Differentially Private Communication of Measurement Anomalies in the Smart Grid

Nikhil Ravi, *Graduate Student Member, IEEE*, Anna Scaglione, *Fellow, IEEE*, Sean Peisert, *Senior Member, IEEE*, Parth Pradhan, *Senior Member, IEEE*

Abstract—In this paper, we present a framework based on differential privacy (DP) for querying electric power measurements to detect system anomalies or bad data. Our DP approach conceals consumption and system matrix data, while simultaneously enabling an untrusted third party to test hypotheses of anomalies, such as the presence of bad data, by releasing a randomized sufficient statistic for hypothesis-testing. We consider a measurement model corrupted by Gaussian noise and a sparse noise vector representing the attack, and we observe that the optimal test statistic is a chi-square random variable. To detect possible attacks, we propose a novel DP chi-square noise mechanism that ensures the test does not reveal private information about power injections or the system matrix. The proposed framework provides a robust solution for detecting bad data while preserving the privacy of sensitive power system data.

Index Terms—bad data attacks, differential privacy, smart grids, energy internet, hypothesis testing

I. INTRODUCTION

POWER systems are pivotal in delivering electricity to various sectors, including residential, commercial, and industrial. The effective management of these systems is heavily reliant on the accurate and timely acquisition of data for operational, control, and monitoring purposes [1]–[4]. Such data is essential for numerous critical functions like load forecasting and security assessments. However, as power systems become increasingly digitized and interconnected, they are also more vulnerable to cyber threats. These include the malicious manipulation of measurement data, known as bad data (see fig. 1 for an illustration), which can lead to significant disruptions in operations, financial losses, and even endanger public safety [5]. Consequently, the detection of such malicious activities emerges as a crucial area of focus, underpinning the need for robust cybersecurity measures within the fragmented smart grid.

A. Motivation: Collective Defense

The management of power grids often suffers from fragmentation among multiple Regional Transmission Organizations
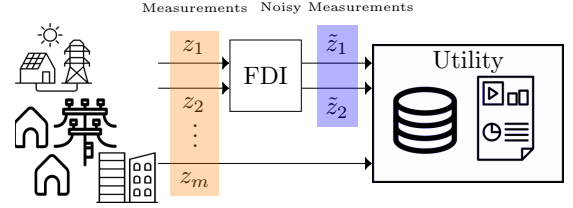
Fig. 1: Illustration of bad data attacks.

(RTOs) and utilities and a division between distribution and transmission, despite the interconnected nature of the grid. Efficient prevention and detection of cyber-physical attacks necessitate collaboration and information sharing among different utilities. The advent of Distributed Energy Resources (DERs) has further complicated the cybersecurity landscape of power systems. The decentralized nature and a mix of ownership between utilities and private stakeholders of DER-rich grids necessitates a collective approach to cyber-defense, emphasizing the importance of collaboration among various stakeholders. This *collective defense* strategy is foundational to ensuring the secure and efficient operation of interconnected systems and devices. Furthermore, this collaborative approach can significantly enhance early detection capabilities, improve understanding of attack methods, develop effective defense mechanisms, implement cost-effective solutions, and ensure regulatory compliance, ultimately bolstering the security and reliability of power systems. Anomalies in local data can serve as warning signs for issues that may have implications for neighboring systems [6], yet achieving such a collaborative cybersecurity framework in the fragmented grid management infrastructure is fraught with challenges, particularly in the realms of information sharing, technological solutions, collaborative partnerships, and coordinated incident response [7].

Information sharing is crucial for collective defense but is hindered by several factors, including concerns about data privacy related to customers, proprietary issues, and security [8]. Many grid operators are hesitant to share sensitive measurement data, such as Advanced Metering Infrastructure (AMI) and Phasor Measurement Units (PMU), further exacerbated by the fragmented landscape of grid devices and systems. This landscape now encompasses Home/Building Energy Management Systems, DER aggregators, and IoT devices, complicating the construction of a comprehensive and effective cyber-defense posture. Moreover, even sharing data with law enforcement agencies and regulators can encounter obstacles due to privacy considerations. This fragmentation

and the myriad of concerns lead to critical cybersecurity information remaining siloed or entirely unshared, significantly compromising the grid's overall security and underscoring the importance of safeguarding against unintended disclosure of private data within the power system sector.

Without a dedicated platform for real-time information exchange and automated decision-making, utilities struggle to identify new threats, learn from incidents in other territories, and coordinate responses to cyber-attacks. This lack of structured incident response, coupled with insufficient knowledge of protecting operational technology systems, leaves the grid vulnerable to sophisticated cyber-attacks. Recognizing these challenges, the United States government emphasizes the importance of technology software equipped with a collective defense capability that enables rapid sharing of insights and detections with the Federal government, participants, and other trusted (by the US government) organizations, thereby enhancing grid resilience against sophisticated cyber-attacks [9]. Moreover, the recent initiatives by the White House and the Department of Energy to fund research into innovative cyber-physical collective defense methodologies are timely and crucial. This funding aims to foster the development of defense strategies that are effective even in scenarios involving potentially untrustworthy third parties, thereby signaling a significant step towards bolstering the resilience and integrity of our energy infrastructure [10], [11].

In this context, differential privacy (DP) [12] mechanisms emerge as a promising solution to the challenges faced in implementing a collective defense strategy. By enabling the secure sharing of data among grid operators and with third-party entities, DP addresses key concerns around data privacy and proprietary information. DP mechanisms, by introducing controlled statistical "noise" to the data, protect sensitive information while maintaining its utility for analytical purposes, offering a balance between privacy and accuracy. This method surpasses traditional anonymization techniques by providing mathematical guarantees on the amount of information leakage, allowing for the optimization of queries relevant to the energy sector and the design of differentially private databases for analysis and research. Such practicality in applying DP to energy datasets fosters increased data sharing, enhancing stakeholder comfort and safeguarding privacy, trade secrets, and sensitive information. Before summarizing our contributions, next, we provide a brief review of the literature on anomaly detection, DP, and DP anomaly detection for smart grid applications.

B. Literature Review

1) Anomaly Detection in the Smart Grid: Efforts to combat bad data attacks on power systems have led to the development of a diverse array of detection algorithms. Data-driven strategies, such as those employing machine learning techniques like distributed Support Vector Machines for stability-focused detection [13], and real-time electricity theft detection [14], leverage large datasets to identify anomalies indicative of false data injection (FDI) attacks. Anomaly identification techniques utilizing Multiclass SVMs have shown efficiency

[15], though their computational demand limits broader application. Artificial neural networks and their extensions into deep learning have gained popularity for their high detection accuracy in identifying FDI attacks [16]–[21], yet they suffer from extensive training times. Attempts to mitigate these computational challenges have led to innovative solutions, such as the integration of artificial bee colony algorithms with differential evolution theory [22]. Despite the advantages of data-driven methods, their effectiveness is curtailed by the need for extensive, often centralized, datasets, leading to challenges in time, cost, and privacy. The reliance on large local datasets or detailed system information introduces substantial data transmission burdens, while the absence of effective privacy-preserving measures and the risks associated with centralized data processing highlight the need for a more collaborative detection approach.

2) DP for Smart Grids: The landscape of cybersecurity within power systems has seen various strategies aimed at preserving data privacy and enhancing system resilience. Traditional approaches such as access control and anonymization have been instrumental in initial efforts to protect sensitive data. Access control methods, as discussed in [23]–[25], attempt to regulate data access, often resulting in binary outcomes of either complete access or total restriction. Meanwhile, anonymization techniques, highlighted in [26], seek to obscure personal identifiers within datasets. Despite their intentions, these techniques have been criticized for their susceptibility to reidentification attacks, a vulnerability exposed by [27]. Regulatory attempts to navigate the privacy challenges inherent in electric grid data management, such as the "15/15 Rule" [28], aim to balance consumer privacy with the public's right to access data. However, the effectiveness of these policies has been questioned, with critiques pointing out a lack of scientific underpinning and insufficient privacy safeguards, as elaborated in [29].

In response to these challenges, DP has emerged as a robust alternative, particularly in the context of smart grids. DP's capacity to protect individual privacy while enabling aggregated data analysis offers a solution to the limitations of previous approaches. The survey by Ul Hassan et al. [30] provides a thorough overview of DP applications across cyber-physical systems, with a notable emphasis on smart grids. Research in DP for smart grids has primarily focused on three areas: grid demand response, smart building operations, and grid data collection with fog computing. In demand response, DP methods like data masking using Laplacian noise have been explored to protect consumer data without compromising utility operations [31]. For smart buildings, which are integral to urban development, DP is used to secure sensor data streams and Internet traffic, ensuring the privacy of inhabitants against potential intrusions [32]–[37]. DP Smart meter load monitoring has been studied in [38]. Integrating DP with fog computing has shown the potential to enhance privacy and operational efficiency in smart grids. This approach safeguards data during transmission and storage in fog nodes, protecting against privacy breaches without significant impacts on system performance [39], [40]. The employment of DP in demand-data reporting has been shown to preserve the

confidentiality of individual consumption patterns while still providing aggregated data useful for grid optimization [41]. Further applications of DP in clustering load profiles have enabled the grouping of similar consumption patterns without infringing on the privacy of individual users [29]. This same study illustrates how DP techniques can facilitate the sharing of load profiles for research purposes, effectively balancing the need for data utility with privacy considerations.

Despite progress, securing grid users' data remains a complex issue, with DP providing a promising path forward across various scenarios. However, areas like fault information transmission, load profiling, and billing information privacy still demand attention to achieve comprehensive privacy protection in smart grid applications.

3) DP Bad Data Detection in Smart Grids: A handful of studies have explored the application of DP for bad data and FDI attack detection within smart grids, addressing the critical balance between privacy protection, system security, and data utility. Hossain et al. [42] delve into the dual role of DP in smart grids, noting its capacity to safeguard user privacy while potentially enabling integrity attacks through privacy-preserving noise. They propose a tailored DP design strategy focused on mitigating the effects of FDI attacks. Their work extends to assessing the viability of DP in smart grid environments, especially under adversarial conditions, and evaluates the implications on the quality of service. Specifically, they analyze the sum query on a database of measurements from a μPMU dataset with the addition of simple Laplacian noise, from which they derive optimal strategies for both attack and defense scenarios. Gaboardi et al. [43] introduce a novel approach to conducting chi-squared tests for goodness of fit and independence that adhere to DP constraints. Their method is innovative in that it modifies classical statistical tests to incorporate DP mechanisms, thus ensuring the privacy of sensitive data. The study presents both Monte Carlo-based and asymptotically aligned tests that adjust for DP-induced noise, highlighting a methodological advancement in integrating privacy preservation within statistical analysis. Lin, et. al. [44] present a federated learning-based algorithm for distributed and privacy-preserving FDI attack detection that allows state owners to collaboratively generate a global detection model without extensive data transmission, thus protecting data privacy by integrating artificial Gaussian noise into the local model estimations.

C. Contributions

Building on recent advancements and addressing enduring challenges in cyber-defense, our work specifically focuses on enabling bad data detection by entities that may not be fully trusted, all while preserving the privacy of critical system data. This focus marks a distinct departure from existing approaches that primarily enhance the detection mechanisms of FDI attacks or apply DP in a general context. Our contributions, tailored to this unique challenge within the realm of power systems' security and privacy, include:

- A novel chi-squared noise DP mechanism that enhances privacy in querying grid measurements for detecting bad

data and anomalies. Applied to the norm of the residual error of the power systems' state estimate [5], [45], [46], this mechanism is versatile enough to be used for any quadratic queries following a chi-square distribution. This approach enables bad data detection by third parties without compromising the confidentiality of system states or matrices.
- An approximation of the chi-squared mechanism to a Gaussian mechanism for stochastic queries in large systems, optimizing the balance between privacy preservation and analytical utility.

In contrast to existing approaches detailed in Section I-B3, our methodology eschews the direct application of DP noise to measurements in favor of targeting the residual, as formalized in Section II-B. This strategy not only simplifies analytical processes but also demands lower privacy budgets to effectively protect system data. Furthermore, while prior studies consider deterministic and static measurement models, our framework innovatively accounts for the stochastic nature of measurements. By focusing on ensemble-based privacy rather than individual measurement instantiations, we offer a unique contribution to the DP landscape. This methodological innovation permits the precise tailoring of a chi-square DP mechanism for quadratic queries, marking a notable advancement in privacy-preserving techniques for power systems. Our approach underscores the utility of privacy-preserving techniques in supporting collective defense strategies, filling gaps in information sharing and technology, and promoting a more resilient grid against cyber threats in DER-rich environments. An illustration of our proposed mechanism is shown in fig. 2.
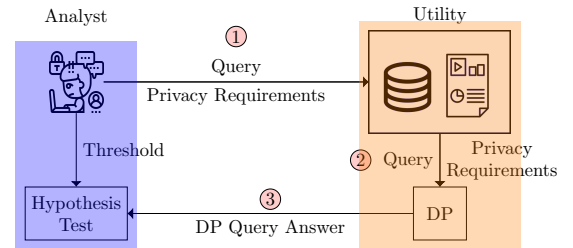
Fig. 2: Illustration of our proposed DP mechanism for FDA detection.

The remainder of the paper is organized as follows. Section II defines the measurement model, performs preliminary analysis on the least squares residual of the state estimation problem, and introduces the threat model. In Section III, we first introduce the definitions related to DP before presenting our novel DP mechanism and its Gaussian approximation for sharing the residuals with third parties. Section IV showcases the numerical results, and finally, Section V concludes the paper.

Notation: Boldfaced lowercase (uppercase, respectively) letters denote vectors (matrices, respectively), and x_i (X_{ij}, respectively) denotes the ith element of vector $\boldsymbol{x}$ (the ij^{th} entry of matrix $\boldsymbol{X}$, respectively). Calligraphic letters denote sets, and $|\cdot|$ represents the cardinality of a set. Furthermore, $[N]$ denotes the set of integers $0, 1, \ldots, N-1$.

II. PRELIMINARIES, THREAT MODEL AND PROBLEM STATEMENT

In power systems operations, state estimation algorithms are used to fit the observed measurements collected from the system and make informed decisions. State estimation algorithms need to be robust to a variety of errors arising from measurement errors, modeling errors, uncertainty in the model parameters, and bad (maliciously placed or otherwise) data. In this paper, we are motivated by the problem of bad data injection attacks on the observed measurements, although the method applies to detecting other anomalies. Traditionally, the analysis of residuals in state estimation has been utilized to detect the presence of bad data: data is considered "bad" if the error with respect to the model is higher than what is statistically consistent with the measurement noise. Incorrect estimation of system states can compromise operations and control decisions. In this section, we briefly introduce the measurement model and the residual-based bad data detection (BDD) algorithm.

A. Measurement Model and False Data Attack

The measurement model relates the observed measurements z to the system states x and the noise; for additive noise such models can be expressed as:

$$z = h(x_o) + \eta + a. \tag{1}$$

Here, $z \in \mathbb{R}^m$ represents a vector of observed grid measurements, which can include various quantities such as bus injections, bus voltages, and also possibly the so-called pseudo-measurements. The *ground-truth* system states, denoted by $x_o \in \mathbb{R}^n$, correspond to the voltages at different buses in the power system. The function h reflects the physical model that ties the state to the quantity measured in the noiseless case and depends on the power system parameters, including the properties of the lines and transformers. The measurement noise is captured by the random vector $\eta \in \mathbb{R}^m$, which is assumed to follow a Gaussian distribution with mean 0 and covariance $\sigma^2 I$[1]. Additionally, the vector a represents the deterministic sparse vector of bad data injections. The sparsity of a implies that only a subset of the measurements is targeted by the adversary. If a particular meter is affected by an adversary, the corresponding entry a_i in a is non-zero.

It is important to note that any other errors arising from modeling and uncertainty in model parameters are combined with the measurement errors η and assumed to be independent of the system parameters. By considering this measurement model, we investigate the effects of bad data injections on the observed measurements and describe one of the widely used methods to detect such attacks.

Remark 1. *The measurement model in eq. (1) is exactly linear for Phasor Measurement Units (PMUs) whose model is:*

$$z = \begin{bmatrix} v \\ i \end{bmatrix} = \begin{bmatrix} I \\ Y \end{bmatrix} v + \eta, \tag{2}$$

where v is the vector of voltages at the grid buses, i is the vector of corresponding currents, and Y is the admittance matrix. In this linear model, the voltages vector is also the state x_o and the function h can be expressed as $h(x) = Hx$, where H is the matrix shown in eq. (2). Any linearized power flow model, such as those proposed in [47], [48], including the DC power flow models, are also special cases.

For the non-linear AC power flow model, the analysis relies on a first-order approximation of the measurement model, substituting the Jacobian matrix (refer to Appendix A) with the system matrix H, as we will discuss in the next section.

B. BDD via Weighted Least Squares

In this section, we will review classical bad data outlier detection methods to define test statistics that can be shared to determine if the system is experiencing an anomaly, without directly sharing the measurements z, which could potentially reveal system and state information.

The Weighted Least Squares (WLS) method is commonly used to estimate the system state by minimizing the weighted sum of squared residuals (WSSR), where the weights are determined by the inverse of the covariance matrix. In the case where the observations are pre-whitened, as assumed in eq. (1), we can consider the equivalent Least Squares (LS) problem without loss of generality:

$$x^\star = \arg\min_x \sigma^{-2} \|z - h(x)\|^2, \tag{3}$$

where $x^\star$ is the state estimate given measurements z that follow the measurement model in eq. (1), and x is the optimization variable.

Next, for convenience's sake, we analyze the linear measurement model but show in Appendix A that the non-linear model can be linearized by utilizing the Jacobian matrices computed at the current state. For the linear case, the WSSR can be written as:

$$\mathfrak{q}(z) = \sigma^{-2} \|z - Hx^\star\|^2 := \sigma^{-2} \|P(\eta + a)\|^2, \tag{4}$$

where we use that $x^\star = \left(H^T H\right)^{-1} H^T z$ and set $P := I - H \left(H^T H\right)^{-1} H^T = P^2$ as the orthogonal projection (or hat) matrix. It is used to project the observed measurements onto the space spanned by the columns of the Jacobian matrix. In the context of detecting bad data, the matrix P is utilized to compute the weighted sum of squared residuals (WSSR) and plays a crucial role in determining the statistical properties of the residual test statistic.

We formulate the detection of bad data as a hypothesis-testing problem with two hypotheses. The null hypothesis $\mathcal{H}_0^{(a)}$ represents the absence of an attack, where $a = 0$. The alternative hypothesis $\mathcal{H}_1^{(a)}$ corresponds to the presence of an attack, indicating that $a \neq 0$. We write this formally as:

$$\mathcal{H}_0^{(a)} : a = 0 \qquad \text{(no attack)} \tag{5a}$$
$$\mathcal{H}_1^{(a)} : a \neq 0 \qquad \text{(attack)} \tag{5b}$$

The query to perform the hypothesis test is the residual test statistic $\mathfrak{q}(z)$ in eq. (4), which allows the analyst to compare the WSSR to a threshold τ. If the WSSR is below τ, we

---

[1]This is without loss of generality since it is always possible to pre-whiten the noise.

accept the null hypothesis $\mathcal{H}_0^{(a)}$, indicating that there is no attack. Otherwise, if the WSSR exceeds τ, we reject $\mathcal{H}_0^{(a)}$ and accept the alternative hypothesis $\mathcal{H}_1^{(a)}$, suggesting the presence of an attack, i.e.:

$$\mathbb{q}(\boldsymbol{z}) \underset{\mathcal{H}_0^{(a)}}{\overset{\mathcal{H}_1^{(a)}}{\gtrless}} \tau. \tag{6}$$

By sharing the WSSR, we allow the analyst to freely choose the threshold τ and determine the optimal trade-off between the probability of false alarm (accepting $\mathcal{H}_1^{(a)}$ when $\mathcal{H}_0^{(a)}$ is true) and the probability of detection (correctly accepting $\mathcal{H}_1^{(a)}$ when $\mathcal{H}_1^{(a)}$ is true). Both probabilities are influenced by the specific values of the bad data vector $\boldsymbol{a}$ and the chosen threshold τ.

Under the assumption of Gaussian additive noise $\boldsymbol{\eta}$ for hypothesis $\mathcal{H}_1^{(a)}$, the WSSR follows a non-central chi-square distribution with r degrees of freedom, where r is the rank of the matrix $\boldsymbol{P}$. The WSSR is centered at $\theta^2 = \sigma^{-2}\|\boldsymbol{Pa}\|^2$, which represents the squared norm of the projection of the bad data vector onto the subspace orthogonal to the columns of $\boldsymbol{H}$, i.e.:

$$\mathbb{q}(\boldsymbol{z}) \sim \chi_r^2\left(\sigma^{-2}\|\boldsymbol{Pa}\|^2\right), \tag{7}$$

where $r = \text{rank}(\boldsymbol{P}) = m - n$. For the null hypothesis, the WSSR follows a central chi-square distribution with r degrees of freedom.

Remark 2. *The stochastic nature of $\mathbb{q}(\boldsymbol{z})$ will play a fundamental role in the development of our privacy mechanism, which diverges from conventional differential privacy methods designed for deterministic queries. The details of our mechanism and its privacy considerations will be discussed in the following sections.*

C. Special Case of the Measurement Model

When $m < n$, the null space is empty, and therefore there are no residuals to share. We propose two approaches. First, we suggest estimating the system state by employing a regularized weighted least squares (RWLS) objective, given by:

$$\boldsymbol{x}^\star = \arg\min_{\boldsymbol{x}} \left(\sigma^{-2}\|\boldsymbol{z} - \boldsymbol{h}(\boldsymbol{x})\|^2 + \lambda\|\boldsymbol{x}\|^2\right), \tag{8}$$

where λ is the regularization parameter that for $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{Hx}$ is solved by:

$$\boldsymbol{x}^\star = \left(\boldsymbol{H}^T\boldsymbol{H} + \lambda\sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{H}^T\boldsymbol{z}, \tag{9}$$

and the WSSR is given by:

$$\mathbb{q}(\boldsymbol{z}) = \frac{\boldsymbol{z}\boldsymbol{P}_\lambda^T\boldsymbol{P}_\lambda\boldsymbol{z}}{\sigma^2} \sim \chi_r^2(\theta_\lambda^2), \tag{10}$$

where r is the rank of $\boldsymbol{P}_\lambda$ and

$$\boldsymbol{P}_\lambda := \left[\boldsymbol{I} - \boldsymbol{H}\left(\boldsymbol{H}^T\boldsymbol{H} + \lambda\sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{H}^T\right], \tag{11}$$

$$\theta_\lambda^2 := (\boldsymbol{Hx} + \boldsymbol{a})^T\boldsymbol{P}_\lambda^2(\boldsymbol{Hx} + \boldsymbol{a})/\sigma^2. \tag{12}$$

Note that eq. (9) reduces to the ordinary least squares solution of $\boldsymbol{x}^\star = \left(\boldsymbol{H}^T\boldsymbol{H}\right)^{-1}\boldsymbol{H}^T\boldsymbol{z}$ when the measurement model contains redundant measurements, i.e., when $m > n$, and by setting $\lambda = 0$.

The second option is an alternative form of regularization based on Graph Signal Processing (GSP) [49]. In the context of a given system, the phasors vector $\boldsymbol{v}$ can be regarded as *low-pass graph signal* [49]. This implies that their empirical covariance matrix has dominant components in the space spanned by the $\kappa < n$ least significant eigenvectors of the system matrix $\boldsymbol{Y}$. In other words, we can approximate $\boldsymbol{v}$ as $\boldsymbol{v} \approx \boldsymbol{U}_\kappa\tilde{\boldsymbol{v}}_\kappa$, where $\boldsymbol{U}_\kappa$ is an $n \times \kappa$ matrix. We can update the linear model as $\boldsymbol{z} = \boldsymbol{H}'\boldsymbol{x}_o' + \boldsymbol{\eta} + \boldsymbol{a}$, where $\boldsymbol{H}' = \boldsymbol{HU}_\kappa$ and $\boldsymbol{x}_o' = \tilde{\boldsymbol{v}}_\kappa$. The LS solution can still be applied as long as $\kappa < m$. Even when $m > n$, this method is useful because it can handle stealth attacks (see Remark 4 in the next subsection). However, it should be noted that the residual in this case may reveal system information, as clarified next.

D. Threat Model

Depending on m, n, and r, if the WSSR is published to an analyst, issues relating to the disclosure of the state (only when $m < n$) and the system matrix may arise. This can be observed in eq. (7) and eq. (10) where the WSSR depends on the system matrix in the former and both the system matrix and the system state in the latter. The privacy leakage is summarized in Table I:

System Matrix Size	System Matrix	System State
$m > n$	Disclosed	Secure
$m < n$	Disclosed	Disclosed

TABLE I: Privacy protection of the system state and matrix.

The publication of system matrices or system states to third parties is a threat to the security and resilience of the electric grid system as they provide valuable information about the system's vulnerabilities. Publishing information about the system's topology, load distribution, or power flow patterns can help attackers identify critical infrastructure components. This reconnaissance information can then be used to plan targeted attacks (such as power outages, equipment damage, or even physical attacks on infrastructure components) that result in severe consequences, including service disruptions and financial losses. Attackers can also use the system's state information to identify vulnerabilities in the system's control systems, such as SCADA systems or energy management systems, and exploit them for cyberattacks.

The publication of system matrices or system states can also pose a privacy risk to energy consumers. The grid system collects and analyzes vast amounts of data on energy consumption patterns, which can be used to infer a consumer's daily routines, lifestyle, and even location. Such information can be exploited for social engineering attacks, such as phishing or spear-phishing, or other forms of cybercrime. Moreover, the privacy of energy consumers is a fundamental right, and any compromise to this right can erode public trust in the operators. For all the aforementioned reasons the system matrix or system states must remain confidential.

As discussed in the prior sections, traditional rules of thumb adopted by specific industries have flawed or no quantification of privacy guarantees, and anonymization often fails in the

presence of substantial side information. In this paper, we address the threat posed by a third-party analyst who may be able to deduce the system state x or the system matrix H by analyzing the residual query [44].

Remark 3 (Internal and External Threats). *We do not consider insiders (of the organization that stores the data) with legitimately acquired access to the data as threats. Instead, we are concerned with the inference of a data point's involvement after a particular aggregate query has been published to an external, untrustworthy third party.*

Remark 4 (Stealth Attacks). *An additional area of concern is stealth attacks where the attacker injects a sparse vector. Here, non-zero entries of the attack vector that correspond to the sensors being attacked are modeled such that residual in eq. (49) is unaffected even with the perturbed state:*

$$Pa = 0 \qquad (13)$$

Here, the attacker can alter the algorithm's output without any change in the loss function of the state estimation problem eq. (3). These types of attacks are only possible when a malicious agent possesses complete knowledge about the system and a non-trivial null space of P exists. Detecting and mitigating such attacks is challenging, particularly in the absence of a specially imposed structure on the actual measurement vectors. The literature proposes various methodologies to detect stealth attacks. As mentioned in the previous section, we recommend using the GSP-based method presented in [49]. This approach aligns well with our measurement model description and is highly effective in detecting stealth attacks.

In the next section, after formally defining the concept of DP, we introduce our proposed DP mechanism for sharing the test statistic. This mechanism aims to safeguard the differential privacy of both the system matrix and the state.

III. DIFFERENTIALLY PRIVATE BAD DATA DETECTION

Motivated by overcoming the disclosure issues of BDD algorithms, in this section, we describe a novel methodology for the publication of a differentially privatized test statistic $\mathbb{q}(z)$ and show how to adjust the performance guarantees of the hypothesis test to account for the loss in accuracy due to the DP mechanism. Before describing our novel methodology, we first provide a brief description of differential privacy.

A. Preliminaries

In the context of a dataset z and a query $\mathbb{q}$, we use the notation $\tilde{\mathbb{q}}(z)$ to represent the differential private answer to the query $\mathbb{q}$. The random outcome of the query, post the application of the DP mechanism, is denoted as $\tilde{q}$, which belongs to the set $\mathcal{Q}$ and follows a distribution $f(\tilde{q}|z)$. This distribution is a probability density function for continuous random queries or a probability mass function for discrete random variables. The common definition of differential privacy from [12], [50] is:

Definition 1 ((ϵ, δ)-Differential privacy). *A randomized mechanism $\tilde{\mathbb{q}}$ is (ϵ, δ)-differentially private if for all neighboring datasets z and z' that differ in one point, for any arbitrary event pertaining to the outcome of the query, the randomized mechanism satisfies the following inequality*

$$\forall \mathcal{S}, \quad Pr(\tilde{\mathbb{q}}(z) \in \mathcal{S}) \leq \exp(\epsilon) Pr(\tilde{\mathbb{q}}(z') \in \mathcal{S}) + \delta, \qquad (14)$$

where $Pr(\mathcal{A})$ denotes the probability of the event $\mathcal{A}$, for some privacy budget $\epsilon \geq 0$ and $\delta \in [0, 1]$.

Note that, since δ is a bound that may not be tight, smaller values of δ are possible. Hence, (ϵ, δ) guarantees are sufficient but not necessary conditions. A second definition in terms of the privacy leakage function is:

Definition 2 ((ϵ, δ)-Probabilistic Differential privacy). *The so-called privacy leakage function $L_{zz'}$ is the log-likelihood ratio between the two hypotheses that the query outcome $\tilde{q}$ is the answer generated by the data z or the data z' that differ by one element. Mathematically:*

$$L_{zz'}(\tilde{q}) := \log \frac{f(\tilde{q}|z)}{f(\tilde{q}|z')}, \qquad (15)$$

A randomized mechanism $\tilde{\mathbb{q}}(z)$ is (ϵ, δ) differentially private for z if and only if:

$$\sup_{zz'} Pr\left(|L_{zz'}(\tilde{q})| \leq \epsilon \right) \geq 1 - \delta. \qquad (16)$$

It can be shown that (ϵ, δ)-PDP is a strictly stronger condition than (ϵ, δ)-DP [51].

B. DP for Stochastic Queries

Earlier in this section, we provided an overview of two definitions of DP. However, we highlight an important caveat regarding these definitions – they primarily focus on protecting the DP of individual elements, denoted as z, within a database z. They aim to conceal the presence or absence of each element in z. Traditionally, mechanisms derived based on these definitions assume that the database is deterministic, lacking any stochastic aspects in its entries. For instance, consider an averaging query on a database consisting of the income of a group of people or biographical details of individuals surveyed for a census [52].

However, in this paper, we address measurement vectors z that arise from a stochastic measurement model based on the physics of the electric grid, as discussed in Section II-B and Remark 2. In this scenario, our primary focus is on safeguarding the DP of the system configuration that generates the measurement model, rather than focusing on an individual instantiation of its measurements. This approach was motivated by the reasons detailed in Section II-D. It is worth noting that the data owner possesses knowledge of the system configuration only in scenarios where the electric utility itself is the data owner. In all other cases, neither the data owner nor the external analyst has access to the system configuration.

For instance, suppose we aim to protect the DP of the elements in the matrix H. In this case, each system configuration gives rise to an *ensemble of query instantiations*. By considering the stochastic nature of the measurements

and focusing on the DP of the system rather than individual measurements, our paper introduces a novel perspective in the realm of differential privacy mechanisms.

The traditional definitions of DP include a neighboring database at distance one (or, in other words, differing in one element). Similarly, we define a distance one neighbor to the system matrix as follows:

Definition 3 (Distance one neighborhood). *Consider a system matrix $\boldsymbol{H} \in \mathbb{R}^{m \times n}$. The distance one neighborhood of $\boldsymbol{H}$ is defined as the set of all matrices that differ from $\boldsymbol{H}$ in exactly one row. More formally,*

$$\left\{ \boldsymbol{H}' \in \mathbb{R}^{m \times n} \mid \boldsymbol{H}' = \boldsymbol{H} + \boldsymbol{e}\boldsymbol{\Delta}_h^T \right\}, \quad (17)$$

where $\boldsymbol{H}$ and $\boldsymbol{H}'$ differ in the mth row (without loss of generality), $\boldsymbol{e} \in \boldsymbol{R}^m$ is a coordinate vector with its mth entry set to 1 and all other entries set to 0. Additionally, $\boldsymbol{\Delta}_h := (\boldsymbol{h}' - \boldsymbol{h})$, where $\boldsymbol{h}$ and $\boldsymbol{h}'$ represent the mth rows of $\boldsymbol{H}$ and $\boldsymbol{H}'$, respectively.

Consequently, when considering a matrix $\boldsymbol{H}'$ from the distance one neighborhood of $\boldsymbol{H}$, the corresponding vector of measurements $\boldsymbol{z}'$ differs from $\boldsymbol{z}$ in exactly one element:

$$z_i' = \begin{cases} z_i, & i < m, \\ \boldsymbol{h}'^T\boldsymbol{x} + a + \eta, & i = m \end{cases} \Rightarrow \boldsymbol{z}' = \boldsymbol{H}'\boldsymbol{x} + \boldsymbol{a} + \boldsymbol{\eta}. \quad (18)$$

In turn, the distance one neighborhood WSSR is given by:

$$\mathbb{q}(\boldsymbol{z}') = \sigma^{-2}(\boldsymbol{a} + \boldsymbol{\eta})^T \boldsymbol{P}'(\boldsymbol{a} + \boldsymbol{\eta}), \quad (19)$$

where $\boldsymbol{P}' = \boldsymbol{I} - \boldsymbol{H}'\boldsymbol{H}'^\dagger$. This implies that the residual follows the non-central chi-square distribution with $r := m - n$ degrees of freedom and non-centrality parameter $\theta'^2 := \sigma^{-2}\|\boldsymbol{P}'\boldsymbol{a}\|^2$:

$$\mathbb{q}(\boldsymbol{z}') \sim \chi_r^2(\theta'^2). \quad (20)$$

Remark 5. *In our framework, it is important to note that the term differential in differential privacy arises from the need to conceal whether a measurement is a result of the system configuration of $\boldsymbol{H}$ or one of its neighboring configurations $\boldsymbol{H}'$. This not only hides the origin of the measurement as part of an ensemble but also enables a more traditional interpretation in terms of the differential of the actual measurement vector, as illustrated in eq. (18).*

Finally, to derive a differentially private mechanism for answering the residual query, we will rely on Definition 2, which provides a direct statistical interpretation. As (ϵ, δ) values approach zero, the log-likelihood ratio, which serves as a sufficient statistic for determining whether the randomized answer is generated from neighboring datasets $\boldsymbol{z}$ or $\boldsymbol{z}'$, produces mostly incorrect or unreliable outcomes. In other words, there is a non-zero probability of the test yielding incorrect results. This trade-off in terms of answer accuracy needs to be carefully considered.

C. Differentially Private Chi-Squared Noise Mechanism

As seen in eqs. (7) and (10), the WSSR query is a non-central chi-square random variable. In this section, we propose a novel additive noise DP mechanism where the WSSR is treated with a random noise drawn from the chi-squared distribution as follows:

$$\tilde{\mathbb{q}}(\boldsymbol{z}) = \mathbb{q}(\boldsymbol{z}) + \nu \quad \text{where} \quad \nu \sim \chi_{r'}^2(0), \quad (21)$$

which implies that $\tilde{\mathbb{q}}(\boldsymbol{z})$ is also a non-central chi-square random variable with $\tilde{r} := r + r'$ DoF and centered at θ^2, i.e.:

$$\tilde{\mathbb{q}}(\boldsymbol{z}) \sim \chi_{\tilde{r}}^2(\theta^2). \quad (22)$$

With this in mind, we state the following theorem guaranteeing the (ϵ, δ)-DP of the chi-square noise mechanism with its proof in appendix B.

Theorem 1 (Chi-square mechanism is (ϵ, δ)-DP). *The mechanism in eq. (21) is (ϵ, δ)-DP for all pairs of neighboring measurement sets $\boldsymbol{z}$ and $\boldsymbol{z}'$ differing in exactly one measurement, where the guarantee δ is given by:*

$$\delta = \max_{\theta, \theta'}\left[Q_{\frac{\tilde{r}}{2}}\left(\theta, \frac{\epsilon}{\theta' - \theta} - \frac{\theta' + \theta}{2}\right) + Q_{\frac{\tilde{r}}{2}}\left(\theta, \frac{\epsilon}{\theta' - \theta} + \frac{\theta' + \theta}{2}\right) \right], \quad (23)$$

where $Q_s(a, b)$ is the Marcum Q-function of order $s > 0$ with $a > 0$ and $b \geq 0$.

The sensitivity analysis is undertaken in appendix C and provides explicit expressions for the δ under assumptions about the system Jacobian.

While this mechanism can be analyzed for smaller m, the analytical calculation to derive δ for larger values of m is not numerically viable, as the Marcum Q-function becomes degenerate in this regime. Thus, in the following section, we provide a Gaussian approximation for the noisy query $\tilde{\mathbb{q}}(\boldsymbol{z})$ that may be used with the Gaussian mechanism for stochastic queries developed in [53] to release the residual query.

D. Gaussian Approximation

The residual query, WSSR, follows a non-central chi-square distribution as discussed in Section II-B. We derive a Gaussian approximation of the WSSR using the following theorem, first proved by [54, Theorem 1]. This provides us with a method for dealing with systems with large m values.

Theorem 2 (Gaussian Approximation of $\mathbb{q}(\boldsymbol{z})$). *Given a measurement vector $\boldsymbol{z}$ and the linear measurement model $\boldsymbol{z} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{a} + \boldsymbol{\eta}$ with $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, and the singular value decomposition of $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, then the following statements hold:*

(a) The WSSR, $\mathbb{q}(\boldsymbol{z})$, is a chi-squared-type mixture:

$$\mathbb{q}(\boldsymbol{z}) = \sum_{i=1}^m d_i z_{U,i}^2, \quad \text{with} \quad z_{U,i}^2 \overset{ind}{\sim} \chi_1^2(\theta_i^2), \quad \forall i \in [m],$$

where

$$\boldsymbol{z}_U := \boldsymbol{U}^T\boldsymbol{z}, \quad \boldsymbol{\theta} := \boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{x} + \boldsymbol{U}^T\boldsymbol{a}, \quad (24)$$

$$\boldsymbol{D} := \left(\boldsymbol{I} - \boldsymbol{\Sigma}\left(\lambda\sigma^2\boldsymbol{I} + \boldsymbol{\Sigma}^T\boldsymbol{\Sigma} \right)^{-1}\boldsymbol{\Sigma}^T \right)^2, \quad \text{and} \quad (25)$$

$$\boldsymbol{d} := \text{diag}(\boldsymbol{D}). \quad (26)$$

(b) The cumulants of $\mathbb{q}(\boldsymbol{z})$ for $\ell = 1, 2, \ldots$ are given by:

$$\mathcal{K}_\ell(\mathbb{q}(\boldsymbol{z})) = 2^{\ell-1}(\ell - 1)! \sum_{i=1}^m d_i^\ell(1 + \ell \cdot \theta_i^2). \quad (27)$$

Let $\zeta := \frac{8\mathcal{K}_2^3(\mathbb{q}(\boldsymbol{z}))}{\mathcal{K}_3^2(\mathbb{q}(\boldsymbol{z}))}$ *and* $\rho := \max_{1 \leq i \leq m} \frac{2d_i^2(1+2\theta_i^2)}{\mathcal{K}_2(\mathbb{q}(\boldsymbol{z}))}$. *Then, for the normalized* $\mathbb{q}(\boldsymbol{z})$ *given by* $\bar{q}(\boldsymbol{z}) = \frac{\mathbb{q}(\boldsymbol{z}) - \mathcal{K}_1(\mathbb{q}(\boldsymbol{z}))}{\sqrt{\mathcal{K}_1(\mathbb{q}(\boldsymbol{z}))}}$, *the following inequality is satisfied:*

$$\sup_q |f_{\bar{q}(\boldsymbol{z})}(q) - \phi(q)| < 0.1323 \left(4 + \frac{0.2503}{(1 - 8\rho)^2} \right) \zeta^{-\frac{1}{2}} \tag{28}$$

when $\rho < 1/8$, *where* $\phi(q)$ *is the density function of a standard normal random variable.*

(c) *If either (i)* $\rho \to 0$ *or (ii)* $\max_{1 \leq i \leq m}(1 + 2\theta_i^2) < \infty$ *and* $\zeta \to \infty$, *is satisfied, then* $\bar{q}(\boldsymbol{z}) \to \mathcal{N}(0, 1)$.

Using theorem 2, we can show that:

$$\mathbb{q}(\boldsymbol{z}) \sim \mathcal{N}(\theta_z, \sigma_z^2), \quad \text{where} \tag{29}$$

$$\theta_z = \text{Tr}(\boldsymbol{D}) + \boldsymbol{\theta}^T \boldsymbol{D} \boldsymbol{\theta}, \quad \text{and} \tag{30}$$

$$\sigma_z^2 = 2\text{Tr}(\boldsymbol{D}^2) + 4\boldsymbol{\theta}^T \boldsymbol{D}^2 \boldsymbol{\theta}. \tag{31}$$

The Gaussian DP mechanism that is used in literature is (ϵ, δ)-DP for a deterministic query. However, in our case, the query is stochastic and, moreover, the variances of the query under two neighboring measurement sets are not the same, i.e., $\sigma_z^2 \neq \sigma_{z'}^2$. Thus, in the following subsection, we derive the DP guarantees for a stochastic query.

IV. PERFORMANCE METRICS AND RESULTS

In Section III-A, we presented our chi-square noise mechanism and its Gaussian approximation for publishing residuals of a BDD algorithm. Adding DP noise inevitably corrupts the residual, which will affect the utility of the residual in bad data detection and lead to degraded performance of the hypothesis test in eq. (6). In this section, the performance of the hypothesis test with the DP noisy residual is quantified through the Receiver Operating Characteristic (ROC) analysis.

The ROC curve is a graphical representation of the trade-off between the probability of detection (denoted by P_d – it is the probability that a true anomaly is correctly identified as such) and the probability of false alarm (denoted by P_{fa} – it is the probability that a normal data point is incorrectly identified as an anomaly) of a hypothesis test.

A perfect hypothesis test would have a P_d of 1 and a P_{fa} of 0, which would mean that all true anomalies are identified as anomalies and all normal data points are identified as normal data. However, in practice, no hypothesis test is perfect, so there is always a trade-off between the two.

The area under the ROC curve (AUROC) indicates how well the hypothesis test can distinguish between anomalies and normal data. A higher AUROC indicates that the test is better at distinguishing between anomalies and normal data.

A. Performance of the chi-square noise mechanism

In this section, we first derive the probabilities of detection and false alarm for the hypothesis test without any DP noise, that is, with the use of the true residual $\mathbb{q}(\boldsymbol{z})$. We then do the same for the residual with the DP noise, $\tilde{\mathbb{q}}(\boldsymbol{z})$.

Suppose the operator sets a probability of false alarm of α, then from the definition of the hypothesis test in eq. (6), we get:

$$\alpha =: P_{fa} = Pr\{\mathbb{q}(\boldsymbol{z}) > \tau \mid \mathcal{H}_0\}, \tag{32}$$

and since $\mathbb{q}(\boldsymbol{z})$ is a central chi-square random variable with r degrees of freedom under $\mathcal{H}_0$, we have:

$$\alpha = 1 - \mathcal{P}\left(\frac{r}{2}, \frac{\tau}{2}\right) := \mathcal{Q}\left(\frac{r}{2}, \frac{\tau}{2}\right), \tag{33}$$

where $\mathcal{P}(\cdot, \cdot)$ is the regularized gamma function and $\mathcal{Q}(\cdot, \cdot)$ is its complementary. Using this relation, we may calculate the threshold to be set as:

$$\tau = 2\mathcal{Q}^{-1}(\alpha, r/2). \tag{34}$$

Similarly, under $\mathcal{H}_1$, $\mathbb{q}(\boldsymbol{z})$ is a non-central central chi-square random variable with a non-centrality parameter of $\theta^2 = \sigma^{-2}\|\boldsymbol{Pa}\|^2$ and r degrees of freedom. Then, the probability of detection is given by:

$$P_d = Pr\{\mathbb{q}(\boldsymbol{z}) > \tau \mid \mathcal{H}_1\} = \mathcal{Q}_{r/2}(\sigma^{-1}\|\boldsymbol{Pa}\|, \sqrt{\tau}),$$
$$= \mathcal{Q}_{r/2}\left(\sigma^{-1}\|\boldsymbol{Pa}\|, \sqrt{2\mathcal{Q}^{-1}(\alpha, r/2)}\right), \tag{35}$$

where $\mathcal{Q}_{\{\cdot\}}(\cdot, \cdot)$ is the Marcum Q-function.

In a similar vein, we may compute these probabilities with the DP residual. First, recall that:

$$\mathcal{H}_0 : \tilde{\mathbb{q}}(\boldsymbol{z}) \sim \chi_{\tilde{r}}^2(0) \quad \text{(no attack)} \tag{36a}$$

$$\mathcal{H}_1 : \tilde{\mathbb{q}}(\boldsymbol{z}) \sim \chi_{\tilde{r}}^2(\theta^2) \quad \text{(attack)} \tag{36b}$$

Suppose $\tilde{r} = r+1$, then, for a threshold of $\tau = 2\mathcal{Q}^{-1}(\alpha, r/2)$, the probabilities are thus calculated as:

$$\tilde{P}_{fa} := Pr\{\tilde{\mathbb{q}}(\boldsymbol{z}) > \tau \mid \mathcal{H}_0\} = \mathcal{Q}\left(\frac{\tilde{r}}{2}, \mathcal{Q}^{-1}(\alpha, r/2)\right), \tag{37a}$$

$$\tilde{P}_d := Pr\{\tilde{\mathbb{q}}(\boldsymbol{z}) > \tau \mid \mathcal{H}_1\} = \mathcal{Q}_{\frac{\tilde{r}}{2}}\left(\theta, \sqrt{2\mathcal{Q}^{-1}(\alpha, r/2)}\right)$$

$$= \mathcal{Q}_{\frac{\tilde{r}}{2}}\left(\sigma^{-1}\|\boldsymbol{Pa}\|, \sqrt{2\mathcal{Q}^{-1}(\alpha, r/2)}\right). \tag{37b}$$

As shown in eqs. (33) to (35) and eq. (37), the additional degree of freedom required by the DP noise is the main reason for the change in performance. This is because the DP noise increases the variance of the residual, which makes it more difficult to distinguish between normal data and anomalies.

B. Performance of the Gaussian approximation

A line of analysis is similar to the one undertaken in section IV-A leads to the following false alarm and detection probabilities for the hypothesis test. Recall that $\mathbb{q}(\boldsymbol{z}) \sim \mathcal{N}(\theta_z, \sigma_z^2)$. The moments of the query vary depending on the hypothesis $\mathcal{H}_0$ or $\mathcal{H}_1$. Then let $\theta_{z,h}$ and $\sigma_{z,h}^2$ denote the mean and variance under hypothesis $\mathcal{H}_h$, for $h = 0, 1$. They are given by:

$$\theta_{z,h} = \text{Tr}(\boldsymbol{D}) + \boldsymbol{\theta}_h^T \boldsymbol{D} \boldsymbol{\theta}_h, \tag{38}$$

$$\sigma_{z,h} = 2\text{Tr}(\boldsymbol{D}^2) + 4\boldsymbol{\theta}_h^T \boldsymbol{D}^2 \boldsymbol{\theta}_h, \tag{39}$$

where:

$$\boldsymbol{\theta}_h = \boldsymbol{\Sigma} \boldsymbol{V}^T \boldsymbol{x} + h \cdot \boldsymbol{U}^T \boldsymbol{a} \quad \text{for } h = 0, 1. \tag{40}$$

The false alarm probability is given by:

$$\alpha =: P_{fa} = \mathcal{Q}\left(\frac{\tau - \theta_{z,0}}{\sigma_{z,0}}\right), \tag{41}$$

where $\mathcal{Q}(\cdot)$ is the Gaussian Q function. The threshold can be calculated as:

$$\tau = \theta_{z,0} + \sigma_{z,0}\mathcal{Q}^{-1}(\alpha). \tag{42}$$

Similarly, the probability of detection is given by:

$$P_d = \mathcal{Q}\left(\frac{\sigma_{z,0}\mathcal{Q}^{-1}(\alpha) - \Delta_{\theta_z}^{(a)}}{\sigma_{z,1}}\right), \tag{43}$$

where $\Delta_{\theta_z}^{(a)} = \theta_{z,1} - \theta_{z,0}$. Similarly, for the DP residual, we have the following false alarm and detection probabilities:

$$\tilde{P}_{fa} = \mathcal{Q}\left(\frac{\tau - \tilde{\theta}_{z,0}}{\tilde{\sigma}_{z,0}}\right) = \mathcal{Q}\left(\frac{\tau - (\theta_{z,0} + \theta_{\nu|z})}{\sqrt{\sigma_{z,0}^2 + \sigma_{\nu|z}^2}}\right), \tag{44}$$

$$\tilde{P}_d = \mathcal{Q}\left(\frac{\sigma_{z,0}\mathcal{Q}^{-1}(\alpha) - (\Delta_{\theta_z}^{(a)} + \theta_{\nu|z})}{\sqrt{\sigma_{z,1}^2 + \sigma_{\nu|z}^2}}\right). \tag{45}$$

C. Numerical Results

In this section, we provide a comprehensive analysis of the performance of our detection algorithm in the presence of DP noise and compare it with an approach involving input perturbation. We have previously discussed the limitations associated with directly perturbing the measurement vector using input perturbation to protect the system and its state. To further explore this, we consider an input perturbation scenario with σ and σ_w denoting the standard deviation of the measurement error and DP noise, which is added using the Gaussian mechanism. Recall that in this scenario, the observed measurement vector $\boldsymbol{z}$ is given by $\boldsymbol{Hx} + \boldsymbol{a} + \boldsymbol{\eta} + \boldsymbol{w}$.
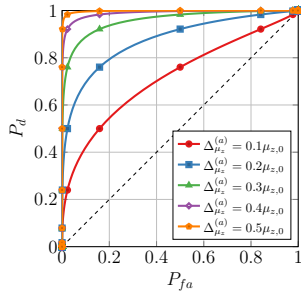
Fig. 3: Sensitivity of the test's performance to attack strength.

Throughout this section, we present a detailed analysis of our detection algorithm's performance under different conditions. In fig. 3, we examine the ROC for binary hypotheses $\mathcal{H}_{\cdot}^{(a)}$ for various values of $\Delta\theta_z^{(a)}$. To establish this, we set the means and variances as follows: $\theta_{z,0} = 10$, $\theta_{z,1} = \theta_{z,0} + \Delta_{\theta_z}^{(a)}$, $\sigma_{z,0} = 1$, $\sigma_{z,1} = 2$. This analysis reveals that the ROC curve deteriorates as the difference $\Delta_{\theta_z}^{(a)}$ between the means of the two hypotheses decreases. This is an expected outcome, as the hypothesis test is more likely to accept the null hypothesis when the difference between the means is small, even under the alternate hypothesis.

In fig. 4, we focus on the scenario where input perturbation of the measurement vector $\boldsymbol{z}$ is performed before conducting
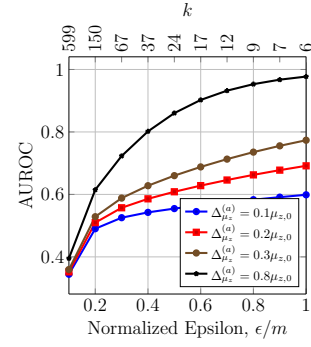
Fig. 4: AUROC vs. DP Normalized privacy budget and noise variance scaling (k-factor) for input-perturbed measurements $\tilde{\boldsymbol{z}} = \boldsymbol{z} + \boldsymbol{w}$.

the hypothesis test. We employ the standard Gaussian mechanism with $\delta = 0.1$ and a sensitivity set to 1. This analysis is conducted for different values of $\Delta_{\theta_z}^{(a)}$ to understand how varying levels of DP noise impact the algorithm's performance when using input perturbation. We plot the AUROC against the DP privacy parameter ϵ and the corresponding k-factor, where the DP noise variance is denoted as $\sigma_w^2 = k\sigma^2$. Notably, we present this information in terms of the normalized (or the per-element) privacy budget, $\epsilon_o := \epsilon/m$, as we are adding noise to each of the m elements in the $\boldsymbol{z}$ vector. As expected, we observe that the AUROC increases with an increase in the per-element privacy budget. This implies that when a smaller standard deviation is used for input perturbation noise, a higher level of performance can be achieved. In practical terms, this figure helps analysts understand the trade-off between the desired level of performance (as defined by the AUROC), the $\Delta_{\theta_z}^{(a)}$, and the allocated privacy budget (ϵ). It provides valuable insights into the resources required to ensure a specific level of performance while safeguarding sensitive information.
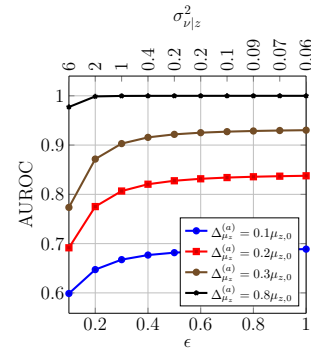
Fig. 5: The AUROC of the hypothesis test for different values of $\sigma_{\nu|z}$ when using our mechanism.

In fig. 5, we illustrate the AUROC of our detection mechanism, incorporating our novel approximate Gaussian DP noise mechanism, across a range of values for $\Delta_{\theta_z}^{(a)}$. Specifically, we set the means and variances as follows: $\theta_{z,0} = 10$, $\theta_{z,1} = 1.3\theta_{z,0}$, $\sigma_{z,0} = 1$, and $\sigma_{z,1} = 4$. It is worth emphasizing that our approach demonstrates superior performance in terms of the privacy budget when compared to input perturbation.

This improved efficiency results from our targeted privacy-preserving strategy, which focuses on perturbing the residual query rather than applying noise to the entire measurement vector. As a consequence, we achieve the desired level of performance while minimizing the expenditure of the privacy budget. It's important to note that in this figure, the AUROC curve is plotted directly against the privacy budget, rather than the normalized privacy budget, further underscoring the efficiency of our mechanism.
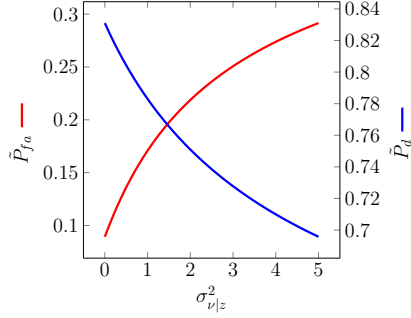

Fig. 6: The probabilities of detection and false alarm of the hypothesis test for different values of $\sigma_{\nu|z}$ when a P_{fa} of 0.05 is required.

In fig. 6, we investigate the performance of the hypothesis test across various values of $\sigma_{\nu|z}$ while maintaining a required P_{fa} of 0.05. As with fig. 5, the means and variances are set as follows: $\theta_{z,0} = 10$, $\theta_{z,1} = 1.3\theta_{z,0}$, $\sigma_{z,0} = 1$, and $\sigma_{z,1} = 4$. This analysis provides insights into how our algorithm behaves under the constraints of a controlled false alarm rate while introducing varying levels of noise (expressed by different $\sigma_{\nu|z}$ values) to the system. As expected, we observe that the probability of false alarm (P_{fa}) increases, and the detection probability (P_d) decreases as we introduce noise with increasing variances. This figure serves to highlight the trade-off between algorithm performance, as defined by P_{fa} and P_d, and the noise variance, and by extension, the privacy budget allocation.

V. CONCLUSION

This paper presents a novel DP chi-squared noise mechanism tailored for power systems, emphasizing residual analysis over direct measurement perturbation. This approach simplifies analytics, requires lower privacy budgets, and introduces a mechanism that considers the stochastic nature of power system measurements—a distinct contribution to DP applications. Our methodology enables precise chi-square DP mechanism application to quadratic queries, enhancing privacy-preserving capabilities within power systems and beyond. By focusing on ensemble-based privacy, this work supports the detection of attacks by third parties without exposing system states or matrices, thus bridging significant gaps in cyber defense. Additionally, our approximation of the chi-squared mechanism to a Gaussian mechanism for stochastic queries illustrates the method's adaptability. We advocate for a collective defense strategy, leveraging a distributed detection framework that reduces data transmission and boosts privacy, addressing the need for cooperative cybersecurity solutions in the increasingly interconnected power grid landscape.

REFERENCES

[1] S. Kalyani and K. S. Swarup, "Particle Swarm Optimization Based K-Means Clustering Approach for Security Assessment in Power Systems," *Expert Systems with Applications*, vol. 38, no. 9, pp. 10 839–10 846, 2011.

[2] W. Wu and M. Peng, "A Data Mining Approach Combining K-Means Clustering With Bagging Neural Network for Short-Term Wind Power Forecasting," *IEEE Internet of Things Journal*, vol. 4, no. 4, pp. 979–986, 2017.

[3] X. Dong, L. Qian, and L. Huang, "Short-Term Load Forecasting in Smart Grid: A Combined CNN and K-Means Clustering Approach," in *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 2017, pp. 119–125.

[4] K. Zor, O. Timur, and A. Teke, "A State-Of-The-Art Review of Artificial Intelligence Techniques for Short-Term Electric Load Forecasting," in *2017 6th International Youth Conference on Energy (IYCE)*. IEEE, 2017, pp. 1–7.

[5] Y. Liu, P. Ning, and M. K. Reiter, "False Data Injection Attacks against State Estimation in Electric Power Grids," *ACM Trans. Inf. Syst. Secur.*, vol. 14, no. 1, jun 2011. [Online]. Available: https://doi.org/10.1145/1952982.1952995

[6] S. Saha, N. Ravi, K. Hreinsson, J. Baek, A. Scaglione, and N. G. Johnson, "A secure distributed ledger for transactive energy: The electron volt exchange (eve) blockchain," *Applied Energy*, vol. 282, p. 116208, 2021.

[7] J. Liu, Y. Xiao, S. Li, W. Liang, and C. P. Chen, "Cyber security and privacy issues in smart grids," *IEEE Communications Surveys & Tutorials*, vol. 14, no. 4, pp. 981–997, 2012.

[8] The White House, "National Cybersecurity Strategy," arpa-e.energy.gov/technologies/programs/grid-data, Mar 2023, (Accessed on 03/21/2024).

[9] Office of Cybersecurity, Energy Security, and Emergency Response, "Considerations for ICS/OT Cybersecurity Monitoring Technologies," https://www.energy.gov/ceser/considerations-icsot-cybersecurity-monitoring-technologies, (Accessed on 03/21/2024).

[10] The White House, "National Security Memorandum on Improving Cybersecurity for Critical Infrastructure Control Systems," https://www.whitehouse.gov/briefing-room/statements-releases/2021/07/28/national-security-memorandum-on-improving-cybersecurity-for-critical-infrastructure-control-systems/, Jul 2021, (Accessed on 03/21/2024).

[11] Office of Cybersecurity, Energy Security, and Emergency Response, "DOE Announces $39 Million in Research Funding to Enhance Cybersecurity of Clean Distributed Energy Resources," 9 2013. [Online]. Available: https://www.energy.gov/ceser/articles/doe-announces-39-million-research-funding-enhance-cybersecurity-clean-distributed

[12] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our Data, Ourselves: Privacy via Distributed Noise Generation," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 2006, pp. 486–503.

[13] M. Esmalifalak, L. Liu, N. Nguyen, R. Zheng, and Z. Han, "Detecting stealthy false data injection using machine learning in smart grid," *IEEE Systems Journal*, vol. 11, no. 3, pp. 1644–1652, 2014.

[14] A. Jindal, A. Dua, K. Kaur, M. Singh, N. Kumar, and S. Mishra, "Decision tree and SVM-based data analytics for theft detection in smart grid," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 3, pp. 1005–1016, 2016.

[15] A. A. Khan, O. A. Beg, M. Alamaniotis, and S. Ahmed, "Intelligent anomaly identification in cyber-physical inverter-based systems," *Electric Power Systems Research*, vol. 193, p. 107024, 2021.

[16] M. R. Habibi, H. R. Baghaee, T. Dragičević, and F. Blaabjerg, "False data injection cyber-attacks mitigation in parallel DC/DC converters based on artificial neural networks," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 68, no. 2, pp. 717–721, 2020.

[17] Z. Liu, J. Tang, Z. Zhao, and S. Zhang, "Adaptive neural network control for nonlinear cyber-physical systems subject to false data injection attacks with prescribed performance," *Philosophical Transactions of the Royal Society A*, vol. 379, no. 2207, p. 20200372, 2021.

[18] E. M. Ferragut, J. Laska, M. M. Olama, and O. Ozmen, "Real-time cyber-physical false data attack detection in smart grids using neural networks," in *2017 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 2017, pp. 1–6.

[19] M. R. Habibi, H. R. Baghaee, T. Dragičević, and F. Blaabjerg, "Detection of false data injection cyber-attacks in DC microgrids based on recurrent neural networks," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 9, no. 5, pp. 5294–5310, 2020.

[20] M. Dehghani, A. Kavousi-Fard, M. Dabbaghjamanesh, and O. Avatefipour, "Deep learning based method for false data injection attack detection in AC smart islands," *IET Generation, Transmission & Distribution*, vol. 14, no. 24, pp. 5756–5765, 2020.

[21] Y. Zhang, J. Wang, and B. Chen, "Detecting false data injection attacks in smart grids: A semi-supervised deep learning approach," *IEEE Transactions on Smart Grid*, vol. 12, no. 1, pp. 623–634, 2020.

[22] L. Yang, Y. Li, and Z. Li, "Improved-ELM method for detecting false data attack in smart grid," *International Journal of Electrical Power & Energy Systems*, vol. 91, pp. 183–191, 2017.

[23] S. Ruj and A. Nayak, "A Decentralized Security Framework for Data Aggregation and Access Control in Smart Grids," *IEEE transactions on smart grid*, vol. 4, no. 1, pp. 196–205, 2013.

[24] M. Wen, R. Xie, K. Lu, L. Wang, and K. Zhang, "Feddetect: A novel privacy-preserving federated learning framework for energy theft detection in smart grid," *IEEE Internet of Things Journal*, vol. 9, no. 8, pp. 6069–6080, 2021.

[25] Y. Chang, J. Li, N. Lu, W. Shi, Z. Su, and W. Meng, "Practical Privacy-Preserving Scheme With Fault Tolerance for Smart Grids," *IEEE Internet of Things Journal*, 2023.

[26] C. Efthymiou and G. Kalogridis, "Smart Grid Privacy via Anonymization of Smart Metering Data," in *2010 first IEEE international conference on smart grid communications*. IEEE, 2010, pp. 238–243.

[27] A. Narayanan and V. Shmatikov, "Robust De-Anonymization of Large Sparse Datasets," in *29th IEEE Symposium on Security and Privacy*, May 2008.

[28] Public Utility Commission of the State of Colorado, "Decision No. R11-0922," *Proposed Rules Relating to Smart Grid Data Privacy for Electric Utilities*, 2011.

[29] N. Ravi, A. Scaglione, S. Kadam, R. Gentz, S. Peisert, B. Lunghino, E. Levijarvi, and A. Shumavon, "Differentially Private-Means Clustering Applied to Meter Data Analysis and Synthesis," *IEEE Transactions on Smart Grid*, vol. 13, no. 6, pp. 4801–4814, 2022.

[30] M. U. Hassan, M. H. Rehmani, and J. Chen, "Differential Privacy Techniques for Cyber Physical Systems: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 1, pp. 746–789, 2019.

[31] P. Barbosa, A. Brito, and H. Almeida, "A technique to provide differential privacy for appliance usage in smart metering," *Information Sciences*, vol. 370, pp. 355–367, 2016.

[32] Y. Chen, A. Machanavajjhala, M. Hay, and G. Miklau, "Pegasus: Data-adaptive differentially private stream processing," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 1375–1388.

[33] J. Liu, C. Zhang, and Y. Fang, "Epic: A differential privacy framework to defend smart homes against internet traffic analysis," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1206–1217, 2018.

[34] P. Pappachan, M. Degeling, R. Yus, A. Das, S. Bhagavatula, W. Melicher, P. E. Naeini, S. Zhang, L. Bauer, A. Kobsa *et al.*, "Towards privacy-aware smart buildings: Capturing, communicating, and enforcing privacy policies and preferences," in *2017 IEEE 37th International Conference on Distributed Computing Systems Workshops (ICDCSW)*. IEEE, 2017, pp. 193–198.

[35] D. Eckhoff and I. Wagner, "Privacy in the smart city—applications, technologies, challenges, and solutions," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 1, pp. 489–516, 2017.

[36] R. Jia, R. Dong, S. S. Sastry, and C. J. Spanos, "Privacy-enhanced architecture for occupancy-based hvac control," in *Proceedings of the 8th international conference on cyber-physical systems*, 2017, pp. 177–186.

[37] S. Ghayyur, Y. Chen, R. Yus, A. Machanavajjhala, M. Hay, G. Miklau, and S. Mehrotra, "Iot-detective: Analyzing iot data under differential privacy," in *Proceedings of the 2018 International Conference on Management of Data*, 2018, pp. 1725–1728.

[38] M. Jawurek, F. Kerschbaum, and G. Danezis, "Sok: Privacy technologies for smart grids–a survey of options," *Microsoft Res., Cambridge, UK*, vol. 1, pp. 1–16, 2012.

[39] C. Xu, J. Ren, D. Zhang, and Y. Zhang, "Distilling at the edge: A local differential privacy obfuscation framework for iot data analytics," *IEEE Communications Magazine*, vol. 56, no. 8, pp. 20–25, 2018.

[40] H. Cao, S. Liu, L. Wu, Z. Guan, and X. Du, "Achieving differential privacy against non-intrusive load monitoring in smart grid: A fog computing approach," *Concurrency and Computation: Practice and Experience*, vol. 31, no. 22, p. e4528, 2019.

[41] H.-Y. Tran, J. Hu, and H. R. Pota, "Smart meter data obfuscation with a hybrid privacy-preserving data publishing scheme without a trusted third party," *IEEE Internet of Things Journal*, vol. 9, no. 17, pp. 16 080–16 095, 2022.

[42] M. T. Hossain, S. Badsha, and H. Shen, "Privacy, security, and utility analysis of differentially private cpes data," in *2021 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2021, pp. 65–73.

[43] M. Gaboardi, H. Lim, R. Rogers, and S. Vadhan, "Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing," in *International conference on machine learning*. PMLR, 2016, pp. 2111–2120.

[44] W.-T. Lin, G. Chen, and X. Zhou, "Privacy-preserving federated learning for detecting false data injection attacks on power system," *Electric Power Systems Research*, vol. 229, p. 110150, 2024.

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

[46] J. Zhao, A. Gómez-Expósito, M. Netto, L. Mili, A. Abur, V. Terzija, I. Kamwa, B. Pal, A. K. Singh, J. Qi *et al.*, "Power system dynamic state estimation: Motivations, definitions, methodologies, and future work," *IEEE Transactions on Power Systems*, vol. 34, no. 4, pp. 3188–3198, 2019.

[47] M. Baran and F. F. Wu, "Optimal Sizing of Capacitors Placed on a Radial Distribution System," *IEEE Transactions on Power Delivery*, vol. 4, no. 1, pp. 735–743, 1989.

[48] M. E. Baran and F. F. Wu, "Network Reconfiguration in Distribution Systems for Loss Reduction and Load Balancing," *IEEE Transactions on Power delivery*, vol. 4, no. 2, pp. 1401–1407, 1989.

[49] R. Ramakrishna and A. Scaglione, "Detection of False Data Injection Attack Using Graph Signal Processing for the Power Grid," in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2019, pp. 1–5.

[50] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis," in *Theory of Cryptography Conference*. Springer, 2006, pp. 265–284.

[51] D. McClure, "Relaxations of Differential Privacy and Risk/Utility Evaluations of Synthetic Data and Fidelity Measures," Ph.D. dissertation, Duke University, 2015.

[52] S. Garfinkel, "Differential Privacy and the 2020 US Census," *MIT Case Studies in Social and Ethical Responsibilities of Computing*, no. Winter 2022, jan 24 2022, https://mit-serc.pubpub.org/pub/differential-privacy-2020-us-census.

[53] R. Ramakrishna, A. Scaglione, T. Wu, N. Ravi, and S. Peisert, "Differential privacy for class-based data: A practical gaussian mechanism," *IEEE Transactions on Information Forensics and Security*, pp. 1–1, 2023.

[54] J.-T. Zhang, "Approximate and Asymptotic Distributions of Chi-Squared–Type Mixtures With Applications," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 273–285, 2005.

[55] D. Ross, "Inequalities for Special Functions," *SIAM Review*, vol. 14, no. 3, p. 494, 1972.

[56] D. J. Bordelon, "Inequalities for Special Functions (D. K. Ross)," *SIAM Review*, vol. 15, no. 3, pp. 665–670, 1973.

APPENDIX A
LINEARIZATION OF THE NON-LINEAR MEASUREMENT MODEL

This is the case in which z contains power flow measurements and, thus $h(x_o)$ are the AC power flow equations. Here, the WSSR may be written as:

$$q(z) := \sigma^{-2}\|z - h(x^\star)\|^2 = \sigma^{-2}\|h(x_o) + \eta + a - h(x^\star)\|^2$$
$$\simeq \sigma^{-2}\|H(x_o - x^\star) + \eta + a\|^2, \tag{46}$$

where the approximation relies on the assumption that $\|x_o - x^\star\|$ is small and the Taylor expansion of h around $x^\star$:

$$h(x_o) \simeq h(x^\star) + H(x_o - x^\star), \tag{47}$$

where $\boldsymbol{H} = \frac{\mathrm{d}\boldsymbol{h}}{\mathrm{d}\boldsymbol{x}}$ is the system Jacobian matrix[2]. Also, at the minimizer, $\boldsymbol{x}^\star$, the gradient of the objective in eq. (3) is zero:

$$
\begin{aligned}
&\boldsymbol{0} = \boldsymbol{\sigma}^{-2}\boldsymbol{H}^T(\boldsymbol{z}-\boldsymbol{h}(\boldsymbol{x}^\star)) = \boldsymbol{\sigma}^{-2}\boldsymbol{H}^T[\boldsymbol{h}(\boldsymbol{x}_o)+\boldsymbol{\eta}+\boldsymbol{a}-\boldsymbol{h}(\boldsymbol{x}^\star)],\\
&\Rightarrow\ \boldsymbol{H}^T\left[\boldsymbol{h}(\boldsymbol{x}_o)-\boldsymbol{h}(\boldsymbol{x}^\star)\right] = -\boldsymbol{H}^T\left(\boldsymbol{\eta}+\boldsymbol{a}\right),\\
&\Rightarrow\ (\boldsymbol{x}_o-\boldsymbol{x}^\star) \simeq -\left(\boldsymbol{H}^T\boldsymbol{H}\right)^{-1}\boldsymbol{H}^T\left(\boldsymbol{\eta}+\boldsymbol{a}\right) \quad (48)
\end{aligned}
$$

where the last equation follows from eq. (47). Finally, from eq. (46) and eq. (48), we have:

$$
\mathbb{q}(\boldsymbol{z}) \simeq \sigma^{-2}\|\boldsymbol{P}(\boldsymbol{a}+\boldsymbol{\eta})\|^2, \quad (49)
$$

where $\boldsymbol{P} := \boldsymbol{I} - \boldsymbol{H}\left(\boldsymbol{H}^T\boldsymbol{H}\right)^{-1}\boldsymbol{H}^T = \boldsymbol{P}^2$ is the orthogonal projection (or hat) matrix. Note that this takes the same form as the residual in the linear case, albeit with the system matrix replaced by the Jacobian.

APPENDIX B
CHI-SQUARE MECHANISM DP PROOF

Letting $s := r + r'$, we have that:

$$
\tilde{\mathbb{q}}(\boldsymbol{z}) \sim \chi_s^2(\theta^2) \quad \text{and} \quad \tilde{\mathbb{q}}(\boldsymbol{z}') \sim \chi_s^2(\theta^2 + \Delta_{\theta^2}) \quad (50)
$$

The log-likelihood ratio of $\tilde{\mathbb{q}}(\boldsymbol{z})$ and $\tilde{\mathbb{q}}(\boldsymbol{z}')$ is given by:

$$
\begin{aligned}
L &= \log\frac{f_{\tilde{\mathbb{q}}(\boldsymbol{z})}(\tilde{q})}{f_{\tilde{\mathbb{q}}(\boldsymbol{z}')}(\tilde{q})} = \log\frac{\frac{1}{2}e^{-(\tilde{q}+\theta^2)/2}\left(\frac{\tilde{q}}{\theta^2}\right)^{\frac{s}{4}-\frac{1}{2}}I_{\frac{s}{2}-1}(\sqrt{\theta^2\tilde{q}})}{\frac{1}{2}e^{-(\tilde{q}+\theta'^2)/2}\left(\frac{\tilde{q}}{\theta'^2}\right)^{\frac{s}{4}-\frac{1}{2}}I_{\frac{s}{2}-1}(\sqrt{\theta'^2\tilde{q}})}\\
&= \frac{\theta'^2-\theta^2}{2}+\left(\frac{s}{4}-\frac{1}{2}\right)\log\frac{\theta'^2}{\theta^2}+\log\frac{I_{\frac{s}{2}-1}(\sqrt{\theta^2\tilde{q}})}{I_{\frac{s}{2}-1}(\sqrt{\theta'^2\tilde{q}})}, \quad (51)
\end{aligned}
$$

where $I_a(b)$ is the modified Bessel function of the first kind.

At this stage, it is important to mention the following theorem on the ratio of modified Bessel functions of the first kind that was independently proved by authors of [55], [56]:

Theorem 3. *For all $a > 0$ and $0 < x < y$, the following inequalities hold:*

$$
e^{x-y}\left(\frac{x}{y}\right)^a < \frac{I_a(x)}{I_a(y)} < e^{y-x}\left(\frac{x}{y}\right)^a. \quad (52)
$$

Next, consider the event $|L| < \epsilon$. Its probability may be written as follows:

$$
P\left(|L| \leq \epsilon\right) = P(L \leq \epsilon) - P(L \leq -\epsilon). \quad (53)
$$

In order to find a lower bound for the probability of occurrence of this event, we shall find a lower bound for the first term

[2]We abuse the notation to denote both the Jacobian of $\boldsymbol{h}$ and the linear measurement model's system matrix by the symbol $\boldsymbol{H}$.

and an upper bound for the second term in eq. (53). We have for $\theta'^2 > \theta^2 > 0$:

$$
P(L \leq \epsilon) = P\left[\left(\begin{array}{c}\dfrac{\theta'^2-\theta^2}{2}+\left(\dfrac{s}{4}-\dfrac{1}{2}\right)\log\dfrac{\theta'^2}{\theta^2}\\ +\log\dfrac{I_{\frac{s}{2}-1}(\sqrt{\theta^2\tilde{q}})}{I_{\frac{s}{2}-1}(\sqrt{\theta'^2\tilde{q}})}\end{array}\right) \leq \epsilon\right]
$$

$$
\geq P\left[\left(\begin{array}{c}\dfrac{\theta'^2-\theta^2}{2}+\left(\dfrac{s}{4}-\dfrac{1}{2}\right)\log\dfrac{\theta'^2}{\theta^2}\\ +\log\left[e^{\sqrt{\theta'^2\tilde{q}}-\sqrt{\theta^2\tilde{q}}}\left(\dfrac{\sqrt{\theta^2\tilde{q}}}{\sqrt{\theta'^2\tilde{q}}}\right)^{\frac{s}{2}-1}\right]\end{array}\right) \leq \epsilon\right]
$$

$$
= P\left(\tilde{q} \leq \left(\epsilon/(\theta'-\theta)-(\theta'+\theta)/2\right)^2\right). \quad (54)
$$

Similarly, an upper bound for the second term in eq. (53) is given by:

$$
P(L \leq -\epsilon) \leq P\left(\tilde{q} \geq \left(\epsilon/(\theta'-\theta)+(\theta'+\theta)/2\right)^2\right). \quad (55)
$$

Thus,

$$
\begin{aligned}
P\left(|L| \leq \epsilon\right) &\geq P\left(\tilde{q} \leq \left(\epsilon/(\theta'-\theta)-(\theta'+\theta)/2\right)^2\right)\\
&\quad - P\left(\tilde{q} \geq \left(\epsilon/(\theta'-\theta)+(\theta'+\theta)/2\right)^2\right)\\
&=: 1-\delta, \quad (56)
\end{aligned}
$$

APPENDIX C
SENSITIVITY ANALYSIS

We are interested in the deviation in $\boldsymbol{P}$ when an element in $\boldsymbol{z}$ is changed. In order to find this deviation, we need to first write $\boldsymbol{P}'$ as a function of $\boldsymbol{P}$. Since $\boldsymbol{H}^T\boldsymbol{e} = \boldsymbol{h}$ (the mth row of $\boldsymbol{H}$), we can write the following:

$$
\begin{aligned}
\boldsymbol{H}'^T\boldsymbol{H}' &= \left(\boldsymbol{H}^T+\boldsymbol{\Delta}_h\boldsymbol{e}^T\right)\left(\boldsymbol{H}+\boldsymbol{e}\boldsymbol{\Delta}_h^T\right)\\
&= \boldsymbol{C}_0+\boldsymbol{H}^T\boldsymbol{e}\boldsymbol{\Delta}_h^T+\boldsymbol{\Delta}_h\boldsymbol{e}^T\boldsymbol{H}+\boldsymbol{\Delta}_h\boldsymbol{e}^T\boldsymbol{e}\boldsymbol{\Delta}_h^T\\
&= \boldsymbol{C}_0+\boldsymbol{h}\boldsymbol{\Delta}_h^T+\boldsymbol{\Delta}_h\boldsymbol{h}^T+\boldsymbol{\Delta}_h\boldsymbol{\Delta}_h^T = \boldsymbol{C}_0-\boldsymbol{h}\boldsymbol{h}^T+\boldsymbol{h}'\boldsymbol{h}'^T\\
&= \left(\boldsymbol{I}+\boldsymbol{C}_1\right)\boldsymbol{C}_0\left(\boldsymbol{I}+\boldsymbol{C}_1\right)^T = \boldsymbol{C}_2\boldsymbol{C}_0^{-1}\boldsymbol{C}_2^T, \quad (57)
\end{aligned}
$$

where $\boldsymbol{C}_0 := \boldsymbol{H}^T\boldsymbol{H}$, $\boldsymbol{C}_1 := \boldsymbol{\Delta}_h\boldsymbol{h}^T\boldsymbol{C}_0^{-1}$ and $\boldsymbol{C}_2 := \left(\boldsymbol{C}_0+\boldsymbol{\Delta}_h\boldsymbol{h}^T\right)$. Using the Sherman-Morrison formula, the inverse of $\boldsymbol{C}_2$ may be written as follows:

$$
\boldsymbol{C}_2^{-1} = \left(\boldsymbol{C}_0+\boldsymbol{\Delta}_h\boldsymbol{h}^T\right)^{-1} = \boldsymbol{C}_0^{-1}\left(\boldsymbol{I}-\frac{\boldsymbol{C}_1}{c_0}\right), \quad (58)
$$

where $c_0 := 1+\boldsymbol{h}^T\boldsymbol{C}_0^{-1}\boldsymbol{\Delta}_h$. Consequently, the inverse of $\boldsymbol{H}'^T\boldsymbol{H}'$ is:

$$
\left(\boldsymbol{H}'^T\boldsymbol{H}'\right)^{-1} = \boldsymbol{C}_2^{-T}\boldsymbol{C}_0\boldsymbol{C}_2^{-1} = \boldsymbol{C}_0^{-1}+\boldsymbol{C}_3, \quad (59)
$$

where

$$
\boldsymbol{C}_3 := -\frac{\boldsymbol{C}_0^{-1}\boldsymbol{C}_1}{c_0}-\frac{\boldsymbol{C}_1^T\boldsymbol{C}_0^{-1}}{c_0}+\frac{\boldsymbol{C}_1^T\boldsymbol{C}_0^{-1}\boldsymbol{C}_1}{c_0^2} \quad (60)
$$

Finally, we can write $\boldsymbol{P}'$ in terms of $\boldsymbol{P}$:

$$
\begin{aligned}
\boldsymbol{P}' &= \boldsymbol{I}-\boldsymbol{H}'\boldsymbol{H}'^\dagger = \boldsymbol{I}-\boldsymbol{H}'\left(\boldsymbol{H}'^T\boldsymbol{H}'\right)^{-1}\boldsymbol{H}'^T\\
&= \boldsymbol{I}-\left(\boldsymbol{H}+\boldsymbol{e}\boldsymbol{\Delta}_h^T\right)\left(\boldsymbol{C}_0^{-1}+\boldsymbol{C}_3\right)\left(\boldsymbol{H}^T+\boldsymbol{\Delta}_h\boldsymbol{e}^T\right)\\
&= \boldsymbol{I}-\boldsymbol{H}\boldsymbol{C}_0^{-1}\boldsymbol{H}^T+\boldsymbol{C}_4 = \boldsymbol{P}+\boldsymbol{C}_4, \quad (61)
\end{aligned}
$$

where

$$C_4 := -HC_0^{-1}\Delta_h e^T$$
$$- \left[HC_3 + e\Delta_h^T \left(C_0^{-1} + C_3 \right) \right] \left(H^T + \Delta_h e^T \right) \quad (62)$$

is the correction in P if H is rank two corrected.

APPENDIX D
NORMAL APPROXIMATION PROOF

Let $H = U\Sigma V^T$ be the SVD of H, where $U = [u_1, \ldots, u_m] \in \mathbb{R}^{m \times m}, V = [v_1, \ldots, v_n] \in \mathbb{R}^{n \times n}$. Then, the RWLS model's state estimate in eq. (9) can be rewritten as:

$$x^\star = V \left(\lambda\sigma^2 I + \Sigma^T\Sigma \right)^{-1} \Sigma^T U^T z \quad (63)$$

and P_λ can be rewritten as:

$$P_\lambda = U \left(I - \Sigma \left(\Sigma^T\Sigma + \lambda\sigma^2 I \right)^{-1} \Sigma^T \right) U^T. \quad (64)$$

Thus, the WSSR is given by:

$$\mathfrak{q}(z) = \frac{z^T U D U^T z}{\sigma^2} = z_U^T D z_U = \sum_{i=1}^m D_{ii} z_{U,i}^2, \quad (65)$$

where $D := \left(I - \Sigma \left(\lambda\sigma^2 I + \Sigma^T\Sigma \right)^{-1} \Sigma^T \right)^2$, $z_U := \frac{U^T z}{\sigma} \sim \mathcal{N}(\theta, I)$, and $\theta := \Sigma V^T x + U^T a$. This implies that:

$$z_{U,i}^2 \overset{ind}{\sim} \chi_1^2(\theta_i^2), \qquad \forall i \in [m]. \quad (66)$$

Thus, the WSSR is a random variable of chi-squared-type mixtures. The theorem then follows from [54, Theorem 1].