Differential Privacy for Regulatory Compliance in Cyberattack Detection on Critical Infrastructure Systems

Paritosh Ramanan, H. M. Mohaimanul Islam, Abhiram Reddy Alugula
School of Industrial Engineering and Management, Oklahoma State University
Stillwater, Oklahoma, USA
{*paritosh.ramanan, h_m_mohaimanul.islam, aalugul*}*@okstate.edu*

***Abstract*—Industrial control systems are a fundamental component of critical infrastructure networks (CIN) such as gas, water and power. With the growing risk of cyberattacks, regulatory compliance requirements are also increasing for large scale critical infrastructure systems comprising multiple utility stakeholders. The primary goal of regulators is to ensure overall system stability with recourse to trustworthy stakeholder attack detection. However, adhering to compliance requirements requires stakeholders to also disclose sensor and control data to regulators raising privacy concerns. In this paper, we present a cyberattack detection framework that utilizes differentially private (DP) hypothesis tests geared towards enhancing regulatory confidence while alleviating privacy concerns of CIN stakeholders. The hallmark of our approach is a two phase privacy scheme that protects the privacy of covariance, as well as the associated sensor driven test statistics computed as a means to generate alarms. Theoretically, we show that our method induces a misclassification error rate comparable to the non-DP cases while delivering robust privacy guarantees. With the help of real-world datasets, we show the reliability of our DP-detection outcomes for a wide variety of attack scenarios for interdependent stakeholders.**

***Index Terms*—Differential Privacy, Industrial Control Systems, Data-driven attacks, Regulatory Compliance**

1. Introduction

Large-scale Critical Infrastructure Networks (CINs) are characterized by physically interdependent subsystems of varying network sizes that are operated by a diverse group of utility stakeholders governed by a regulatory entity. Data-driven cyberattacks targeting key operational technology (OT) components, such as industrial control systems (ICS), of several utilities, have been shown to cause devastating cascading failures that threaten overall network stability [1]. In the United States, there have been several attempts aimed at establishing information-sharing and analysis centers (ISACs) as a means for detecting network-wide ICS attacks through secure aggregation of sensor data as well as local cyber-incident alarms [2]. However, efforts to establish regulatory compliance initiatives like ISACs have severely fallen short of expectations primarily due to the tepid response from stakeholders [3]. Privacy concerns of utilities as well as lack of trust and credibility of reported information form the core set of obstacles that threaten the feasibility of ISAC-like compliance frameworks. Therefore, in this paper, we specifically focus on developing a subsystem-level ICS attack detection frameworks that relies on privacy-preserving disclosures of underlying datasets by utilities. Additionally, our proposed framework can be leveraged by ISAC-like entities to verify compliance of reported detection outcomes with respect to the disclosed datasets leading to increased trust and credibility in regulatory outcomes.

Enhancing the cyber resilience of interdependent CINs in order to limit disruptions from cascading impacts is a key research priority, as outlined by the Cybersecurity and Infrastructure Security Agency (CISA) [4]. In such cases, regulatory entities like ISACs play a critical role in stemming the impacts of data-driven ICS attacks by coordinating cyber-incident response [3] accompanied by timely dissemination of insights. However, the capabilities of ISACs are only as good as the quality of the alarms and the associated datasets reported from the utility stakeholders. In fact, poor quality data collected by ISACs without lack of sufficient context impedes situational awareness, hampers decision-making, and increases false positives occasionally resulting in unnecessary system upgrades [5]. Therefore, enabling ISACs to verify that the reported alarms comply with the underlying datasets of the utilities can help pave the way for improved situational awareness and reduced false alarm rates across the entire network.

The ability to verify compliance of alarms on the basis of underlying datasets is challenging due to privacy concerns of utility stakeholders. Privacy concerns also impede real-time information sharing among CIN stakeholders [6]. It has been demonstrated that operational data from CINs can be used to identify industrial customer demands [7], reveal operational costs of strategic CIN [8], and identify systemic vulnerabilities [9]. There is also considerable trepidation among stakeholders regarding the perceived misuse of information obtained from cyber information-sharing programs [6] by governmental agencies. In some cases CIN stakeholders are also concerned about risks to their organizational reputation [10] in the event of data leaks. As a result, the need of the hour is a private, trustworthy framework for detecting data-driven ICS attacks [11] while enabling data-driven compliance verification of the reported alarms.

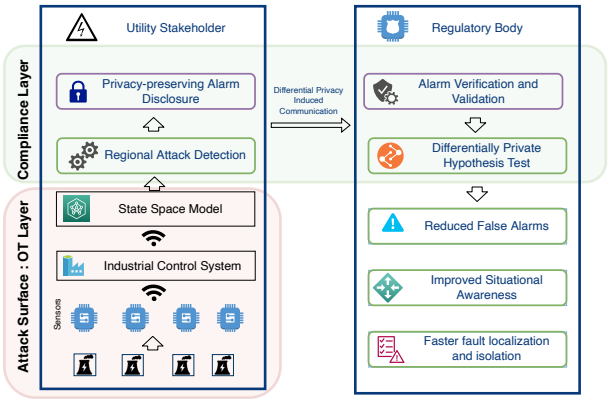


Figure 1: Compliance-driven, private ICS attack detection

Conventionally, ICS frameworks capture IoT and sensor data from assets [12] across each utility. ICS frameworks utilize state-space models to characterize the operational aspect of utilities [13] and detect anomalies and deviations from steady state conditions. A majority of these anomaly detection frameworks utilize statistical hypothesis tests on state space residuals computed using sensor data [14]. As a result, alarms at the utility level are an outcome of anomaly detection frameworks that is inherently characterized by local sensor data as well as the associated covariance matrices. Statistical analysis of residuals can also yield significant insights into the nature of detected anomalies such as attack diagnosis and distinguishing attacks from routine equipment failure [14]. As a result, statistical methods are widely used as a first line of defense in identifying potential attack driven anomalies in ICS frameworks.

In this paper, we consider the interaction between the utility stakeholders and their regulatory counterparts such as ISACs. The goal of the utility is to *convince* an ISAC regarding alarm validity through privacy-preserving disclosures of the underlying residuals and the detection algorithm used for alarms. Our proposed approach therefore uses differential privacy (DP) driven disclosures of data as a means to drive the transparent validation of the alarms. A schematic of our proposed approach is depicted in Figure 1. However, two major bottlenecks arise with this approach. First, statistical hypothesis tests, commonly used for data-driven ICS attack detection, require privacy-preserving disclosures of local covariance matrices to generate alarms [12]. Second, using differentially private data disclosures can itself severely affect the alarm outcomes leading to misclassification with respect to original alarm statistics [15], [16]. Consequently, ensuring compliance and transparency through data sharing introduces a high degree of statistical complexity driven by privacy constraints.

In this paper, we target both these bottlenecks by adopting a two phase DP mechanism to introduce transparency and regulatory compliance for ICS based attack detection. In the first phase, we focus on DP based disclosures of covariance matrices in order to characterize the utility-level detection framework. The second phase involves privacy-preserving disclosures of temporal state-space residual values by leveraging Gaussian differential privacy frameworks. We derive (ϵ, δ) differential privacy guarantees on data-driven ICS attack detection frameworks. Our contributions can be summarized as follows:

- We develop an algorithmic framework for privacy-preserving hypothesis tests that are compatible with DP disclosures of covariance and state-space residuals.
- We develop two distinct implementation modes that enable regulatory bodies like ISACs to independently verify utility detection outcomes using DP disclosures of high-dimensional ICS data.
- We derive strong privacy guarantees governing residual and covariance disclosures in order to obtain DP equivalent levels of significance and test statistic distributions for the corresponding hypothesis tests.
- We theoretically characterize the impact of DP on detection quality by analyzing the DP-induced levels of significance and associated test statistics.

Our proposed framework is evaluated using a generalizable state-space modeling framework that utilizes a non-linear Kalman Filter based approach. Our experimental results are demonstrated using real-world ICS data [17] consisting of diverse attack scenarios on several heterogenous subsystems. The key takeaway of our research is that DP based disclosures offer a viable alternative for establishing regulatory compliance standards, help achieve higher degree of situational awareness, trust and credibility in CINs while providing strong privacy guarantees for utility stakeholders.

2. Related Work

Attacks like DoS, DDoS, phishing, which specifically taget IT systems can often be effectively detected and isolated by monitoring network traffic [18], [19], [20]. On the other hand, data-driven attacks form a more significant threat to ICSs due to their ability to impact information, communication, and the underlying physical systems [21] leading to significant damage to critical infrastructure. In these attacks, sensor data is manipulated in order to effect damage through malicious control actions and incorrect state estimations leading to degraded asset performance.

Several model-based detection mechanisms have been proposed for ICS attacks involving data manipulation [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34]. A popular state-space estimation modeling framework used for ICS attack detection is the Kalman Filter based technique [13], [12], [35], [36], [14]. The Kalman Filter based detection algorithms rely on residuals computed on the basis of the estimated (or predicted) and observed measurements or states followed by a statistical testing procedure [37], [38]. Due to the presence of robust physics-based models in industrial IoT, the Kalman Filter based methods form a powerful class of techniques for anomaly detection. Using a Kalman Filter based state space estimation, one can conduct attack diagnosis [14], distinguish between routine faults and attacks using degradation models [35] and use decentralized methodologies to raise network-wide alarms [36].

On the other hand, recent works have investigated the use of DP as a means to circumvent the privacy-related impediments to enable data sharing in CINs [9]. DP is a widely used method to protect the privacy of data sets intended to be communicated through public domains [39], [40]. DP-driven approaches involve injecting a randomized noise in order to obfuscate the real underlying data record [41], [42]. The injected randomized noise can be designed so as to facilitate theoretical guarantees bounding the loss of privacy [39]. DP thereby ensures that the probability of extracting the real value from a noisy data set by any external entity remains remarkably low. Most DP approaches applied in the context of CINs are geared toward the public release of operational data for benchmarking purposes [43], [44], [45], [7], [9], [46], [47]. The use of DP for protecting the input signals in a Kalman Filter based state space modeling framework has also received considerable attention as well [48], [49], [50].

Additionally, there have been several approaches that have focused on differential privacy in the context of statistical hypothesis tests [15], [51]. Such methods have typically relied on developing DP versions of popular statistical hypothesis tests such as Wilcoxon-signed test [15], goodness-of-fit tests[51], [16], [52], F-test for linear regression significance [53]. However, a majority of them target categorical datasets [16], [52] or are nonparametric in nature [15]. Therefore, while most of these methods provide strong methodological foundations, they are not geared towards temporal datasets that result from state-space modeling approaches such as Kalman Filter. *As a result, there exists a critical methodological gap for DP based approaches that can specifically cater to anomaly detection for temporal state-space models.*

3. State Space Modeling for ICS

Operational modeling of utility stakeholder level ICS is the preliminary step for building a robust detection framework that can be ultimately leveraged for regulatory compliance. In that regard, an exceptional operational model must possess two critical estimation capabilities. First, operational models of utility ICS must incorporate transition functions that can help estimate the future state of the system based on existing sensor data. Second, such models must also be capable of yielding stable and accurate estimations of process and sensor noise distributions which are critical in helping make future state estimations more robust. The core idea is that an exceptional operational model can be utilized to statistically distinguish between normal and anomalous ICS behavior enabling accurate local attack detection which will in turn drive regulatory compliance.

3.1. Non-Linear State Space Formulation

For instituting a local ICS attack detection framework, we begin by discussing a generalizable state-space modeling framework for characterizing utility level ICS operations that are non-linear in nature. Our generalizable state-space model considers a sensor-driven non-linear system at time t, where $x_t \in \mathbb{R}^m$ represents the latent space embedding, $u_t \in \mathbb{R}^m$ represents the control action and $y_t \in \mathbb{R}^d$ represents noisy sensor measurements from asset sensors.

$$x_{t+1} = g(x_{t-1}, u_{t-1}) + v_t, \tag{1}$$

$$y_t = h(x_t) + w_t, \tag{2}$$

In Equations (1), (2), $\boldsymbol{g}, \boldsymbol{h}$ are the state transition and observation functions respectively. Collectively, $\boldsymbol{g}, \boldsymbol{h}$ the represents the relationship between the measurements y_t and the state x_t. The process and measurement noises at time t are denoted by $v_t \in \mathbb{R}^m$, and $w_t \in \mathbb{R}^m$ respectively. The process and measurement noises follow multivariate normal distributions with zero mean implying that $v_t \sim N(0, Q_t), w_t \sim N(0, R_t)$, where Q_t, R_t represent the covariance matrices respectively. Such a type of modeling framework has also been used extensively in prior art [54], [28], [24]. A Non Linear Kalman Filter can be used to model the state space of the stakeholder ICS as represented by Equations (1) and (2).

3.2. Non-linear Kalman Filter Estimation

In the non-linear model given in Equations (1), (2), an extended Kalman Filer based model denoted by $\mathcal{K}$ can be formulated using the following equations.

$$\hat{x}_{t|t-1} = g(\hat{x}_{t-1|t-1}, u_{t-1}), \tag{3}$$

$$r_t = y_t - h(\hat{x}_{t|t-1}), \tag{4}$$

$$P_{t|t-1} = G_t P_{t|t-1} G_t^T + Q_{t-1}, \tag{5}$$

$$S_t = (H_t P_{t|t-1} H_t^T + R_t)^{-1}, \tag{6}$$

$$K_t = P_{t|t-1} H_t^T S_t \tag{7}$$

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t r_t \tag{8}$$

$$P_{t|t} = (I - K_t H_t) P_{t|t-1} \tag{9}$$

In Equations (5),(7), $P_{t|t-1}, P_{t|t}$ represents the predicted and the updated covariance estimates respectively, while S_t represents the residual covariance at t. The state transition and the observation matrices at t given by F_t, H_t respectively are computed using the Jacobians of g, h as denoted by the following equations.

$$G_t = \left.\frac{\partial g}{\partial x}\right|_{x_{t-1|t-1}, u_t} \text{ and } H_t = \left.\frac{\partial h}{\partial x}\right|_{x_{t|t-1}} \tag{10}$$

3.3. Learning the Non Linear Kalman Filter

The stakeholder level ICS system can be modeled as the Non Linear Kalman Filter (NLKF) described in Section 3.2. The NLKF model requires a generalizable framework that is known *a priori* and can characterize the stakeholder level ICS system dynamics including the estimation of process and measurement noise covariance estimation based on current and historic sensor measurements and state estimates.

However, constructing a generalizable framework that efficiently captures ICS system dynamics is challenging due to the tedious nature of estimating and fitting parametrized probability distributions to process and measurement noise with covariance Q_t, R_t respectively. Without accurate knowledge of these covariance matrices, the Kalman gain cannot be computed precisely resulting in

erroneous posteriori state estimates predictions which can compromise the detection quality for data driven attacks.

Therefore, we utilize a machine learning framework comprising of Long and Short Term Memory (LSTM) based recurrent neural networks complemented by feed-forward layers that attempt to cast the ICS process towards a Non Linear Kalman Filter based setting. More precisely, the Non Linear Kalman Filter based LSTM (NLKF) design involves the posteriori state estimates of the prior time step $\hat{x}_{t|t}$ in addition to the observed sensor measurements y_t to *predict* the process and measurement noise covariance Q_t, R_t.

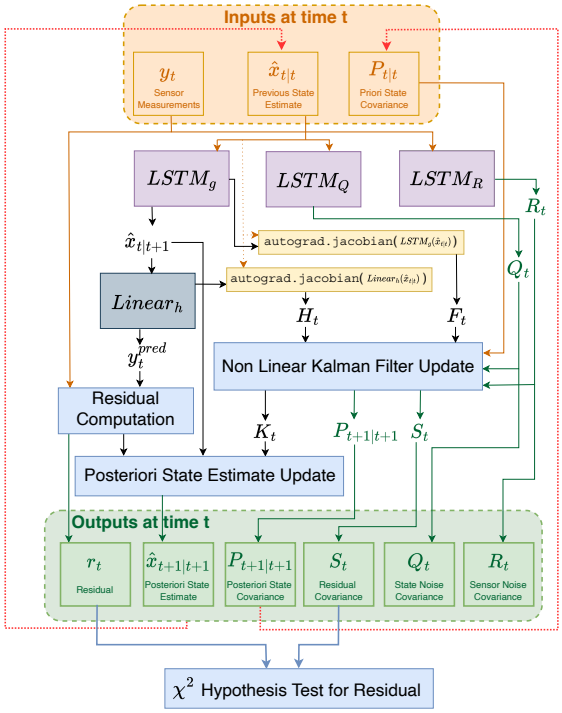


Figure 2: Framework for learning Non Linear Kalman Filter

Figure 2 depicts the NLKF framework that is used for learning the state space model. The NLKF framework consists of three LSTM ($LSTM_g$, $LSTM_Q$ and $LSTM_R$) and one feed-forward Linear layers ($Linear_h$) that are collectively used to learn the non-linear interdependencies between the sensor measurements y_t, and the priori and posteriori state estimates $\hat{x}_{t|t-1}$ and $\hat{x}_{t|t}$. Specifically, $LSTM_g$ learns the relationship between the current state estimate $\hat{x}_{t|t}$ and the *a priori* state estimate $\hat{x}_{t|t+1}$. $LSTM_Q$ and $LSTM_R$ predict the state and measurement covariance matrices based on $\hat{x}_{t|t}$ and y_t respectively. We use $Linear_h$ to map the lower dimensional *a priori* state estimates to the high dimensional sensor data space in order to predict y_t^{pred} such that residual $r_t = y_t - y_t^{pred}$ can be computed. Using the autograd functionality in machine learning frameworks like PyTorch, we compute the Jacobians H_t and G_t for the $LSTM_g$ and $Linear_h$ respectively. Finally, the estimated Jacobians, sensor noise, state noise and state covariance matrices enable us to calculate the extended Kalman gain matrix K_t and the posterior state covariance $P_{t+1|t+1}$ according to Equations (5)-(7). Therefore, the NLKF framework provides us with the fundamental ability to generate residual r_t estimates along with temporally sound predictions of the covariance matrices Q_t, R_t, S_t as well.

3.4. Hypothesis Tests for Anomaly Detection

Estimation of covariance matrices enables us to implement statistical hypothesis tests on the corresponding covariates themselves. As a result, we can leverage the NLKF framework, for orchestrating the χ^2 hypothesis tests, that can detect covert attacks [12] using S_t and r_t. Additionally, the same methodology can also be used for detecting false data injection attacks using sensor data measurements y_t and covariance R_t[13].

To establish the theoretical underpinnings of the χ^2 hypothesis test, we consider without any loss of generality, the predicted residual covariance matrix S_t and the residual r_t at time t. S_t is symmetric, which means that $S_t = V_t \Lambda_t V_t^T$, where V_t is describes the set of orthonormal eigenvectors and Λ_t is a diagonal matrix representing a set of eigenvalues. As a result, we can obtain the principal component decomposition of S_t. We know that $(S_t)^{-1/2} = (V \Lambda_t^{-1} V_t^T)^{1/2} = V_t \Lambda_t^{-1/2}$. Given a residual vector $r_t \sim N(0, S_t)$, we can realize $\tau_t = (S_t)^{-1/2}(r_t) = S_t \Lambda_t^{-1/2}(r_t)$ such that $\tau_t \sim N(0, I)$ [12], [55]. Therefore, to detect attacks, the utility stakeholder level χ^2-hypothesis test can be formulated as follows:

$$H_0 : S_t^r = S_t \tag{11}$$

$$H_A : S_t^r \neq S_t^{true} \tag{12}$$

The null hypothesis H_0 represents the condition that the covariance of the reported residual vector r_t denoted by S_t^r is equal to the expected covariance S_t predicted by the NLKF framework. In other words, the null hypothesis tests whether the residual vector r_t indeed follows the parametrized distribution $N(0, S_t)$. The underlying insight behind H_0 is that perturbations due to abnormal sensor readings or faulty computation of state estimates, will result in residual vectors that adhere to a covariance matrix that is statistically different from the one predicted by the NLKF framework. Further, we can also state that if H_0 holds, then the standardized vector of principal component (PC) scores is given by τ_t implies that $||\tau_t||_2^2 \sim \chi_p^2$, where $p < d$ is the degrees of freedom corresponding to the number of principal components used [55].

$$\rho = \begin{cases} 1, & \text{if } T_{\chi^2,t} = \|\tau_t\|_2^2 > \chi_{m,\alpha}^2, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Based on Equation (13), an alarm $\rho = 1$ is triggered when the test statistic $T_{\chi^2,t} = ||\tau_t||_2^2 > \chi_{m,\alpha}^2$, where α is the level of significance, while $\rho = 0$ otherwise [13].

4. Privacy Preserving Detection Scheme

Detecting attacks requires continuous monitoring of the residual vector. However, the alarms might need to be publicly validated by a third party (such as a nodal authority or overseer agency) due to regulatory and compliance reasons. As a result, residual vectors and the covariance matrices

need to be disclosed to facilitate public validation of alarm values so as to meet regulatory compliance. However, a critical challenge arises in the inability to provide strong guarantees regarding the privacy loss stemming from the public disclosure of these data.

4.1. Privacy Scope and Objectives

We focus on a utility based subsystems and its associated state-space model as detailed in Section 3. Our privacy scope particularly focuses on the disclosure of the S_t and r_t resulting from NLKF framework as a means to validate the alarm value ρ_t. Our objective is to bound the privacy loss that results from the public disclosure of S_t and r_t while simultaneously bounding the miss-classification probability of alarms as a direct consequence of ensuring privacy.

4.2. Differential Privacy Primer

We provide a brief summary of the following useful concepts pertaining to differential privacy.

Definition 1. *[39] : A randomized mechanism $\mathcal{M} : \mathbb{R}^n \rightarrow \mathbb{R}^k$ is said to preserve (ϵ, δ)- differential privacy with respect for all points $x_1, x_2 \in \mathbb{R}^n$, the following holds*

$$\mathbb{P}(\mathcal{M}(x_1) \in \mathcal{O}) \leq e^{\epsilon}\mathbb{P}(\mathcal{M}(x_2) \in \mathcal{O}) + \delta$$

In other words, Definition 1 ensures that for any two points x_1, x_2 in domain $\mathbb{R}^n$, the probability that mechanism $\mathcal{M}(x_1)$ leads to an output in the set $\mathcal{O} \subseteq \mathbb{R}^k$ is upper bounded by the probability that mechanism $\mathcal{M}(x_2)$ leads to an output in the set $\mathcal{O} \subseteq \mathbb{R}^k$ scaled by exponentiation of the privacy loss $\epsilon > 0$ with an addition of $\delta > 0$.

For DP mechanisms, we characterize the concept of adjacency with respect to a multi-variate function $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$. Any two elements $x_1, x_2 \in \mathbb{R}^d$ are deemed to be adjacent, if they differ in at most 1 element. Consequently, the sensitivity of a function depends on the maximum difference in function values caused by adjacent elements. Definition 2 formalizes the concept of sensitivity.

Definition 2. *: The l_k-sensitivity of a d dimensional function $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is defined as $\Delta_{kf} = \max_{x,y} ||f(x) - f(y)||_k$, where x and y are adjacent elements.*

Most common DP formulations utilize $k = \{1, 2\}$ denoting l_1 or l_2 sensitivity assumptions respectively. In this paper, we adopt a two phase approach towards differential privacy wherein each phase is applied sequentially and pertains to protecting the privacy of the covariance matrix and residual vectors respectively. More specifically, we leverage a DP-based method [56] for privacy preserving covariance matrix factorization while employing a Gaussian Differential Privacy (GDP) approach for residual vectors.

Definition 3. *[39] A mechanism is deemed to follow Gaussian Differential Privacy (GDP) if it injects independent noise $e_i \sim \mathcal{N}(0, \sigma^2)$, where $\sigma > 0$ is the noise scale, to each component of p dimensional function $f : \mathbb{R}^p \rightarrow \mathbb{R}^p$.*

The GDP mechanism provides (ϵ, γ_r)-DP when $\epsilon \in (0, 1)$, $c^2 > 2ln(1.25/\gamma_r)$ and $\sigma \geq c\Delta_2/\epsilon$. As a result, we have Equations (14) and (15). Since the l_2 sensitivity Δ_2 is used in the context of GDP of residuals, we use $\Delta_2 = \Delta_r$ for notational clarity.

$$\sigma > \frac{\Delta_r}{\epsilon}\sqrt{2ln(\frac{1.25}{\gamma_r})} \tag{14}$$

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{p} e_{i,t}\Big| \geq \theta_r\Big) \leq \gamma_r \ , \ \theta_r = \frac{\sigma^2\epsilon}{\Delta_r} - \frac{p\Delta_r}{2} \tag{15}$$

4.3. Privacy Preserving Covariance Disclosures

We utilize the Laplacian noise based DP method [56] that factorizes S_t to yield a DP driven perturbed matrix $\hat{S}_t$ as given in Equation (16). However, instead of directly perturbing S_t, we compute its square root factorization $\tilde{S}_t = V_t\Lambda_t^{\frac{-1}{2}}V_t^T$. Next we perturb eigenvalues of $\tilde{S}_t$ as specified in [56] so as to compute $\hat{S}_t = \hat{V}_t\hat{\Lambda}_t^{\frac{-1}{2}}\hat{V}_t^T$. Consequently, we obtain $\hat{S}_t^{\frac{-1}{2}} = \hat{V}_t\hat{\Lambda}^{\frac{-1}{2}}$. Using $\hat{S}_t^{\frac{-1}{2}}$, we can also obtain a perturbed lower dimensional residual $\hat{\tau}_t^{cov}$ and its corresponding test statistic $\hat{T}_{\chi^2,t}^{cov}$ denoted by Equations (17) and (18) respectively. Using Laplacian noise to perturb eigenvalues as described in [56] ensures $(\epsilon_{cov}, \gamma_{cov})$-DP for all $\lambda_i, i \in \{1, d\}$ such that $\hat{\lambda}_i = \lambda_i + Lap(\Delta_l/\epsilon_{cov})$.

$$\hat{S}_t = \hat{V}_t\hat{\Lambda}_t\hat{V}_t^T \tag{16}$$

$$\hat{\tau}_t^{cov} = (\hat{S}_t)^{-1/2}r_t \tag{17}$$

$$\hat{T}_{\chi^2,t}^{cov} = (\hat{\tau}^{cov})_t^T\hat{\tau}_t^{cov} = ||\hat{\tau}_t^{cov}||_2^2 \tag{18}$$

Further, we note that the perturbed test statistic $\hat{T}_{\chi^2,t}^{cov}$ is computed using a DP-induced covariance matrix $\hat{S}_t$ *and an unperturbed* residual vector r_t. Consequently, we can state the relations defined by Equations (19) and (20) as stated in [56].

$$E_l = ||\Lambda - \hat{\Lambda}||_1 = \max_i |\hat{\lambda}_i - \lambda_i| \tag{19}$$

$$\mathbb{P}\Big(E_l \leq \frac{\Delta_l}{\epsilon_{cov}} log\Big(\frac{d}{\gamma_{cov}}\Big)\Big) \geq 1 - \gamma_{cov} \tag{20}$$

Equation (20) ensures that the maximum absolute Laplacian noise value given by E_l as defined in Equation (19) is less than $\frac{\Delta_l}{\epsilon_{cov}} log\Big(\frac{d}{\gamma_{cov}}\Big)$ with probability at least $1 - \gamma_{cov}$. Analyzing the effect of DP-induced covariance matrices helps us establish Lemma 4.1.

Lemma 4.1. *The original and the perturbed test statistics $\hat{T}_{\chi^2,t}^{cov}$, $T_{\chi^2,t}$ satisfy the following relation for $p \leq d$:*

$$\mathbb{P}\Big[|\hat{T}_{\chi^2,t}^{cov} - T_{\chi^2,t}| \leq R_t\theta_l\Big] \geq 1 - \gamma_{cov}$$

where $R_t = \sum_{i=1}^{p}(r_t^i)^2$ and $\theta_l = \Big(\frac{\Delta_l}{\epsilon_{cov}} log\Big(\frac{d}{\gamma_{cov}}\Big)\Big)$

Proof of Lemma 4.1 has been presented in Appendix A. Lemma 4.1 enables us to probabilistically bound the observed difference in test statistic value, *only when DP is applied on the covariance matrix*. However, in the interest of full compliance, it is also important to reveal the reduced residual values, such that the regulator can carry out the

entire hypothesis test workflow represented in Equations (11) - (13). Doing so, would require privacy protections on residuals as well, which we explore in the following section.

4.4. Privacy Preserving Residual Disclosure

We consider the unperturbed low dimensional residual representation $\tau_t = (S_t)^{-1/2}(r_t)$ where $\tau_t \in \mathbb{R}^p$ as defined in Section 3.4. We use the GDP mechanism presented in Definition 3 to perturb τ_t according to Equation (21).

$$\hat{\tau}_t^{res} = \tau_t + e_t, e_t \sim N(0, \sigma^2 I) \tag{21}$$

$$\hat{T}_{\chi^2,t}^{res} = (\hat{\tau}_t^{res})^T \hat{\tau}_t^{res} = ||\hat{\tau}_t^{res}||_2^2 \tag{22}$$

$\hat{T}_{\chi^2,t}^{res}$ in Equation (22) represents the test statistic obtained *purely* through the GDP perturbation of τ_t. Generating $\hat{T}_{\chi^2,t}^{res}$ is especially useful in enabling implementation flexibilities for compliance verification methods as discussed in Section 5. The GDP perturbation leads us to Lemma 4.2 which provides probabilistic bounds on the GDP noise vector e_t.

Lemma 4.2. *Under conditions of GDP, for a given value of γ_r, ϵ, the following condition must hold*

$$\mathbb{P}\Big[||e_t||_2^2 \leq \frac{\theta_r^2}{p}\Big] \geq (1-\gamma_r)^p$$

Proof of Lemma 4.2 has been presented in Appendix A. Lemma 4.2 is vital towards deriving the overall privacy implications when the covariance as well as residual perturbations are integrated and presented to the regulator for compliance verification.

We now turn our attention towards characterizing the probability distribution of the GDP induced test statistic $\hat{T}_{\chi^2,t}^{res}$. As a result, we consider the non-central χ^2 distribution with non-centrality parameter μ and k degrees of freedom denoted by $\chi^2(k,\mu)$. Further, we denote $F_{\chi^2,p}^{\tau_t}(x)$ as the CDF of $\chi^2(p, ||\tau_t||_2^2)$. Consequently, we derive Lemma 4.3 which establishes the probability distribution of the variance scaled test statistic $(\hat{T}_{\chi^2,t}^{res}/\sigma^2)$ under GDP provisions.

Lemma 4.3. *Under GDP,* $(\frac{\hat{T}_{\chi^2,t}^{res}}{\sigma^2}) \sim \chi^2(p, \frac{||\tau_t||_2^2}{\sigma^2})$

Proof of Lemma 4.3 has been presented in Appendix A. Lemma 4.3 provides a distributional characterization of the perturbed residual test statistic under GDP conditions. Therefore, it serves as a precursor to Theorems 4.1 helping establish probabilistic bounds between perturbed and original low dimensional residual representations. Lemma 4.3 also proves to be an important enabler for computing DP-informed level of significance for carrying out the χ^2 hypothesis test at the regulator as represented in Theorem 4.2.

Theorem 4.1. *For an (ϵ, δ)-DP Gaussian mechanism on a given τ_t, the following result holds*

$$\mathbb{P}\Big[L \leq (\hat{T}_{\chi^2,t}^{res} - T_{\chi^2,t}) \leq U \Big| ||e_t||_2^2 \leq \frac{\theta_r^2}{p}\Big] \geq (F_{\chi^2,p}^{\tau_t}(U) - F_{\chi^2,p}^{\tau_t}(L))(1-\gamma_r)^p$$

with $L = \frac{\theta_r}{\sigma^2 p}\Big(\theta_r - 2\sum_{i=1}^{p}\tau_{i,t}\Big)$ *and* $U = \frac{\theta_r}{\sigma^2 p}\Big(\theta_r + 2\sum_{i=1}^{p}\tau_{i,t}\Big)$

Proof of Theorem 4.1 has been presented in Appendix A. The main contribution of Theorem 4.1 lies in its ability to link the perturbed test statistic $\hat{T}_{\chi^2,t}^{res}$ as a function of the non-central χ^2 CDF with p degrees of freedom centered on the original test statistic $T_{\chi^2,t}$. As a consequence of Theorem 4.1, we can derive lower and upper bounds on the difference between original and perturbed tests statistics purely as a function of GDP parameters and the unperturbed low dimensional residual representation τ_t.

4.5. Integrating Covariance and Residual Privacy

Lemma 4.1 deals with aspects of privacy when only the covariance undergoes DP, Theorem 4.1 applies to a scenario wherein GDP is applied only to lower dimensional residual representations. However, integrating both of these individual DP steps is important so as to enable the disclosures of both covariance and residuals *separately* by the utility. Separate disclosures are vital to recreate the hypothesis test workflow at the regulator level to help satisfy compliance while preserving privacy of utility level operations. In this section, we address two issues which arise as a consequence of separate disclosures. First, introduction of DP noise across the covariance and residual steps might provide different test results at the regulator compared those of the original test at the utility level. Second, minimizing these miss-classifications at the regulator level requires an alternate version of the test characterized by a level of significance customized for DP noise injected at the covariance and residual levels. Lastly, regulator level tests that consider separate disclosures covariance and residuals require strong privacy bounds that can help inform the implementation modalities.

We consider $T_{\chi^2,t}$, $\hat{T}_{\chi^2,t}^{cov}$ and $\hat{T}_{\chi^2,t}^{res}$ corresponding to τ_t, $\hat{\tau}_t^{cov}$ and $\hat{\tau}_t^{res} = \hat{\tau}_t^{cov} + e_t$ respectively. Further, we consider

$$\Pi_t = \{r_t, \gamma_{cov}, \gamma_{res}, \epsilon_{cov}, \epsilon_l, \epsilon_r, \Delta_l, \Delta_r\} \tag{23}$$

In Equation (23), Π_t denotes the set of observed residuals and DP parameters at time t. We let $\hat{\rho}_t, \rho_t$ denote the alarms raised with and without differential privacy respectively. We first establish Lemma 3 as a means to determine the upper bound the reduced dimensional DP residual τ given the distributional knowledge of the perturbed test statistic.

Lemma 4.4. *Given $\hat{T}_{\chi^2,t}^{res}(\tau) = ||\tau + e||_2^2$, $e \sim N(0, \sigma^2 I)$,*

$$\underset{\tau}{argmax}\ \mathbb{P}\Big[\hat{T}_{\chi^2,t}^{res} > \sigma^2\phi\Big] = \underset{\tau}{argmax}(||\tau||_2)$$

where ϕ belongs to the support of distribution $\chi^2(p, \frac{||\tau||_2^2}{\sigma^2})$

Proof of Lemma 4.4 has been presented in Appendix A.

We now consider the integration of DP driven covariance matrix structure defined in Section 4.3 and the GDP induced residuals discussed in Section 4.4. In order to do so, we consider the sequential application of DP on the covariance matrix followed by the GDP phase on the residual. Adopting a sequential approach enables a seamless framework of the-

oretical analysis that can be used to derive privacy oriented miss-classification rates, alternate levels of significance as well as efficient and customizable implementation strategies. As a result, we have Equations (24)-(25) that represent the *sequential privacy scheme*.

$$\hat{\tau}_t^{cov} = (\hat{S}_t)^{-1/2} r_t \tag{24}$$

$$\hat{\tau}_t^{res} = \hat{\tau}_t^{cov} + e_t, e_t \sim N(0, \sigma^2 I) \tag{25}$$

Next we consider $\chi^{2,NC}_{\hat{\alpha}}$ which denotes the $\hat{\alpha}$ level upper quantile of the non-central χ^2 distribution given by $\chi^2(p, \hat{T}^{cov}_{\chi^2,t})$, where $\hat{T}^{cov}_{\chi^2,t} = ||\hat{\tau}_t^{cov}||_2^2$ forms the covariance privacy induced test statistic defined in Equation (18). We also note that $\chi^{2,NC}_{\hat{\alpha}}$ denotes the $\hat{\alpha}$ level upper quantile representing the level of significance for the hypothesis test carried out with DP measures. Similarly, χ^2_α represents the α level of significance for the non-DP hypothesis test. We obtain Theorem 4.2 which considers a scenario wherein a regulator tries to recreate the hypothesis test on the basis of utility stakeholder disclosures that follow the sequential privacy scheme.

Theorem 4.2. *Under the sequential privacy scheme, the Type-I error rate of the DP hypothesis test with $\hat{\alpha}$ level of significance is upper bounded by $\mathcal{E}_I^{max}(\hat{\alpha})$, where*

$$\mathcal{E}_I^{max}(\hat{\alpha}) \leq \Big[1 - F^{gamma}_{r_{max}}(R_t\theta_l)\Big]\Big[1 - F^{\hat{\tau}_t^{cov}}_{\chi^2,p}\Big(\sigma^2\chi^{2,NC}_{\hat{\alpha}}\Big)\Big] + \Big[F^{ex}(\theta_l)\Big]^p\Big[1 - F^{\hat{\tau}^{cov}_{max,t}}_{\chi^2,p}\Big(\sigma^2\chi^{2,NC}_{\hat{\alpha}}\Big)\Big]$$

where $F^{ex}(.)$, $F^{gamma}_{r_{max}}(.)$ are the CDFs of $Exp(\frac{\epsilon_{cov}}{\Delta_l})$ and $Gamma(p, \frac{\epsilon_{cov}}{\Delta_l r^2_{max}})$ respectively.

Proof of Theorem 4.2 has been presented in Appendix A. Theorem 4.2 helps characterize the Type-I error rate $\mathcal{E}_I(\hat{\alpha})$ of the hypothesis test with $\hat{\alpha}$ level of significance carried out at the regulator. We can see that $\mathcal{E}_I(\hat{\alpha})$ is dependent on θ_l which consists of the covariance DP parameter γ^{cov}, covariance DP induced low dimensional residual $\hat{\tau}_t^{cov}$, its corresponding maxima $\hat{\tau}^{cov}_{max,t}$ computed according to Lemma 4.4 as well as their CDFs $F^{\hat{\tau}_t^{cov}}, F^{\hat{\tau}^{cov}_{max,t}}$ respectively. Additionally, we note that $\mathcal{E}_I^{max}(\hat{\alpha})$ also incorporates the GDP variance parameter σ as well. Therefore, Theorem 4.2 lays the foundation for computing the Type 1-error rate of the sequential privacy scheme that is collectively influenced by the privacy measures at the covariance and residual steps. We specifically derive a Gamma function dependent upper bound in Theorem 4.2 since it is known to provide tighter bounds when attempting to characterize tail-probabilities [57], [58].

A fundamental implication of Theorem 4.2 is that it helps guide the choice for $\hat{\alpha}$ depending on the privacy parameters chosen by the stakeholders which also influence their privacy loss. Usually, stakeholders have an established level of significance α depending on local detection benchmarks and domain expertise. Therefore, estimating the function $(\mathcal{E}_I^{max})^{-1}(\alpha) = \hat{\alpha}$ provides an equivalent DP level of significance as a function of a pre-existing α. The inverse function estimation ensures that the Type-I error of the DP hypothesis test is upper bounded by the non-DP test.

As a consequence, stakeholders can choose a perturbed level of significance $\hat{\alpha}$ such that $\alpha = \mathcal{E}_I^{\max}(\hat{\alpha})$ which can be shared with the regulator. This choice of $\hat{\alpha}$ ensures that the regulator can reconstruct the hypothesis testing workflow using privacy-preserving disclosures of covariance and residuals, while still maintaining a worst-case Type I error rate that does not exceed that of the unperturbed χ^2 test at the stakeholder level. In our framework, given a fixed stakeholder level α, we use the Monte-Carlo simulation method to estimate $\hat{\alpha}$.

We extend Theorem 4.2 to determine bounds on miss-classification rates of the hypothesis tests conducted with and without DP which is presented in Theorem 4.3.

Theorem 4.3. *Given $\hat{\tau}_t^{cov}, \hat{\tau}_t^{res}, \hat{\alpha}$, the miss-classification rates can be given as*

$$\mathbb{P}[\hat{\rho}_t = 0 | \rho_t = 1] \leq \omega_1\Big[1 - F^{\hat{\tau}_t^{cov}}_{\chi^2,p}\Big(\hat{T}_t\Big)\Big] + \omega_2\Big[1 - F^{\hat{\tau}^{cov}_{max,t}}_{\chi^2,p}\Big(\hat{T}_t\Big)\Big]$$

$$\mathbb{P}[\hat{\rho}_t = 1 | \rho_t = 0] \leq F^{\hat{\tau}_t^{cov}}_{\chi^2,p}\Big(\hat{T}_t\Big)\Big[\omega_1 + \omega_2\Big]$$

where,
$\omega_1 = \Big[1 - F^{gamma}_{r_{max},\epsilon_{cov},\Delta_l}(R_t\theta_l)\Big]$, $\omega_2 = \Big[F^{ex}_{\Delta_l,\epsilon_{cov}}(\theta_l)\Big]^p$ *and* $\hat{T}_t = T_{\chi^2,t} + \sigma^2\chi^{2,NC}_{\hat{\alpha}} - \chi^2_\alpha$

Proof of Theorem 4.2 has been presented in Appendix A. Theorem 4.3 formally states the miss-classification rates that can occur with respect to the stakeholder and regulator purely on account of privacy preserving disclosures of covariance and residuals as part of the sequential privacy scheme. Using Theorem 4.3, we can see that when Gamma CDF values decrease and are more sensitive when R_t is small, resulting in tighter bounds on miss-classification rates. On the other hand, with larger R_t values, Gamma CDF increases culminating in lower miss-classification rates.

In addition to Theorem 4.3, we introduce a GDP noise calibration factor $\mu_t = ||\hat{\tau}^{cov}_{max,t}||_2^2/\chi^{2,NC}_{\hat{\alpha}}$ to adjust the GDP noise variance $\sigma_t^2 = \mu_t.\sigma^{min}$. The calibration factor is designed to incentivize more targeted application of noise for computing residual disclosures. This is particularly useful in high residual cases (such as during an attack window) where the DP threshold $\chi^{2,NC}_{\hat{\alpha}}$ might fail to deliver a good detection rate. As a result, we leverage μ_t to improve the power of the test in a dynamic, DP friendly fashion.

5. Privacy Preserving Algorithmic Framework

We delineate our proposed algorithmic framework into two distinct components pertaining to the regulatory bodies and the utility stakeholders. For the utility level component, we focus on the development of a detection framework as well as relevant data disclosures based on residuals observed from the non-linear Kalman Filter model. On the other hand, the regulatory component purely focuses on the verification aspects based on disclosed data. Specifically, we consider a set of $j \in \mathcal{J}$ utilities, where $|\mathcal{J}| = J$. We divide the time horizon into discrete time steps denoted by t that yield a distinct observation of sensor measurements as well as its

associated residual at each utility. Further, we group these time steps into sets of evaluation epochs w, with each epoch consisting of W consecutive, discrete time steps.

As a consequence of the guarantees derived in Section 3, we can derive two distinct implementation modes of our algorithmic framework pertaining to critical region based compliance and p-value driven verification. *It is important to note that both these implementation modalities are mutually exclusive and must be pre-determined with consensus among utilities and regulators*. We present the algorithmic framework for each implementation mode.

5.1. Critical Region Based Verification

For the critical region (CR) compliance verification, we assume that the sole regulatory objective is to verify alarms with respect to differentially private disclosures of residuals and covariance from utilities. Alarm verification can be done by the disclosure of the test statistic by the utility followed by the critical region threshold.

5.1.1. CR based Utility Level Detection

The CR based utility level detection framework can be described on the basis of Algorithm 1. In Algorithm 1, at each time step the utility observes residual values r_t and C_t. This is followed by the computation of the aggregated residual r_w and covariance matrix S_w and the epoch alarm ρ_w for the evaluation epoch w. Based on these quantities, the utility can compute DP driven disclosures of $\hat{\tau}_w^{res}$ and $(\hat{S}_w)^{-1/2}$. Finally, the utility transmits an information tuple Π_w consisting of the perturbed covariance matrix $(\hat{S}_w)$ as well as the transformed DP perturbed residual $\hat{\tau}_w^{rg}$, the critical region threshold $\chi_{\hat{\alpha}}^{2,NC}$ and the detected alarm ρ_w with the regulator. In addition, using the concept of post-processing immunity and composition [39], we can state that the disclosure of $\hat{\tau}_w^{rg}$ preserves (ϵ, δ) privacy as well.

Algorithm 1 Utility Level CR Verification Algorithm

for w=0,1,2,... **do**
  for t=0,1,2,...W **do**
    observe r_t and C_t using NLKF model $\mathcal{K}$
  end for
  compute $r_w = \sum_{t=0}^{W} r_t$ and $S_w = \sum_{t=0}^{W} S_t$
  compute ρ_w based on Equations (13)
  compute $(\hat{S}_w)^{-1/2}$ using Equation (16)
  compute $\hat{\tau}^{cov}$ using Equation (17)
  compute $\hat{\tau}_w^{res} = \hat{\tau}_w^{cov} + e_w$ using Equation (21)
  compute $\hat{\tau}_w^{rg} = r_w + (\hat{S}_w)^{1/2} e_w$ using $\hat{\tau}_w^{res}$
  compute $\hat{\alpha}, \chi_{\hat{\alpha}}^{2,NC}$ using Theorem 4.2.
  transmit $\Pi_w = \left[\hat{S}_w, \hat{\tau}_w^{rg}, \chi_{\hat{\alpha}}^{2,NC}, \rho_w\right]$ with regulator.
end for

5.1.2. CR based Regulatory Level Verification

Algorithm 2 captures the sequence of steps taken by a regulator for critical region based verification. At the regulatory level, we consider the set of utilities given by $\mathcal{J}$. The regulator receives DP-based information tuple Π_w^j for each utility $j \in \mathcal{J}$ at each evaluation epoch. Using information contained in Π_w^j, the regulator can compute the factorization of the DP based covariance matrix $\hat{C}^{-1/2}$ as well as obtain an estimate of the DP-driven test statistic $\hat{T}_w^{j,res}$. The regulator can compute an alarm depending on the value of the test statistic $\hat{T}_w^{j,res}$ and the value of the CR threshold $\chi_{\hat{\alpha}}^{j,2,NC}$ based on the conditions given by (26).

$$\hat{\rho}_w^j = \begin{cases} 1, & \text{if } \hat{T}_w^{j,res} > \chi_{\hat{\alpha}}^{j,2,NC}, \\ 0, & \text{otherwise.} \end{cases} \tag{26}$$

As a consequence of Algorithm 2, the regulator independently obtains an estimate of $\hat{\rho}_w^j$ which can be compared with the reported alarm ρ_w^j. The miss-classification rates pertaining to $\hat{\rho}_w^j$ and ρ_w^j is provided using Theorem 4.3. Additionally, we provide a probabilistic bound on the worst case privacy loss incurred as a consequence of the CR based verification mode captured in Algorithms 2 and 1 in Theorem 5.1.

Algorithm 2 Regulator Level CR Verification Algorithm

for w=0,1,2,... **do**
  for j=0,1,2 ...J **do**
    receive Π_w^j from utility stakeholder j
    factorize $\hat{S}_w^j = \hat{V}_w^j \hat{\lambda}_w^j (\hat{V}_w^j)^T$ compute $(\hat{S}_w^j)^{-1/2}$
    compute $\hat{\tau}_w^{j,res} = (\hat{S}_w^j)^{-1/2} \hat{\tau}_w^{j,rg}$
    compute $\hat{T}_w^{j,res} = ||\hat{\tau}_w^{j,res}||_2^2$
    compute $\hat{\rho}_w^j$ based on Equation (26)
    verify if $\hat{\rho}_w^j = \rho_w^j$
  end for
end for

Theorem 5.1. *The disclosure of $\hat{\tau}_w^{rg}$ incurs a worst case privacy loss ϵ' with the following probabilistic bounds*

$$\mathbb{P}\Big[\epsilon' \geq \mathcal{L}(\Delta_r, \sigma^2, \hat{S}_w)\Big] \leq 1 - (1 - \gamma_r)^p$$

where $\mathcal{L}(\Delta_r, \sigma^2, \hat{S}_w) = \frac{\Delta_r}{\sigma^2} (\mathbf{1}^T \hat{S}_w^{-1} \mathbf{1})^2 \left(\frac{\theta_r^2}{p} + \frac{1}{2 \mathbf{1}^T \hat{S}_w^{-1} \mathbf{1}} \right)$

Theorem 5.1 tells us that if GDP failure risk is low, as indicated by $\gamma_r \to 0$, the likelihood of the overall privacy loss exceeding $\mathcal{L}(\Delta_r, \sigma^2, \hat{S}_w)$ will be negligible. Additionally, lower values of l_2 sensitivity of residuals (denoted by Δ_r) and high GDP noise (denoted by variance σ^2) minimize the lower bound on the worst case privacy loss. Lastly, $\mathbf{1}^T \hat{S}_w^{-1} \mathbf{1}$ can be viewed as a scaled Rayleigh quotient for $\hat{S}_w^{-1}$ computed using the vector $\mathbf{1}$. We can observe generally that increasing covariance privacy noise, characterized by increasing $\Delta_l / \epsilon_{cov}$, leads to a diminished value of the scaled Rayleigh quotient implying a lower worst case privacy loss.

5.2. P-value Based Compliance

For the P-value (PV) compliance verification, the regulatory objective is to ensure that alarms have been computed with the correct p-value at the utility level. In this case, the utility discloses the test statistic distribution parameters as well as the DP-equivalent level of significance for verification of alarms at the regulator level.

5.2.1. PV based Utility Level Detection

To facilitate PV compliance, the utility level detection algorithm is represented by Algorithm 3. Similar to a CR driven setting, the utility computes the values of $\hat{\tau}_w^{cov}, \hat{\tau}_w^{res}, \hat{\alpha}$ and generates an alarm ρ_w. It discloses the information tuple Π_w consisting of the values $\hat{T}_w^{res}, \hat{T}_w^{cov}, \hat{\alpha}_w, \rho_w$ to the regulator.

Algorithm 3 Utility Level PV Verification Algorithm

for w=0,1,2,... **do**
  for t=0,1,2,...W **do**
    observe r_t and C_t using NLKF model $\mathcal{K}$
  end for
  compute $r_w = \sum_{t=0}^{W} r_t$ and $S_w = \sum_{t=0}^{W} C_t$
  compute ρ_w based on Equations (13)
  compute $(\hat{S}_w)^{-1/2}$ using Equation (16)
  compute $\hat{\tau}_w^{cov}$ using Equation (17)
  compute $\hat{\tau}_w^{res} = \hat{\tau}_w^{cov} + e_w$ using Equation (21)
  compute $\hat{T}_w^{res} = ||\hat{\tau}_w^{res}/\sigma||_2^2$ and $\hat{T}_w^{cov} = ||\hat{\tau}_w^{cov}/\sigma||_2^2$
  transmit $\Pi_w = \left[\hat{T}_w^{res}, \hat{T}_w^{cov}, \hat{\alpha}_w, \rho_w\right]$ to regulator.
end for

5.2.2. PV based Regulator Level Detection

The regulator level algorithm for PV compliance verification is given in Algorithm 4. The objective of the regulator in this case is to estimate the non-central chi-square distribution using $\hat{T}_w^{j,cov}$ as the centrality parameter according to 4.3 which can be used to estimate $\chi_{\hat{\alpha}_w}^{j,2,NC}$. On the basis of the alarm condition represented by Equation (26), the regulator can independently obtain and validate the alarm $\hat{\rho}_w^j$ with respect to ρ_w^j for each utility. In the

Algorithm 4 Regulator Level PV Verification Algorithm

for w=0,1,2,... **do**
  for j=0,1,2 ...J **do**
    receive Π_w^j from utility stakeholder j
    obtain $\hat{\alpha}_w$, $\hat{T}_w^{j,res}$ and $\hat{T}_w^{j,cov}$ from Π_w^j
    use $\hat{T}_w^{j,cov}$ to compute $\chi_{\hat{\alpha}_w}^{j,2,NC}$ using Theorem 4.2.
    compute $\hat{\rho}_w^j$ based on Equation (26)
    verify if $\hat{\rho}_w^j = \rho_w^j$
  end for
end for

PV implementation mode the regulator has access to the parametrized probability distribution of the DP test statistic $\hat{T}_w^{j,res}$ denoted by the non-central chi-squared distribution with centrality parameter $\hat{T}_w^{j,cov}$. This enables the regulator to obtain p-value of the DP test statistic based on the perturbed level of significance $\hat{\alpha}$ divulged by the utility. Additionally, we can derive bounds on the privacy loss incurred through the disclosure of $\hat{T}_w^{j,cov}$ in Theorem 5.2.

Theorem 5.2. *The disclosure of $\hat{T}_w^{cov}$ results in an (ϵ', δ') DP mechanism where*

$$\begin{aligned}\epsilon' \geq & \epsilon_{cov} + \frac{\Delta_r^T C^{-1} \Delta_r}{2\sigma^2} \\ \delta' \leq & \Phi\Big(\frac{\sigma^2(\epsilon' - \epsilon_{cov})}{||\Delta_r^T C^{-1}||} - \frac{\Delta_r^T C^{-1}\Delta_r}{2||\Delta_r^T C^{-1}||}\Big) \\ & - \Phi\Big(-\frac{\sigma^2(\epsilon' - \epsilon_{cov})}{||\Delta_r^T C^{-1}||} + \frac{\Delta_r^T C^{-1}\Delta_r}{2||\Delta_r^T C^{-1}||}\Big)\end{aligned} \tag{27}$$

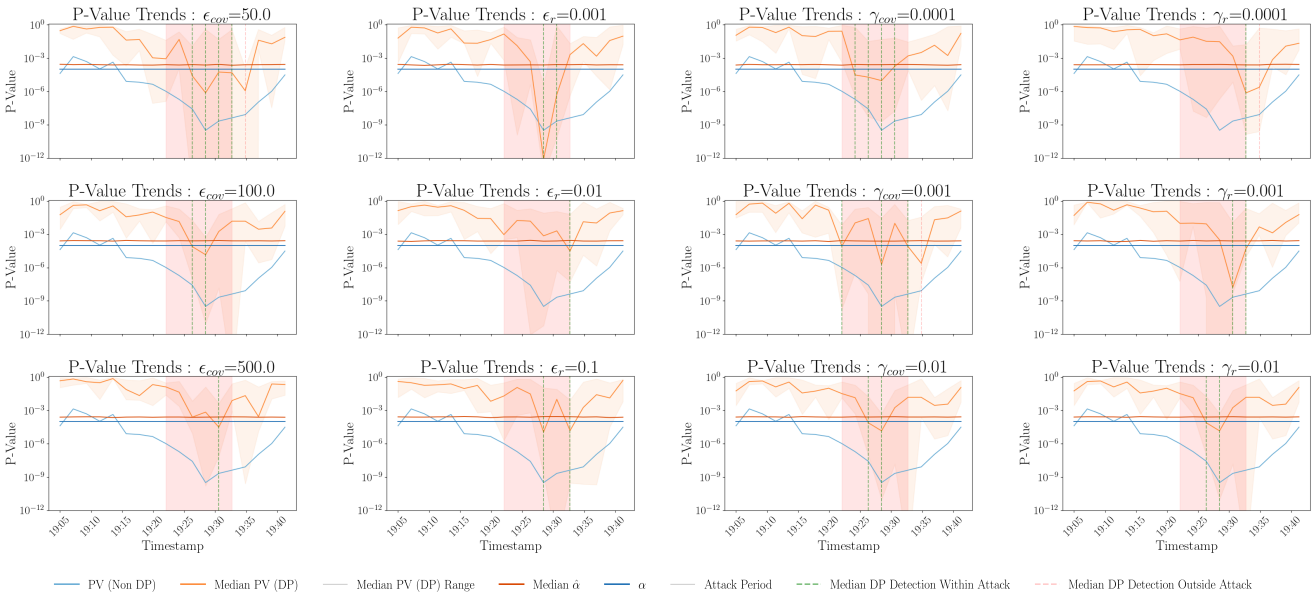
where Φ denotes the CDF of $N(0, \Delta_r^T C^{-1} \Delta_r)$

Theorem 5.2 provides several insights into the privacy implications regarding the disclosure of $\hat{T}_w^{cov}$. First, we observe that with higher GDP covariance (σ^2) and lower sensitivity (Δ_r) individually contribute to ϵ' making it closer to ϵ_{cov}. Additionally, as lower bound on ϵ' approaches ϵ_{cov}, we can also observe that the privacy failure probability δ' also tends towards 0. These observations imply that a higher GDP noise covariance and lower sensitivities while disclosing $\hat{T}_w^{cov}$ increasingly tends towards an $(\epsilon_{cov}, \gamma_{cov})$-DP paradigm.

6. Experimental Results

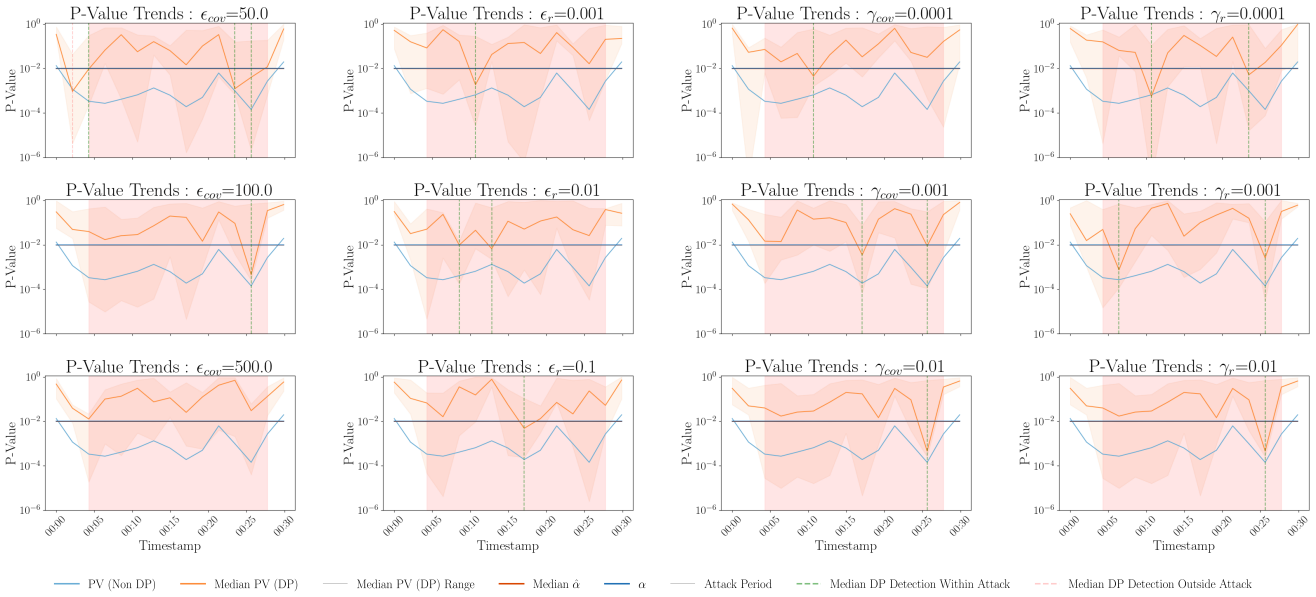
Dataset and Detection Models: For our experiments, we primarily leverage the HAI Security dataset [17], as well as the ORNL power system (ORNL-PS) attack dataset [33], [59]. For both datasets, we trained an NLKF model as described in Section 3.2 using the corresponding state and sensor variables pertaining to each dataset. In order to yield a well-formed non-linear, extended Kalman Filter model for the HAI dataset, we utilized the multi-level LSTM framework provided in [60]. Our experimental strategy revolves around evaluating the variations pertaining to DP failure probabilities γ_{cov}, γ_r, the privacy budgets $\epsilon_{cov}, \epsilon_r$ respectively.

System Implementation Details: All experiments were carried out on a virtual machine (VM) running Ubuntu 24.04 with 100GB of RAM and 16 vCPUs using `Python 3.11` with the detection model inference using `PyTorch 2.7.1`. For evaluating diverse aspects of our proposed framework, we utilized a native as well as a container based execution environment. The native setup was primarily used for evaluating the performance of our framework under various scenarios of differential privacy. The container based execution environment was used to evaluate the distinct implementation modes pertaining to Critical Region Verification (CRV) and the P-Value Compliance (PVC). Specifically, we generated container images representative of the regulator and the utility that replays snippets of the HAI and ORNL-PS datasets under scenarios of attack. In order to evaluate system performance of CRV and PVC mode, we created container images representative of the regulator and the utility that replays attack scenarios of HAI and ORNL-PS datasets. The regulator container service hosts a REST API developed using `Flask` that receives DP-driven disclosures from the corresponding utility container service for each implementation mode and executes the corresponding compliance steps as provided in Algorithms 2 and 4. Con-



(a) ϵ_{cov} (b) ϵ_r (c) γ_{cov} (d) γ_r

Figure 3: HAI Dataset: P-Value trends for varying DP parameter values



(a) ϵ_{cov} (b) ϵ_r (c) γ_{cov} (d) γ_r

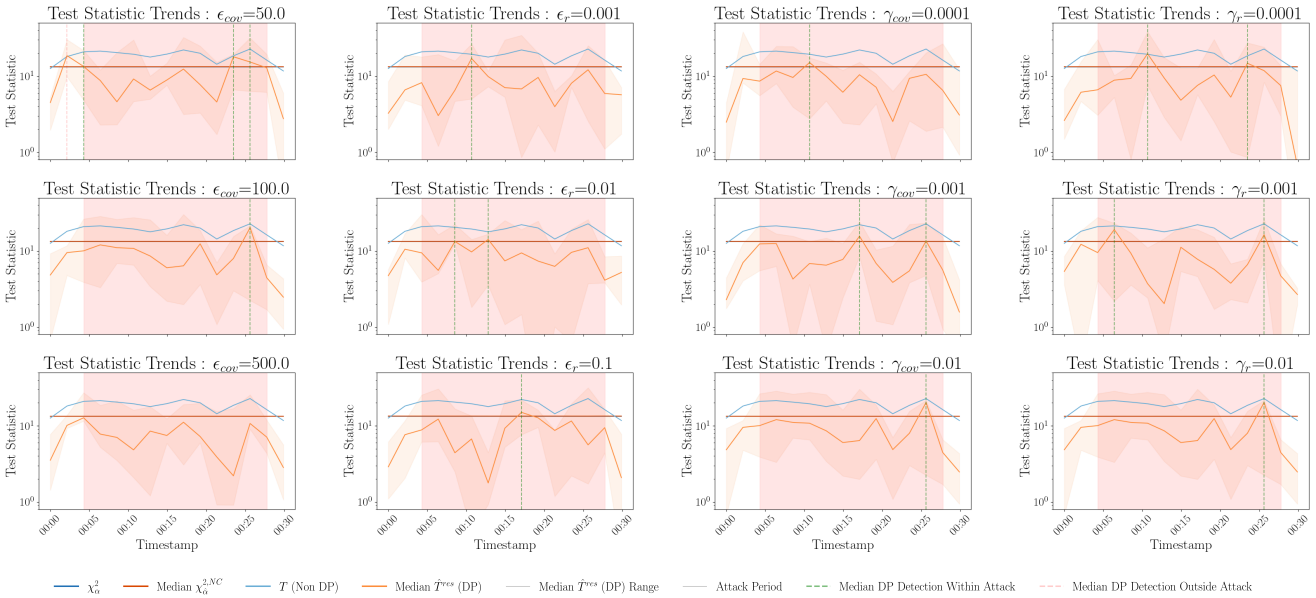
Figure 4: ORNL-PS Dataset: P-Value trends for varying DP parameter values

sequently, we measure the system performance in terms of the CPU utilization and memory consumption of both utility and regulator services under both implementation modes. Docker based quick start scripts and the associated code have been provided as part of the accompanying artifacts to our paper.

DP parameter choice: In our experiments $\frac{\Delta_l}{\epsilon_{cov}}$ denotes the scale value for the Laplacian distribution used for covariance disclosures. For all our experiments we utilize $\sigma = \frac{\Delta_r}{\epsilon_r}\sqrt{2.ln\left(\frac{1.25}{\gamma_r}\right)}$. For the HAI and the ORNL-PS datasets, we used $\Delta_r = 50$, and $\Delta_l = 0.1$. These values were chosen based on rigorous empirical analysis of the maximum 2-norm residual values observed as well as the

(a) ϵ_{cov} (b) ϵ_r (c) γ_{cov} (d) γ_r

Figure 5: HAI Dataset: Test statistic trends for varying DP parameter values



(a) ϵ_{cov} (b) ϵ_r (c) γ_{cov} (d) γ_r

Figure 6: ORNL-PS Dataset: Test statistic trends for varying DP parameter values

maximum eigenvalue square roots observed in the covariance matrices.

Visualization strategy: We present graphs depicting trends that include a *min-max* band comprising the minimum and maximum values of the desired quantity observed over five consecutive runs. The *min-max* encapsulate the median values observed across five independent runs for each experiment. To accommodate the non-linear nature of the model and complexities in the HAI data, we consider a 40 minute window pertaining to the `ap_05` attack scenario. In order to account for the nonlinear nature of the detection model, we designate the first 800 secs from the start `ap_05` as the attack window for the HAI dataset. Since the attack duration in the ORNL-PS datasets exceeds 800 seconds,

TABLE 1: HAI Dataset: Alignment of DP and Non-DP Detection Within Attack Window.

DP Params	DP & Non-DP Detection			Only Non-DP Detection			DP Alignment Rate			Mean $\hat{\alpha}$ (Variance) in $1e-4$ ($1e-10$)		
	200s	400s	600s	200s	400s	600s	200s	400s	600s	200s	400s	600s
$\epsilon_{cov} = 100$	11	26	44	39	24	6	0.22	0.52	0.88	2.56 (4.55)	2.56 (4.55)	2.56 (4.55)
$\epsilon_r = 1e-3$	13	29	46	37	21	4	0.26	0.58	0.92	2.56 (4.5)	2.56 (4.5)	2.56 (4.5)
$\gamma_{cov} = 1e-2$	8	25	44	42	25	6	0.16	0.5	0.88	2.57 (4.6)	2.57 (4.6)	2.57 (4.6)
$\gamma_r = 1e-2$	11	27	45	39	23	5	0.22	0.54	0.9	2.56 (4.39)	2.56 (4.39)	2.56 (4.39)

TABLE 2: ORNL-PS Dataset: Alignment of DP and Non-DP Detection Within Attack Window

DP Params	DP & Non-DP Detection			Only Non-DP Detection			DP Alignment Rate			Mean $\hat{\alpha}$ (Variance) in $1e-4$ ($1e-6$)		
	200s	400s	600s	200s	400s	600s	200s	400s	600s	200s	400s	600s
$\epsilon_{cov} = 100$	28	45	48	22	5	2	0.56	0.90	0.96	1.014 (1.0)	1.014 (1.0)	1.014 (1.0)
$\epsilon_r = 1e-3$	28	45	47	22	5	3	0.56	0.90	0.94	0.976 (5.8)	0.976 (5.8)	0.976 (5.8)
$\gamma_{cov} = 1e-2$	25	42	46	25	8	4	0.5	0.84	0.92	1.014 (1.0)	1.014 (1.0)	1.014 (1.0)
$\gamma_r = 1e-2$	27	44	47	23	6	3	0.54	0.88	0.94	1.014 (1.0)	1.014 (1.0)	1.014 (1.0)

TABLE 3: DP False Alarm Rates Outside Attack Window

Dataset	ϵ_{cov}=100	ϵ_r=1e-3	γ_{cov}=1e-2	γ_r=1e-2
HAI	0.0767	0.0833	0.0783	0.08
ORNL-PS	0.04	0.068	0.053	0.047

we restrict ourselves to results for the first 57 minutes from the beginning of the dataset. Red vertical lines depict false alarms with respect to the median obtained from five independent runs. Green vertical lines reflect the correct detection of an attack.

6.1. Analyzing DP Covariance Disclosures

We begin by analyzing the effects of differentially private disclosures of covariance matrices. Our analysis focuses on both HAI and ORNL-PS datasets specifically examining impacts of ϵ_{cov} and γ_{cov} on the p-value and the test statistic in comparison to the non-DP scenarios.

6.2. Effect of Privacy Budget

We begin by analyzing the p-values of both datasets under varying covariance privacy budget values. Figures 3(a), 4(a) reflect the p-value trends pertaining to HAI and ORNL-PS datasets respectively. The corresponding test statistic trends are depicted in Figures 5(a) and 6(a) respectively.

We note that increasing values of ϵ_{cov} correspond to lower DP noise and a higher privacy budget as well. However, Figures 3(a), 4(a), 5(a) and 6(a) also indicate that with higher privacy budget, there is a much higher variation observed in both p-value and test statistic trends. In other words, with a higher privacy budget, we can observe much less variation in detection quality. Additionally, the consistency of alarms is higher in median terms during the attack period although this trend also results in some false alarms from the median as well. In Figure 4(a), we observe that for $\epsilon_{cov} = 500$ the median does not breach the p-value threshold, although the trends make it clear that the attack detection is still robust with far lesser variations observed.

6.2.1. Effect of Failure Probability

We turn our attention to the effect of failure probability observed when disclosing covariance values. Figures 3(c), 4(c) represent trends in the p-value for HAI and ORNL-PS datasets, while Figures 5(c), 6(c) represent trends in test statistic. γ_{cov} conventionally represents the DP failure probability for covariance disclosures. Ultimately, a higher γ_{cov} represents a higher likelihood of DP being ineffective at hiding. From Figures 3(c), 4(c), 5(c) and 6(c), we observe that with increasing values of γ_{cov}, we again see more variance and in p-value and test statistic trends. However, we can also observe that the median detection quality results in consistent alarms during the attack window albeit with a higher variance. In summary, the framework is capable of providing attack detection with realistic privacy expectations even in cases of high γ_{cov} values, indicating the usefulness of our approach.

6.3. Analysis for DP Residual Disclosures

We now analyze the impact of differentially private residual disclosures in terms of the privacy budget ϵ_r and failure probability γ_r.

6.3.1. Effect of Privacy Budget

We plot the p-value trends for varying values of ϵ_r for both datasets in Figures 3(b) and 4(b) respectively. Similarly, we present test statistic trends for both datasets in 5(b) and 6(b) respectively as well. In general, we see that the spread of values as indicated by the min-max spread increases with decreasing values of privacy budget ϵ_r. This results in a slight increase in median alarm consistency for higher values of ϵ_r as well. Overall, even with a high value of privacy budget, our framework provides a significant privacy guarantee.

6.3.2. Effect of Failure Probability

We present the performance of the DP detection framework in terms of the residual disclosure failure probability γ_r in terms of the p-value and test statistics. For both datasets, we can see that the attack detection performance

TABLE 4: Average Memory Usage (MB) and CPU Utilization (%) for varying implementation modes

Dataset	Implementation Mode	Average Memory (MB) (std dev.)		CPU Utilization (%) (std dev.)	
		Utility	Regulator	Utility	Regulator
ORNL-PS	Critical Region Verification	1014.58 (224.85)	128.96 (0.27)	10.27 (12.34)	10.55 (13.61)
	P-Value Compliance	1013.86 (223.18)	128.29 (0.08)	10.09 (12.07)	9.91 (11.48)
HAI	Critical Region Verification	1093.21 (221.24)	128.80 (0.29)	9.87 (11.15)	9.93 (11.33)
	P-Value Compliance	1101.33 (271.88)	128.02 (0.07)	10.97 (13.38)	10.47 (12.29)

remains robust with steady consistency of the median alarm detection rates as well.

6.4. Analyzing Attack Detection Quality

In this subsection, we analyze the performance of the combined DP attack detection framework considering the residual based DP disclosures in conjunction with the DP based covariance matrices. The combined analysis is meant to provide vital insights into the latency of attack detection at the regulator level with respect to the local utility stakeholder.

Therefore, we examine the ability of the DP based framework to detect the attack within three distinct intervals (200s, 400s and 600s) measured from the start of the attack. The results of these experiments are obtained from five independent runs carried out for each DP parameter combination listed in Tables 1 and 2. We are primarily interested in tracking the alignment of DP and Non-DP detection at the discrete interval values of 200s, 400s and 600s. The alignment problem can be thought of as run instances wherein both DP and Non-DP frameworks successfully detected an attack. Therefore, we present results in terms of alignment of DP and Non-DP Detection, Non-DP only detection, DP alignment rates as well as the mean and variance of the $\hat{\alpha}$.

In Tables 1 and 2 we see a consistent improvement in the alignment rate with increase in duration from the beginning of the attack. The alignment rate measured as a fraction of runs where an attack was detected within the attack window consistently improves from around 0.5 to 0.95. This is powered by the rising number of DP and Non-DP detection instances that is naturally accompanied by falling Non-DP only detection. For both datasets, we see that the mean $\hat{\alpha}$ stays relatively stable at $2.56e-4$ and $1.01e-4$ respectively. The associated variance experiences minor volatility but overall retains stability around $4.55e{-}10$ and $1e-10$ for HAI and ORNL-PS datasets respectively.

In Table 3, we present the false alarm rates of the DP based detection framework for both datasets captured during normal operations. We can see that for all the considered combinations of DP parameters, the false alarm rates stay consistently low. For the HAI dataset, the stability of false alarm rates hover around 8%, while for ORNL-PS dataset, this value exhibits slightly more volatility, ranging from 4% to around 6.8%. Collectively, Tables 1, 2 and 3 demonstrate the robustness of the attack detection quality with respect to the miss-classification and alignment rates of the DP mechanism.

6.5. Analyzing System Performance of Implementation Modes

In Table 4, we provide a comparative analysis of system performance of the CRV and PVC implementation modes with respect to HAI and ORNL-PS datasets. Average CPU utilization consistently stays under 11% for all scenarios with standard deviation ranging from 11.15% to 13.6%. Similarly, memory usage for utility service ranges between 1014 MB and 1101 MB, while remaining very close to 128 MB for the regulator service. From Table 4, we also see that the standard deviation values of CRV memory usage is relatively higher for both datasets. This rise can be explained on the basis of the need to factorize the DP-driven covariance matrix by the regulator in the CRV mode as opposed to a simple compliance check needed in the PVC mode. Overall, Table 4 demonstrates that the system performance in terms of both average memory consumption as well as CPU utilization remain consistent and stable across both datasets and implementation modes.

7. Conclusion

In this paper, we present a differentially private algorithmic framework geared towards regulatory compliance for detecting data-driven attacks in industrial control systems for critical infrastructure networks. Our proposed method leverages statistical tests on residuals arising out of state space modeling at the utility stakeholder level to raise attack alarms. We focus on cases wherein utilities are interested in convincing regulatory bodies regarding the veracity of their respective alarms by disclosing differential privacy induced covariance matrices and residual values. As a result, our proposed framework revolves around a two phase privacy scheme that sequentially perturbs covariance using Laplacian noise followed by a Gaussian differential privacy scheme for residuals. We derive strong privacy guarantees pertaining to the test of residuals in addition to providing tight bounds on the miss-classification rates of alarms as well as equivalent levels of significance. We specifically explore two significant modalities of implementation concerning critical region and p-value based compliance schemes. Additionally, we theoretically characterize the privacy implications of each of the modalities. Using real-world ICS datasets, we characterize the performance of our algorithm with respect to varying privacy parameters under diverse attack scenarios. The experimental results demonstrate that our framework is capable of matching the performance of the non-DP versions in almost all cases while preserving the privacy of utility stakeholders.

References

[1] E. Bompard, D. Wu, and F. Xue, "Structural vulnerability of power systems: A topological approach," *Electric Power Systems Research*, vol. 81, no. 7, pp. 1334–1340, 2011.

[2] J. D. Smith, "Cybersecurity for the operational technology environment (cyote)," tech. rep., Idaho National Lab.(INL), Idaho Falls, ID (United States), 2023.

[3] J. V. Cuffari, "Cisa made progress but resources, staffing, and technology challenges hinder cyber threat detection and mitigation." https://www.oig.dhs.gov/sites/default/files/assets/2023-03/OIG-23-19-Mar23.pdf, 2023.

[4] "Rd&i needs and strategic actions for resilience of critical infrastructure." https://www.cisa.gov/sites/default/files/2023-05/rdi_for_resilience_of_cyber-phys_critical_infrastructure_needs_strategic_actions_508c.pdf, 2023.

[5] J. V. Cuffari, "Additional progress needed to improve information sharing under the cybersecurity act of 2015." https://www.oig.dhs.gov/sites/default/files/assets/2022-08/OIG-22-59-Aug22.pdf, 2022.

[6] A. Nolan, *Cybersecurity and information sharing: Legal challenges and solutions*, vol. 5. Congressional Research Service, 2015.

[7] T. W. Mak, F. Fioretto, L. Shi, and P. Van Hentenryck, "Privacy-preserving power system obfuscation: A bilevel optimization approach," *IEEE Transactions on Power Systems*, 2019.

[8] F. Fioretto, L. Mitridati, and P. Van Hentenryck, "Ppsm: A privacy-preserving stackelberg mechanism," *Nuclear Physics, Section A*, 2019.

[9] F. Fioretto, T. W. Mak, and P. Van Hentenryck, "Privacy-preserving obfuscation of critical infrastructure networks," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 1086–1092, 2019.

[10] L. O. Nweke and S. Wolthusen, "Legal issues related to cyber threat information sharing among private entities for critical infrastructure protection," in *2020 12th International Conference on Cyber Conflict (CyCon)*, vol. 1300, pp. 63–78, IEEE, 2020.

[11] C. Johnson, L. Badger, D. Waltermire, J. Snyder, C. Skorupka, *et al.*, "Guide to cyber threat information sharing," *NIST special publication*, vol. 800, no. 150, 2016.

[12] D. Li, N. Gebraeel, and K. Paynabar, "Detection and differentiation of replay attack and equipment faults in scada systems," *IEEE Transactions on Automation Science and Engineering*, vol. 18, no. 4, pp. 1626–1639, 2020.

[13] D. Li, P. Ramanan, N. Gebraeel, and K. Paynabar, "Deep learning based covert attack identification for industrial control systems," *arXiv preprint arXiv:2009.12360*, 2020.

[14] D. Li, N. Gebraeel, K. Paynabar, and A. S. Meliopoulos, "An online approach to covert attack detection and identification in power systems," *IEEE Transactions on Power Systems*, vol. 38, no. 1, pp. 267–277, 2022.

[15] S. Couch, Z. Kazan, K. Shi, A. Bray, and A. Groce, "Differentially private nonparametric hypothesis testing," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 737–751, 2019.

[16] M. Gaboardi and R. Rogers, "Local private hypothesis testing: Chi-square tests," in *International Conference on Machine Learning*, pp. 1626–1635, PMLR, 2018.

[17] . J.-H. Y. B.-G. M. Hyeok-Ki Shin; Woomyo Lee; Seungoh Choi, "Hai security datasets," 2023.

[18] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, and J. Srivastava, "A comparative study of anomaly detection schemes in network intrusion detection," in *Proceedings of the 2003 SIAM International Conference on Data Mining*, pp. 25–36, SIAM, 2003.

[19] M. Caselli, E. Zambon, J. Amann, R. Sommer, and F. Kargl, "Specification mining for intrusion detection in networked control systems.," in *USENIX Security Symposium*, pp. 791–806, 2016.

[20] N. Ye, Y. Zhang, and C. M. Borror, "Robustness of the markov-chain model for cyber-attack detection," *IEEE Transactions on Reliability*, vol. 53, no. 1, pp. 116–123, 2004.

[21] D. I. Urbina, J. A. Giraldo, A. A. Cardenas, N. O. Tippenhauer, J. Valente, M. Faisal, J. Ruths, R. Candell, and H. Sandberg, "Limiting the impact of stealthy attacks on industrial control systems," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1092–1105, ACM, 2016.

[22] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Transactions on Information and System Security (TISSEC)*, vol. 14, no. 1, p. 13, 2011.

[23] G. Dan and H. Sandberg, "Stealth attacks and protection schemes for state estimators in power systems," in *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*, pp. 214–219, IEEE, 2010.

[24] A. Teixeira, S. Amin, H. Sandberg, K. H. Johansson, and S. S. Sastry, "Cyber security analysis of state estimators in electric power systems," in *49th IEEE Conference on Decision and Control (CDC). Atlanta, GA. DEC 15-17, 2010*, pp. 5991–5998, 2010.

[25] Z.-H. Yu and W.-L. Chin, "Blind false data injection attack using pca approximation method in smart grid," *IEEE Transactions on Smart Grid*, vol. 6, no. 3, pp. 1219–1226, 2015.

[26] A. A. Cárdenas, S. Amin, Z.-S. Lin, Y.-L. Huang, C.-Y. Huang, and S. Sastry, "Attacks against process control systems: risk assessment, detection, and response," in *Proceedings of the 6th ACM symposium on information, computer and communications security*, pp. 355–366, ACM, 2011.

[27] G. Chaojun, P. Jirutitijaroen, and M. Motani, "Detecting false data injection attacks in ac state estimation," *IEEE Transactions on Smart Grid*, vol. 6, no. 5, pp. 2476–2483, 2015.

[28] Y. Mo, R. Chabukswar, and B. Sinopoli, "Detecting integrity attacks on scada systems," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 4, pp. 1396–1407, 2014.

[29] T. Huang, B. Satchidanandan, P. Kumar, and L. Xie, "An online detection framework for cyber attacks on automatic generation control," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 6816–6827, 2018.

[30] D. Van Long, L. Fillatre, and I. Nikiforov, "Sequential monitoring of scada systems against cyber/physical attacks," *IFAC Papers Online*, vol. 48, no. 21, pp. 746–753, 2015.

[31] A. Hoehn and P. Zhang, "Detection of covert attacks and zero dynamics attacks in cyber-physical systems," in *American Control Conference (ACC), 2016*, pp. 302–307, IEEE, 2016.

[32] S. Weerakkody and B. Sinopoli, "Detecting integrity attacks on control systems using a moving target approach," in *Decision and Control (CDC), 2015 IEEE 54th Annual Conference on*, pp. 5820–5826, IEEE, 2015.

[33] S. Pan, T. Morris, and U. Adhikari, "Classification of disturbances and cyber-attacks in power systems using heterogeneous time-synchronized data," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 3, pp. 650–662, 2015.

[34] M. S. Rahman, M. A. Mahmud, A. M. T. Oo, and H. R. Pota, "Multi-agent approach for enhancing security of protection schemes in cyber-physical energy systems," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 2, pp. 436–447, 2017.

[35] D. Li, K. Paynabar, and N. Gebraeel, "A degradation-based detection framework against covert cyberattacks on scada systems," *IISE Transactions*, vol. 53, no. 7, pp. 812–829, 2021.

[36] P. Ramanan, D. Li, and N. Gebraeel, "Blockchain-based decentralized replay attack detection for large-scale power systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 8, pp. 4727–4739, 2021.

[37] T. Irita and T. Namerikawa, "Detection of replay attack on smart grid with code signal and bargaining game," in *2017 American Control Conference (ACC)*, pp. 2112–2117, IEEE, 2017.

[38] D. Li, N. Gebraeel, and K. Paynabar, "Detection and differentiation of replay attack and equipment faults in scada systems," *IEEE Transactions on Automation Science and Engineering*, 2020.

[39] C. Dwork, A. Roth, *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[40] J. Cortés, G. E. Dullerud, S. Han, J. Le Ny, S. Mitra, and G. J. Pappas, "Differential privacy in control and network systems," in *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 4252–4272, IEEE, 2016.

[41] C. Zhang and Y. Wang, "Enabling privacy-preservation in decentralized optimization," *IEEE Transactions on Control of Network Systems*, vol. 6, no. 2, pp. 679–689, 2018.

[42] X. Cao, J. Zhang, H. V. Poor, and Z. Tian, "Differentially private admm for regularized consensus optimization," *IEEE Transactions on Automatic Control*, 2020.

[43] F. Zhou, J. Anderson, and S. H. Low, "Differential privacy of aggregated dc optimal power flow data," in *2019 American Control Conference (ACC)*, pp. 1307–1314, IEEE, 2019.

[44] F. Fioretto and P. Van Hentenryck, "Constrained-based differential privacy: Releasing optimal power flow benchmarks privately," in *International Conference on the Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, pp. 215–231, Springer, 2018.

[45] F. Fioretto, T. W. Mak, and P. Van Hentenryck, "Differential privacy for power grid obfuscation," *IEEE Transactions on Smart Grid*, 2019.

[46] T. W. Mak, F. Fioretto, and P. Van Hentenryck, "Privacy-preserving obfuscation for distributed power systems," *Electric Power Systems Research*, vol. 189, p. 106718, 2020.

[47] V. Dvorkin, P. Van Hentenryck, J. Kazempour, and P. Pinson, "Differentially private distributed optimal power flow," in *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 2092–2097, IEEE, 2020.

[48] J. Le Ny and J. Le Ny, "Differentially private kalman filtering," *Differential Privacy for Dynamic Data*, pp. 55–75, 2020.

[49] Y. Song, C. X. Wang, and W. P. Tay, "Privacy-aware kalman filtering," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4434–4438, IEEE, 2018.

[50] K. H. Degue and J. Le Ny, "On differentially private kalman filtering," in *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 487–491, IEEE, 2017.

[51] 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*, pp. 2111–2120, PMLR, 2016.

[52] R. Rogers and D. Kifer, "A new class of private chi-square hypothesis tests," in *Artificial Intelligence and Statistics*, pp. 991–1000, PMLR, 2017.

[53] D. Alabi and S. Vadhan, "Hypothesis testing for differentially private linear regression," *Advances in Neural Information Processing Systems*, vol. 35, pp. 14196–14209, 2022.

[54] R. S. Smith, "A decoupled feedback structure for covertly appropriating networked control systems," *IFAC Proceedings Volumes*, vol. 44, no. 1, pp. 90–95, 2011.

[55] W.-Y. Hwang, "Chi-square quantile-based multivariate variance monitoring for individual observations," *Communications in Statistics-Simulation and Computation*, vol. 46, no. 7, pp. 5392–5409, 2017.

[56] K. Amin, T. Dick, A. Kulesza, A. Munoz, and S. Vassilvitskii, "Differentially private covariance estimation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[57] C. Dwork, G. N. Rothblum, and S. Vadhan, "Boosting and differential privacy," in *2010 IEEE 51st annual symposium on foundations of computer science*, pp. 51–60, IEEE, 2010.

[58] B. Balle, G. Barthe, M. Gaboardi, J. Hsu, and T. Sato, "Hypothesis testing interpretations and renyi differential privacy," in *International Conference on Artificial Intelligence and Statistics*, pp. 2496–2506, PMLR, 2020.

[59] S. Pan, T. Morris, and U. Adhikari, "Developing a hybrid intrusion detection system using data mining for power systems," *IEEE Transactions on Smart Grid*, vol. 6, no. 6, pp. 3104–3113, 2015.

[60] H. Coskun, F. Achilles, R. DiPietro, N. Navab, and F. Tombari, "Long short-term memory kalman filters: Recurrent neural estimators for pose regularization," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5524–5532, 2017.

[61] Y. Sun, Á. Baricz, and S. Zhou, "On the monotonicity, log-concavity, and tight bounds of the generalized marcum and nuttall q-functions," *IEEE Transactions on Information Theory*, vol. 56, no. 3, pp. 1166–1186, 2010.

Appendix

Proof of Lemma4.1

Proof. We begin by stating Equation (28) which results from the orthonormal factorization of the real and perturbed covariance matrices S_t and $\hat{S}_t$ respectively as defined in Section 3.4.

$$|\tau_t^T \tau_t - (\hat{\tau}^{cov})_t^T \hat{\tau}_t^{cov}| = |r_t^T(\Lambda_t - \hat{\Lambda}_t) r_t| \tag{28}$$

We know that Equation (29) holds when $\hat{\lambda}^{PCA}, \lambda^{PCA}$ denote the diagonal matrices of eigenvalues corresponding to the top $p \leq d$ principal components with and without DP respectively

$$|r_t^T(\Lambda - \hat{\Lambda}) r_t| = \Big| \sum_{i=1}^{p} (r_t^i)^2 (\lambda_i - \hat{\lambda}_i) \Big| \leq E_l \Big| \sum_{i=1}^{p} (r_t^i)^2 \Big| \tag{29}$$

We know that Equation (30) holds with a probability of at least $1 - \gamma_{cov}$

$$E_l \Big| \sum_{i=1}^{p} (r_t^i)^2 \Big| \leq \Big| \sum_{i=1}^{p} (r_t^i)^2 \Big| \Big[\frac{\Delta_l}{\epsilon_{cov}} log\Big(\frac{d}{\gamma_{cov}} \Big) \Big] \tag{30}$$

By combining Equations (29), (30) with Equation (19), we get Equation (31) which holds with probability of at least $1 - \gamma_{cov}$ completing the proof.

$$|\hat{T}_{\chi^2,t}^{cov} - T_{\chi^2,t}| \leq \Big| \sum_{i=1}^{p} (r_t^i)^2 \Big| \Big[\frac{\Delta_l}{\epsilon_{cov}} log\Big(\frac{d}{\gamma_{cov}} \Big) \Big] \tag{31}$$

□

Proof of Lemma 4.2

Proof. As a result of Definition 3, we obtain an (ϵ, γ_r)-DP mechanism such that Equation (32) applies.

$$\mathbb{P}\Big(\Big| \sum_{i=1}^{p} e_{i,t} \Big| \geq \theta_r \Big) \leq \gamma_r, \text{ where, } \theta_r = \frac{\sigma^2 \epsilon}{\Delta} - \frac{p\Delta}{2} \tag{32}$$

Since $\sum_{i=1}^{p}|e_{i,t}| \geq |\sum_{i=1}^{p} e_{i,t}|$, we can also derive the following relations

$$\mathbb{P}\Big(E_r \leq \frac{\theta_r}{p}\Big) \geq 1-\gamma_r, \text{ where, } E_r = \max_{0\leq i\leq p}|e_{i,t}| \quad (33)$$

Equation (33) ensures that the probability of the maximum absolute value of Gaussian noise being less than θ_r/p is at least $1-\gamma_r$. Using Equation (33) we can bound E_r^2 as well.

$$\mathbb{P}\Big(0 \leq E_r^2 \leq \frac{\theta_r^2}{p^2}\Big) \geq 1-\gamma_r \quad (34)$$

Since each of the elements of e_t are independently and identically distributed Equation (35) holds

$$\mathbb{P}\Big(0 \leq \sum_{i=1}^{p} e_{i,t}^2 \leq p.E_r^2 \leq \frac{\theta_r^2}{p}\Big) \geq (1-\gamma_r)^p \quad (35)$$

Setting $||e_t||_2^2 = \sum_{i=1}^{p} e_{i,t}^2$ in Equation (35) completes the proof. □

Proof of Lemma 4.3

Proof. We know that, $\hat{T}^{res}_{\chi^2,t} = \sum_{i=1}^{p}(e_{i,t}+\tau_{i,t})^2$. From GDP conditions, $(e_{i,t}+\tau_{i,t}) \sim N(\tau_{i,t},\sigma^2)$. As a result, $[(e_{i,t}+\tau_{i,t})^2/\sigma^2] \sim \chi^2(1,\tau_{i,t}^2/\sigma^2)$ which implies that the variance scaled perturbed test statistic $\hat{T}^{res}_{\chi^2,t}/\sigma^2 \sim \chi^2(p,||\tau_t||_2^2/\sigma^2)$. □

Proof of Theorem 4.1

Proof. We reformulate each element of $\hat{T}^{res}_{\chi^2,t} - T_{\chi^2,t}$ as

$$\begin{aligned}\hat{T}^{res}_{\chi^2,t} - T_{\chi^2,t} &= \sum_{i=1}^{p} e_{i,t}(e_{i,t}+2\tau_{i,t}) \\ &= 2\sum_{i=1}^{p}(e_{i,t}\tau_{i,t}) + \sum_{i=1}^{p}(e_{i,t})^2\end{aligned} \quad (36)$$

Using Lemma 1, we can assert that $e_{i,t} \leq E_r \leq \frac{\theta_r}{p}$ and $\sum_{i=1}^{p}(e_{i,t})^2 \leq \frac{\theta_r^2}{p}$ with a probability of at least $(1-\gamma_r)^p$. Therefore, combining Equation (36) with Lemma 1, we get

$$2\sum_{i=1}^{p}(e_{i,t}\tau_{i,t}) + \sum_{i=1}^{p}(e_{i,t})^2 \leq \frac{\theta_r}{p}\Big(\theta_r + 2\sum_{i=1}^{p}\tau_{i,t}\Big) \quad (37)$$

$$2\sum_{i=1}^{p}(e_{i,t}\tau_{i,t}) + \sum_{i=1}^{p}(e_{i,t})^2 \geq \frac{\theta_r}{p}\Big(\theta_r - 2\sum_{i=1}^{p}\tau_{i,t}\Big) \quad (38)$$

Using Lemma 2, for $U' = \frac{\theta_r}{p}\Big(\theta_r + 2\sum_{i=1}^{p}\tau_{i,t}\Big)$, $L' = \frac{\theta_r}{p}\Big(\theta_r - 2\sum_{i=1}^{p}\tau_{i,t}\Big)$ we can also assert that

$$\mathbb{P}\Big[\frac{1}{\sigma^2}(\hat{T}^{res}_{\chi^2,t} - T_{\chi^2,t}) \leq \frac{U'}{\sigma^2}\Big|||e_t||_2^2 \leq \frac{\theta_r^2}{p}\Big] = F^{\tau_t}_{\chi^2,p}(\frac{U'}{\sigma^2}) \quad (39)$$

$$\mathbb{P}\Big[\frac{1}{\sigma^2}(\hat{T}^{res}_{\chi^2,t} - T_{\chi^2,t}) \geq \frac{L'}{\sigma^2}\Big|||e_t||_2^2 \leq \frac{\theta_r^2}{p}\Big] = 1 - F^{\tau_t}_{\chi^2,p}(\frac{L'}{\sigma^2}) \quad (40)$$

Computing joint probability using Lemma 1 and setting $U = U'/\sigma^2$ and $L = L'/\sigma^2$ completes the proof

$$\begin{aligned}\mathbb{P}\Big[L \leq (\hat{T}^{res}_{\chi^2,t} - T_{\chi^2,t}) \leq U\Big|||e_t||_2^2 \leq \frac{\theta_r^2}{p}\Big] \\ \geq (F^{\tau_t}_{\chi^2,p}(U) - F^{\tau_t}_{\chi^2,p}(L))(1-\gamma_r)^p\end{aligned}$$

□

Proof of Lemma 4.4

Proof. We know that

$$\mathbb{P}\Big[\hat{T}^{res}_{\chi^2,t} > \phi\Big] = 1 - F^{\tau}_{\chi^2,p}(\phi) \quad (41)$$

The CDF of $\chi^2(k,\mu)$ can be stated as $F^{\tau_t}_{\chi^2,p}(\phi) = 1 - Q_{\frac{k}{2}}(\sqrt{\mu},\sqrt{\phi})$ where $Q_a(b,c)$ is the generalized Marcum Q-function. The generalized Marcum Q-function is strictly increasing in μ [61]. As a result,

$$\max_{\tau}(1 - F^{\tau}_{\chi^2,p}(\phi)) = \max_{\tau}\ Q_{\frac{p}{2}}\Big(\frac{||\tau||_2}{\sigma}, \sqrt{\phi}\Big) \quad (42)$$

As a result, given $\tau_{min} = \underset{\tau}{argmin}(||\tau||_2)$ and $\tau_{max} = \underset{\tau}{argmax}(||\tau||_2)$

$$\tau_{max} = \underset{\tau}{argmax}\ (1 - F^{\tau}_{\chi^2,p}(\phi)) \quad (43)$$

□

Proof of Theorem 4.2

Proof. We consider two cases
Case1: When $|\hat{T}^{cov}_{\chi^2,t} - T_{\chi^2,t}| \geq R_t\theta_l$, we know that

$$\begin{aligned}\mathbb{P}\Big[\hat{T}^{res}_{\chi^2,t} > \sigma^2\chi^{2,NC}_{\hat{\alpha}},\ |\hat{T}^{cov}_{\chi^2,t} - T_{\chi^2,t}| \geq R_t\theta_l\Big] = \\ \mathbb{P}\Big[|\hat{T}^{cov}_{\chi^2,t} - T_{\chi^2,t}| \geq R_t\theta_l\Big]\Big[1 - F^{\hat{\tau}^{cov}_t}_{\chi^2,p}\Big(\sigma^2\chi^{2,NC}_{\hat{\alpha}}\Big)\Big]\end{aligned} \quad (44)$$

Using Lemma 4.1, we can state that

$$\mathbb{P}\Big[|\hat{T}^{cov}_{\chi^2,t} - T_{\chi^2,t}| \geq R_t\theta_l\Big] = \mathbb{P}\Big[|r_t^T(\lambda_{t,i} - \hat{\lambda}_{t,i})r_t| \geq R_t\theta_l\Big] \quad (45)$$

We note that using Lemma 4.1, we can know that since $|\hat{T}^{cov}_{\chi^2,t} - T_{\chi^2,t}| = |r_t^T(\Lambda_t - \hat{\Lambda}_t)r_t|$, we can state that using

$$r_{max} = \underset{r}{argmax}||r||$$

$$|r_t^T(\Lambda_t - \hat{\Lambda}_t)r_t| \leq |\sum_{i=1}^{p}(\lambda_t - \hat{\lambda}_{t,i})r_{t,i}^2| \tag{46}$$

$$|\sum_{i=1}^{p}(\lambda_t - \hat{\lambda}_{t,i})r_{t,i}^2| \leq \sum_{i=1}^{p}|(\lambda_t - \hat{\lambda}_{t,i})|||r_{max}||_2^2 \tag{47}$$

As a result, we know that $(\lambda_t - \hat{\lambda}_{t,i}) \sim Lap(0, \frac{\Delta_l}{\epsilon_{cov}})$, which implies that $r_{max}^2|(\lambda_t - \hat{\lambda}_{t,i})| \sim Exp(\frac{\epsilon_{cov}}{\Delta_l r_{max}^2})$ leading to

$$\Big[\sum_{p}|(\lambda_{t,i} - \hat{\lambda}_{t,i})|\Big]||r_{max}||_2^2 \sim Gamma(p, \frac{\epsilon_{cov}}{\Delta_l r_{max}^2}) \tag{48}$$

Since, we know that $|r_t^T(\Lambda_t - \hat{\Lambda}_t)r_t| = |\hat{T}_{\chi^2,t}^{cov} - T_{\chi^2,t}|$, we can state that, $\Big[\sum_{p}|(\lambda_{t,i} - \hat{\lambda}_{t,i})|\Big]||r_{max}||_2^2 \geq |r_t^T(\Lambda_t - \hat{\Lambda}_t)r_t|$. Probabilistically, this leads us to,

$$\mathbb{P}\Big[\Big(\sum_{p}|(\lambda_{t,i} - \hat{\lambda}_{t,i})|\Big)||r_{max}||_2^2 \geq |\hat{T}_{\chi^2,t}^{cov} - T_{\chi^2,t}| \geq R_t\theta_l\Big] \leq 1 - F_{r_{max},\epsilon_{cov},\Delta_l}^{gamma}(R_t\theta_l) \tag{49}$$

Therefore we get Equation (44), where $F_{r_{max},\epsilon_{cov},\Delta_l}^{gamma}(.)$ represents the CDF of $Gamma(p, \frac{\epsilon_{cov}}{\Delta_l r_{max}^2})$

$$\mathbb{P}\Big[\hat{T}_{\chi^2,t}^{res} > \sigma^2\chi_{\hat{\alpha}}^{2,NC},\ |\hat{T}_{\chi^2,t}^{cov} - T_{\chi^2,t}| \geq R_t\theta_l\Big] \leq \Big[1 - F_{r_{max},\epsilon_{cov},\Delta_l}^{gamma}(R_t\theta_l)\Big]\Big[1 - F_{\chi^2,p}^{\hat{\tau}_t^{cov}}\Big(\sigma^2\chi_{\hat{\alpha}}^{2,NC}\Big)\Big] \tag{50}$$

Case2: For the case when $|\hat{T}_{\chi^2,t}^{cov} - T| \leq R_t\theta_l$ Using Lemma 4.1, we can state that

$$\mathbb{P}\Big[|\hat{T}_{\chi^2,t}^{cov} - T_{\chi^2,t}| \leq R_t\theta_l\Big] = \mathbb{P}\Big[|r_t^T(\lambda_{t,i} - \hat{\lambda}_{t,i})r_t| \leq R_t\theta_l\Big] \tag{51}$$

We note that $|r_t^T(\lambda_{t,i} - \hat{\lambda}_{t,i})r_t| \leq R_t\theta_l = r_t^T diag(\theta_l)r_t$ holds for all r. This implies that $|(\lambda_{t,i} - \hat{\lambda}_{t,i})| \leq \theta_l$, and as a result,

$$\mathbb{P}\Big[|r_t^T(\Lambda_t - \hat{\Lambda}_t)r_t| \leq R_t\theta_l\Big] = \mathbb{P}\Big[|\Lambda_t - \hat{\Lambda}_t| \leq \theta_l\Big] \tag{52}$$

We also know that $|\Lambda_t - \hat{\Lambda}_t| \sim Exp(\frac{\epsilon_{cov}}{\Delta_l})$, Therefore $\mathbb{P}\Big[|\Lambda_t - \hat{\Lambda}_t| \leq \theta_l\Big] = [F_{\Delta_l,\epsilon_{cov}}^{ex}(\theta_l)]^p$, where $F_{\Delta_l,\epsilon_{cov}}^{ex}(.)$ is the CDF of the exponential distribution. Therefore, we can state that,

$$\mathbb{P}\Big[\hat{T}_{\chi^2,t}^{res} > \sigma^2\chi_{\hat{\alpha}}^{2,NC},\ |\hat{T}_{\chi^2,t}^{cov} - T_{\chi^2,t}| \leq R_t\theta_l\Big] = \Big[F_{\Delta_l,\epsilon_{cov}}^{ex}(\theta_l)\Big]^p\Big[1 - F_{\chi^2,p}^{\hat{\tau}_t^{cov}}\Big(\sigma^2\chi_{\hat{\alpha}}^{2,NC}\Big)\Big] \tag{53}$$

Using Lemma 4.4, we get

$$\mathbb{P}\Big[\hat{T}_{\chi^2,t}^{res} > \sigma^2\chi_{\hat{\alpha}}^{2,NC},\ |\hat{T}_{\chi^2,t}^{cov} - T_{\chi^2,t}| \leq R_t\theta_l\Big] \leq \Big[F_{\Delta_l,\epsilon_{cov}}^{ex}(\theta_l)\Big]^p\Big[1 - F_{\chi^2,p}^{\hat{\tau}_{max,t}^{cov}}\Big(\sigma^2\chi_{\hat{\alpha}}^{2,NC}\Big)\Big] \tag{54}$$

□

Proof of Theorem 4.3

Proof. We know that $\hat{\rho}_t = 1$ is triggered when $\hat{T}_{\chi,t}^{res} > \sigma^2\chi_{\hat{\alpha}}^{2,NC}$, and $\rho_t = 0$ when $T_{\chi^2,t} < \chi_\alpha^2$. Therefore,

$$\hat{T}_{\chi^2,t}^{res} - T_{\chi^2,t} > \sigma^2\chi_{\hat{\alpha}}^{2,NC} - \chi_\alpha^2 \tag{55}$$

$$\hat{T}_{\chi^2,t}^{res} > T_{\chi^2,t} + \sigma^2\chi_{\hat{\alpha}}^{2,NC} - \chi_\alpha^2 \tag{56}$$

As a result, we can state that

$$\mathbb{P}[\hat{\rho}_t = 1|\rho_t = 0, \hat{\tau}_{\chi^2,t}^{cov}] = \mathbb{P}[\hat{T}_{\chi^2,t}^{res} > T_{\chi^2,t} + \sigma^2\chi_{\hat{\alpha}}^{2,NC} - \chi_\alpha^2] \tag{57}$$

$$\mathbb{P}[\hat{\rho}_t = 1|\rho_t = 0, \hat{\tau}_{\chi^2,t}^{cov}] = 1 - F_{\chi^2,p}^{\hat{\tau}_t^{cov}}\Big(T_{\chi^2,t} + \sigma^2\chi_{\hat{\alpha}}^{2,NC} - \chi_\alpha^2\Big) \tag{58}$$

We can use the same technique for upper bounding as employed in Cases 1 and 2 in Theorem 4.2 in Equations (44) and (53) for the event $\hat{T}_{\chi^2,t}^{res} > T_{\chi^2,t} + \sigma^2\chi_{\hat{\alpha}}^{2,NC} - \chi_\alpha^2$. This leads us to Equations (59) and (60).

$$\mathbb{P}\Big[\hat{T}_{\chi^2,t}^{res} > T_{\chi^2,t} + \sigma^2\chi_{\hat{\alpha}}^{2,NC} - \chi_\alpha^2, |\hat{T}_{\chi^2,t}^{cov} - T_{\chi^2,t}| \geq R_t\theta_l\Big] \leq \Big[1 - F_{r_{max},\epsilon_{cov},\Delta_l}^{gamma}(R_t\theta_l)\Big]\Big[1 - F_{\chi^2,p}^{\hat{\tau}_t^{cov}}\Big(T_{\chi^2,t} + \sigma^2\chi_{\hat{\alpha}}^{2,NC} - \chi_\alpha^2\Big)\Big] \tag{59}$$

$$\mathbb{P}\Big[\hat{T}_{\chi^2,t}^{res} > T_{\chi^2,t} + \sigma^2\chi_{\hat{\alpha}}^{2,NC} - \chi_\alpha^2, |\hat{T}_{\chi^2,t}^{cov} - T_{\chi^2,t}| \leq R_t\theta_l\Big] \leq \Big[F_{\Delta_l,\epsilon_{cov}}^{ex}(\theta_l)\Big]^p\Big[1 - F_{\chi^2,p}^{\hat{\tau}_{max,t}^{cov}}\Big(T_{\chi^2,t} + \sigma^2\chi_{\hat{\alpha}}^{2,NC} - \chi_\alpha^2\Big)\Big] \tag{60}$$

Summing up Equations (59) and (60) leads us to the following where $\hat{T}_t = T_{\chi^2,t} + \sigma^2\chi_{\hat{\alpha}}^{2,NC} - \chi_\alpha^2$

$$\mathbb{P}[\hat{\rho}_t = 1|\rho_t = 0] \leq \Big[1 - F_{r_{max},\epsilon_{cov},\Delta_l}^{gamma}(R_t\theta_l)\Big]\Big[1 - F_{\chi^2,p}^{\hat{\tau}_t^{cov}}\Big(\hat{T}_t\Big)\Big] + \Big[F_{\Delta_l,\epsilon_{cov}}^{ex}(\theta_l)\Big]^p\Big[1 - F_{\chi^2,p}^{\hat{\tau}_{max,t}^{cov}}\Big(\hat{T}_t\Big)\Big] \tag{61}$$

Similarly, $\hat{\rho}_t = 0$ when $\hat{T}_{\chi,t}^{res} < \sigma^2\chi_{\hat{\alpha}}^{2,NC}$, and $\rho_t = 1$ when $T_{\chi^2,t} > \chi_\alpha^2$. Therefore,

$$\hat{T}_{\chi^2,t}^{res} - T_{\chi^2,t} < \sigma^2\chi_{\hat{\alpha}}^{2,NC} - \chi_\alpha^2 \tag{62}$$

$$\hat{T}_{\chi^2,t}^{res} < T_{\chi^2,t} + \sigma^2\chi_{\hat{\alpha}}^{2,NC} - \chi_\alpha^2 \tag{63}$$

In a similar fashion as Case 1, we can obtain,

$$\mathbb{P}[\hat{\rho}_t = 0|\rho_t = 1] = \mathbb{P}[\hat{T}_{\chi^2,t}^{res} < \hat{T}_t] = F_{\chi^2,p}^{\hat{\tau}_t^{cov}}\Big(\hat{T}_t\Big) \tag{64}$$

Therefore, we obtain Equations (65) and (66)

$$\mathbb{P}\Big[\hat{T}_{\chi^2,t}^{res} < \hat{T}_t,\ |\hat{T}_{\chi^2,t}^{cov} - T_{\chi^2,t}| \leq R_t\theta_l\Big] \leq \Big[F_{\Delta_l,\epsilon_{cov}}^{ex}(\theta_l)\Big]^p F_{\chi^2,p}^{\hat{\tau}_t^{cov}}\Big(\hat{T}_t\Big) \tag{65}$$

$$\mathbb{P}\Big[\hat{T}_{\chi^2,t}^{res} < \hat{T}_t,\ |\hat{T}_{\chi^2,t}^{cov} - T_{\chi^2,t}| \geq R_t\theta_l\Big] \leq \Big[1 - F_{r_{max},\epsilon_{cov},\Delta_l}^{gamma}(R_t\theta_l)\Big]F_{\chi^2,p}^{\hat{\tau}_t^{cov}}\Big(\hat{T}_t\Big) \tag{66}$$

Combining both equations leads us to

$$\mathbb{P}[\hat{\rho}_t = 0|\rho_t = 1] = \mathbb{P}\Big[\hat{T}^{res}_{\chi^2,t} < \hat{T}_t\Big] \leq F^{\hat{T}^{cov}_t}_{\chi^2,p}\Big(\hat{T}_t\Big)\Big(\Big[1 - F^{gamma}_{r_{max},\epsilon_{cov},\Delta_l}(R_t\theta_l)\Big] + \Big[F^{ex}_{\Delta_l,\epsilon_{cov}}(\theta_l)\Big]^p\Big) \tag{67}$$

□

Proof of Theorem 5.1

Proof. Based on the definition of ϵ-DP, we know that the effective residual noise is given by $\hat{C}_w^{1/2}e_w \sim N(0, \sigma^2\hat{S}_w)$

$$\Big|ln\Big[\frac{exp[-e_w^T\hat{C}_w^{-1}e_w/(2\sigma^2)]}{exp[-(e_w+\Delta_r)^T\hat{C}_w^{-1}(e_w+\Delta_r)/(2\sigma^2)]}\Big]\Big| \leq \epsilon' \tag{68}$$

Consolidating terms we obtain

$$\Big|\frac{\Delta_r(2.\mathbf{1}^T\hat{V}_w^T\hat{\Lambda}_w\hat{V}_we_w + \Delta_r\mathbf{1}^T\hat{S}_w\mathbf{1})}{2\sigma^2}\Big| \leq \epsilon' \tag{69}$$

$$\Big|\mathbf{1}^T\hat{V}_w^T\hat{\Lambda}_w\hat{V}_we_w + \frac{\Delta_r\mathbf{1}^T\hat{S}_w\mathbf{1}}{2}\Big| \leq \frac{\sigma^2\epsilon'}{\Delta_r} \tag{70}$$

Examining the LHS of Equation (70), we can see that

$$|\mathbf{1}^T\hat{V}_w^T\hat{\Lambda}_w\hat{V}_we_w| \leq ||\mathbf{1}^T\hat{V}_w^T\hat{\Lambda}_w||_2^2||\hat{V}_we_w||_2^2 \leq (\mathbf{1}^T\hat{C}_w^{-1}\mathbf{1})^2||e_w||^2 \tag{71}$$

This implies that

$$\Big|\mathbf{1}^T\hat{V}_w^T\hat{\Lambda}_w\hat{V}_we_w + \frac{\Delta_r\mathbf{1}^T\hat{S}_w\mathbf{1}}{2}\Big| \leq |\mathbf{1}^T\hat{V}_w^T\hat{\Lambda}_w\hat{V}_we_w| + \frac{\Delta_r\mathbf{1}^T\hat{C}_w^{-1}\mathbf{1}}{2} \tag{72}$$

From Equation (72), we obtain,

$$|\mathbf{1}^T\hat{V}_w^T\hat{\Lambda}_w\hat{V}_we_w| + \frac{\Delta_r\mathbf{1}^T\hat{C}_w^{-1}\mathbf{1}}{2} \leq (\mathbf{1}^T\hat{C}_w^{-1}\mathbf{1})^2\Big(||e_w||^2 + \frac{1}{2.\mathbf{1}^T\hat{C}_w^{-1}\mathbf{1}}\Big) \tag{73}$$

Characterizing the worst case (maximizing) ϵ' results in obtaining the upper bound

$$\frac{\sigma^2\epsilon'}{\Delta_r} \geq (\mathbf{1}^T\hat{C}_w^{-1}\mathbf{1})^2\Big(||e_w||^2 + \frac{1}{2.\mathbf{1}^T\hat{C}_w^{-1}\mathbf{1}}\Big) \tag{74}$$

$$\implies \epsilon' \geq \frac{\Delta_r}{\sigma^2}(\mathbf{1}^T\hat{C}_w^{-1}\mathbf{1})^2\Big(||e_w||^2 + \frac{1}{2.\mathbf{1}^T\hat{C}_w^{-1}\mathbf{1}}\Big) \tag{75}$$

Using Lemma 2, we know that the probability that $||e_w||_2^2 \geq \frac{\theta_r^2}{p}$ occurs is at most $1-(1-\gamma_r)^p$. Therefore, under conditions of Lemma 2, we can see that further maximizing ϵ' leads to a probabilistic lower bound of

$$\mathbb{P}\Big[\epsilon' \geq \frac{\Delta_r}{\sigma^2}(\mathbf{1}^T\hat{C}_w^{-1}\mathbf{1})^2\Big(\frac{\theta_r^2}{p} + \frac{1}{2.\mathbf{1}^T\hat{C}_w^{-1}\mathbf{1}}\Big)\Big] \leq 1-(1-\gamma_r)^p \tag{76}$$

□

Proof of Theorem 5.2

Proof. We consider the probability that T_w^{cov} is obtained with a residual r_w^1 and DP-based eigenvalue matrix given by Λ_w^1. To establish differential privacy, we must now consider the probability of obtaining T_w^{cov} when r_w^2 is realized and Λ_w^2 is the corresponding DP-based eigenvalue matrix. Using the definitions of DP, under conditions that r_w^1, r_w^2 and Λ_w^1, Λ_w^2 are adjacent pairs of realizations. Therefore without loss of generality we can state that:

$$ln\Big|\frac{exp(\frac{-r^TC^{-1}r}{2}).\mathbb{P}(\Lambda_w^1)}{exp(\frac{-(r+\Delta)^TC^{-1}(r+\Delta)}{2}).\mathbb{P}(\Lambda_w^2))}\Big| \leq \epsilon' \tag{77}$$

Equation (77) results in the following relationships

$$\Big|ln\Big(exp\Big[\frac{-2\Delta^TC^{-1}r + \Delta^TC^{-1}\Delta}{2\sigma^2}\Big].exp(\epsilon)\Big)\Big| \leq \epsilon' \tag{78}$$

$$\Big|\frac{-2\Delta^TC^{-1}r + \Delta^TC^{-1}\Delta}{2\sigma^2} + \epsilon\Big| \leq \epsilon' \tag{79}$$

We can now use ϵ' to bound the RHS such that the following holds

$$||r|| \leq \frac{\sigma^2(\epsilon'-\epsilon)}{||\Delta^TC^{-1}||} - \frac{\Delta^TC^{-1}\Delta}{2||\Delta^TC^{-1}||} \tag{80}$$

Since $r \sim N(0, C)$, we can state that $\Delta^TC^{-1}r \sim N(0, \Delta^TC^{-1}\Delta)$. As a consequence we obtain the following relationship where Φ is the CDF for $N(0, \Delta^TC^{-1}\Delta)$

$$\mathbb{P}\Big[||r|| \leq \frac{\sigma^2(\epsilon'-\epsilon)}{||\Delta^TC^{-1}||} - \frac{\Delta^TC^{-1}\Delta}{2||\Delta^TC^{-1}||}\Big] = \Phi\Big(\frac{\sigma^2(\epsilon'-\epsilon)}{||\Delta^TC^{-1}||} - \frac{\Delta^TC^{-1}\Delta}{2||\Delta^TC^{-1}||}\Big) - \Phi\Big(-\frac{\sigma^2(\epsilon'-\epsilon)}{||\Delta^TC^{-1}||} + \frac{\Delta^TC^{-1}\Delta}{2||\Delta^TC^{-1}||}\Big) \tag{81}$$

To ensure that privacy loss associated with disclosure of $\hat{T}_w^{cov}$ is bounded by ϵ' we would need a δ' such that

$$\delta' \leq \Phi\Big(\frac{\sigma^2(\epsilon'-\epsilon)}{||\Delta^TC^{-1}||} - \frac{\Delta^TC^{-1}\Delta}{2||\Delta^TC^{-1}||}\Big) - \Phi\Big(-\frac{\sigma^2(\epsilon'-\epsilon)}{||\Delta^TC^{-1}||} + \frac{\Delta^TC^{-1}\Delta}{2||\Delta^TC^{-1}||}\Big) \tag{82}$$

□