.

[12] E. M. Lourenço, A. S. Costa, and K. A. Clements, "Bayesian-based hypothesis testing for topology error identification in generalized state estimation," *IEEE Trans. Power Systems*, vol. 19, pp. 1206–1215, May 2004.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] H. M. Merrill and F. C. Schweppe, "Bad data suppression in power system state estimation," *IEEE Trans. Power Apparatus and Systems*, vol. PAS-90, pp. 2718–2725, 1971.

[19] 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.

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