Adversarial Classification Under Differential Privacy

Jairo Giraldo
University of Utah
jairo.giraldo@utah.edu

Alvaro A. Cardenas
UC Santa Cruz
alvaro.cardenas@ucsc.edu

Murat Kantarcioglu
UT Dallas
muratk@utdallas.edu

Jonathan Katz
George Mason University
jkatz2@gmail.com

***Abstract*—The last decade has seen a growing interest in *adversarial classification*, where an attacker tries to mislead a classifier meant to detect anomalies. We study this problem in a setting where anomaly detection is being used in conjunction with *differential privacy* to protect personal information. We show that a strategic attacker can leverage the additional noise (introduced to ensure differential privacy) to mislead the classifier beyond what the attacker could do otherwise; we also propose countermeasures against such attacks. We then evaluate the impact of our attacks and defenses in road traffic congestion and smart metering examples.**

I. Introduction

The large-scale collection of user data has enabled a variety of new services, from better online search recommendations, to improved transportation services with crowdsourced vehicle routing applications like Waze. This fine-grained collection of user data provides many benefits to society, but it also raises privacy concerns, and to address these concerns, many techniques have been proposed, including **differential privacy** (DP) [1]. DP adds *noise* to sensitive data, or computations done on sensitive data, in order to ensure privacy while not overly degrading the utility of the data.

Although privacy is an important concern for systems collecting user data, it is unfortunately not the only risk, as widespread vulnerabilities [2], [3], [4] can be exploited by attackers to inject false data into the system. For example, attackers can inject fake data in crowdsourced vehicular services [5], [6], [7], [8] to cause non-existent congestion alarms and accidents that never happened, triggering the services to automatically reroute traffic because of false information.

When the integrity of the data cannot be fully trusted, anomaly detection can provide defense-in-depth solutions to mitigate the impact of false data injection attacks. Anomaly detection is at the heart of several information security problems, including intrusion detection, fraud detection, and detecting false data injected into a system. While the classical problem in anomaly detection assumes that the anomalies are not adaptive or strategic, there is a growing interest to design attack-resilient anomaly detection algorithms. In particular, **adversarial classification** deals with anomaly detection in a setting where the attacker knows the classification algorithm being used and actively tries to avoid detection (while still attacking the system) [9]. Adversarial classification has been explored in a variety of settings, including network intrusion detection [10], host-based intrusion detection [11], [12], misbehavior in wireless networks [13], and detection of false data in cyber-physical systems [14], [15], [16], [17].

Network and Distributed Systems Security (NDSS) Symposium 2020
23-26 February 2020, San Diego, CA, USA
ISBN 1-891562-61-4
https://dx.doi.org/10.14722/ndss.2020.23047
www.ndss-symposium.org

As the previous paragraphs suggest, there is a growing need to deploy DP mechanisms to protect the privacy of individuals, while at the same time, there is also a need to develop anomaly detection algorithms over DP data that are resilient to evasion attacks. In this paper we formulate this problem, and analyze the trade-offs between (1) the utility of the data, (2) the privacy provided by DP, and (3) the security of the anomaly detector against evasion attacks. We then show that DP makes adversarial classification attacks easier, and then show how to design new attack-detection algorithms with better resilience to such attacks.

While prior work has considered trade-offs between privacy and utility in the context of statistical data analysis [18], [19], [20], [21], [22], adversarial classification needs to consider not only the trade-off between three components: (1) **utility** (making accurate estimates with DP data), (2) **privacy** (prevent the identification of personal data from individuals), and (3) **security** (detecting false data injection attacks).

The adversary model in this paper does not seek to violate *privacy*; rather, the adversary exploits privacy mechanisms that are implemented in a system and weaponizes them to degrade the *utility* of the system, while at the same time trying to evade the anomaly detection algorithm that looks for maliciously injected data.

Contributions. To the best of our knowledge, we are the first to (1) formulate the problem of adversarial classification in a system that uses DP to protect the privacy of its users, (2) find *optimal false-data-injection attacks* that degrade the anomaly detection capabilities of the system, while allowing the attacker to remain undetected by "hiding" false data in DP noise, and (3) design *optimal attack-detection defenses* to minimize the impact of such attacks.

The rest of the paper is organized as follows: in section II we give a brief motivation for our problem formulation the adversary model. Our main contributions are then presented in sections III and IV; in section III we prove theorems showing how an attacker can leverage DP to design optimal attacks against an anomaly detection model while remaining stealthy; and in section IV, we show how we can design an optimal defense against the sophisticated attacker introduced in the previous section. In section V we extend our results for time-series data. We then apply our theoretical results to a transportation problem in section VI and to a power grid

problem in section VII. Finally in section VIII, we discuss related work, and in section IX we conclude the paper.

II. Motivation and Problem Statement

We begin with an intuitive overview illustrating how differential privacy can affect anomaly detection. Suppose we have a central entity monitoring a sequence of real values (e.g., sensor readings from a device that measures the number of cars passing on a given segment of a street), $y(0), y(1), y(2), \ldots$. Assume further that, under normal conditions, it will always be the case that $|y(k-1) - y(k)| \leq c$ for some constant c, and so a simple form of anomaly detection is to raise an alarm if this condition is ever violated.

An attacker who is able to modify the data $y(0), y(1) \ldots, y(N)$ sent to the monitor will be able to bias those values and make them, say, artificially larger; however, *if the attacker wishes to remain undetected* it will be limited to biasing each value by at most c relative to the previous value, and in particular can only achieve $y(N) \leq y(0) + c \cdot N$.

Now suppose that the sequence of values is protected using a differentially private mechanism in which noise $\eta(k)$ is added to each value $y(k)$ to obtain a perturbed value $\bar{y}(k)$ that is then sent to the monitor. Assume further that the parameters are such that with 99% probability $|\bar{y}(i) - \bar{y}(i+1)| \leq 1.1c$, and the anomaly-detection algorithm is modified to raise an alarm only if this condition is violated. The attacker can leverage this change in at least two ways. First, it can now bias the values by a larger amount without being detected. Furthermore, the anomaly-detection algorithm now has a nonzero false-positive rate, and the attacker can try to exploit that as well, e.g., by modifying data in a more extreme way that raises the alarm but only a small percentage of the time.

We now provide a formal model for the scenario described above.

A. Problem Statement

The type of systems that we consider in this work are composed of four main elements: 1) a source of data $\mathcal{N}$, 2) an attack-detection algorithm $\mathcal{D}$, 3) a differential privacy mechanism $\mathcal{M}$, and 4) an adversary $\mathcal{A}$.

The randomized **source of data** $\mathcal{N}$ (i.e., nature) generates a value y; we write this as $y \leftarrow \mathcal{N}$. An **adversary** $\mathcal{A}$ that can inject false data into the system (e.g., by hacking an IoT sensor); we write this as $y^a \leftarrow \mathcal{A}(y)$. To detect attacks, an attack-detection algorithm $\mathcal{D}$ takes as input a value y and outputs 1 to raise an alert if it views the data as suspicious, and outputs 0 if it deems the value to be normal. The performance of the classifier against a specific adversary $\mathcal{A}$ is usually characterized by its **false positive rate** $\Pr_{y \leftarrow \mathcal{N}}[\mathcal{D}(y) = 1]$ and **true positive rate** $\Pr_{y^a \leftarrow \mathcal{A}}[\mathcal{D}(y^a) = 1]$.

We assume the attacker wishes to remain undetected (i.e., to output y^a such that $\mathcal{D}(y^a) = 0$) while otherwise ensuring that the injected data is as far as possible from the true data (e.g., $y^a \gg y$ or $y^a \ll y$).

Attack-detection algorithms are usually analyzed without differential privacy, but as differential privacy becomes more prevalent, we need to start studying its effect in adversarial classification problems. We now assume that to protect the privacy of the users, a **differential privacy (DP) mechanism** $\mathcal{M}$ adds randomness to the values y and generates a new value $\bar{y}$ to guarantee (ϵ, δ)-differential privacy (δ can be zero)

$$\bar{y} \leftarrow \mathcal{M}(y, (\epsilon, \delta)).$$

When the DP mechanism is introduced, the attacker $\mathcal{A}$ can leverage the information about the additional structured noise to inject false data into the system:

$$\bar{y}^a \leftarrow \mathcal{A}(\bar{y}, y).$$

The false positive rate is then

$$P^{fa} := \Pr_{y \leftarrow \mathcal{N}; \bar{y} \leftarrow \mathcal{M}(y)}[\mathcal{D}(\bar{y}) = 1],$$

and the probability of detection is

$$P^{d} := \Pr_{y \leftarrow \mathcal{N}; \bar{y}^a \leftarrow \mathcal{A}(y, \bar{y})}[\mathcal{D}(\bar{y}^a) = 1].$$

We want to design attack detection algorithms $\mathcal{D}$ with high true positive rates and low false positive rates, even when faced with a strategic attacker $\mathcal{A}$ trying to evade our classifier.

B. Adversary Model

We consider an attacker that gives false data to consumers of a query that is expected to contain DP noise.

The adversary model for classical DP is one where the attacker is a man-in-the-middle, as shown in Figure 1. This adversary model represents an attacker that has hijacked the secure connection between the database and the client, but has not compromised either end point. For example the attacker can present a fake certificate to the client to become a man-in-the-middle between the database and the client, and the attacker then replaces the query response (which is sent with the added DP noise) with malicious values.

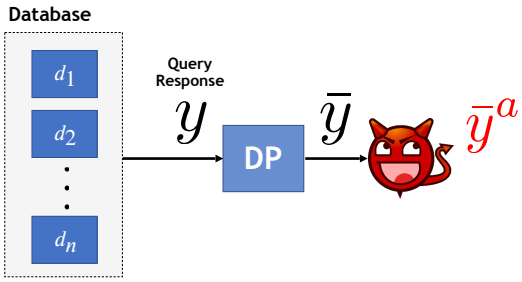

Figure 1. The adversary can be a man-in-the-middle between the source of DP data and the receiver, as illustrated above, or it can compromise one of the sources of information as illustrated in Figure 2.

The adversary model for local DP considers an attacker that has compromised a subset of the information sources (e.g., sensors) as illustrated in Figure 2. In this case the attacker has either compromised a subset of the sensors delivering the data, or in the case of crowdsourced data, the attacker could own a subset of the devices sending false data [8].

We follow the conservative approach preferred in cryptography where to test the security of the algorithms we develop, we consider worst-case attackers. In particular, we give the

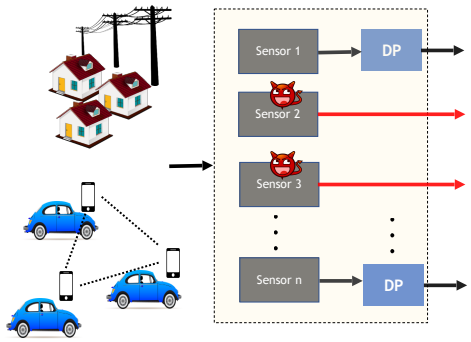

Figure 2. An adversary can compromise a sensor and inject false data that is hard to distinguish from a measurement with DP-noise.

attacker access to the raw data and the DP data to generate the attack ($y^a \leftarrow \mathcal{A}(\bar{y}, y)$). This powerful adversary models the case when the attacker compromises the data source of a distributed DP mechanism, as the attacker can then see the original data, and knows the noise that would be added to it.

In practice, not all attackers will be as powerful; however, as we show later in the paper, our proposed attack-detection algorithm will perform much better against *weaker* adversaries that have less knowledge about the system—i.e., adversaries without access to the raw data, DP parameters, or both, will be detected more easily by our proposed defenses.

Notice also that our adversary model covers local and non-local DP use-cases, as we are agnostic to the way in which the adversary is able to modify the records of a database. In the next section we will use a toy example with classical DP, and in our time-series use-cases at the end of the paper, we will use local DP.

Goal of the Attacker:

Recall that our attacker is not the classical curious attacker from DP because in this paper we focus on the integrity of the data protected by DP; instead, our adversary tries to leverage DP noise to attack the utility of the system while bypassing our security mechanism (attack-detection).

Our attacker has two main objectives:

- U To maximize the damage to the system as much as possible (by sending false data), and
- S To remain undetected by evading our anomaly detection system $\mathcal{D}$.

To model the first objective, we point out that in most cases, the attacker will want to deviate an estimate of a value. For example, for a query to a database containing the average height of a population, the attacker may want to increase (or decrease) the reported average height as much as possible. This type of attack has happened before, in the case of users submitting false data to Waze [8] (attackers wanted the application to show a heavy traffic jam in their neighborhood so that no cars would be routed to their streets). Therefore the objective function U should depend on how much deviation the adversary can induce to the real (or differentially private) value y.

To model the second objective of the attacker, we need to incorporate the true positive rate; that is, the ability of the anomaly detector to correctly detect the attack. The attacker wants this probability to be zero or close to zero, so an intuitive objective of the attacker is to minimize the probability of being detected.

Therefore, the attacker has two objectives that are in conflict with each other. To maximize the damage to the system, the attacker needs to send the user of the system false values as large (or small) as possible; however, these large values would be detected immediately as anomalies by our classifier. On the other hand, to minimize the probability of detection, the attacker needs to send false data as close to the real values as possible, but that perhaps would not create the desired effect by the attacker in the first objective. As such, the attacker needs to balance these two objectives in what is usually called **multi-criteria optimization**.

To maximize a utility U while at the same time minimizing a conflicting utility S, we only need to set one of these objectives as a constraint, and then maximize (or minimize) the other. The answer as to which objective we chose to maximize and which one we use as a constraint can be exchanged and we still get the same result thanks to the duality principle [23], [24]. Therefore, without loss of generality we assume that the attacker wants to maximize the damage to the system subject to the following constraint about the stealthiness of the attack:

Definition 2.1: We say that an attack is *stealthy* if the probability of raising an alarm when there is an ongoing attack is close to (or perhaps even lower than) the probability of raising an alarm during normal operations (i.e., a security analyst notices no operational difference between the statistical behavior of alarms under normal conditions and under an attack). More precisely, we say the attack is stealthy if $P^d - P^{fa} \leq \xi$, for $\xi > 0$ the desired level of stealthiness.

Therefore, the adversary has to solve the following optimization problem:

Problem 1: Adversary's Goal

$$\begin{aligned} &\max_{\mathcal{A}} \; U(\mathcal{A}, \mathcal{D}) \\ &s.t. \\ &P^d - P^{fa} \leq \xi. \end{aligned} \tag{1}$$

Goal of the Defender: We now show how to design better attack-detection algorithms against a strategic attacker hiding her attack in DP noise. We need to design $\mathcal{D}$ (and $\mathcal{M}$) so that 1) differential privacy is maintained (for any attacker without access to the raw data), 2) we have an upper bound on the false positive rate, and 3) the attacker's maximum achievable payoff is as low as possible subject to the first two constraints. In other words, the problem for the defender is the following:

Problem 2: Defender's Goal

$$\begin{aligned} &\min_{\mathcal{D}} \max_{\mathcal{A}} \; U(\mathcal{A}, \mathcal{D}) \\ &s.t. \\ &P^d - P^{fa} \leq \xi. \end{aligned} \tag{2}$$

It is important to note that (to model realistic adversarial conditions) this is a leader-follower game, where the defender

has to make the first move (minimize the maximum damage the attacker can do) and disclose the selected attack-detection algorithm $\mathcal{D}^*$ (which can be a randomized algorithm) and the attacker then, given the fixed $\mathcal{D}^*$, will find an optimal attack strategy (maximize the payoff function with a fixed $\mathcal{D}^*$).

The next sections show concrete examples of how to solve these problems. In section III, we first show how optimal attacks can be designed by solving Problem 1, and then we show how a new resilient classifier can be designed in section IV. In particular, in section IV, we show that our solution to Problem 2 satisfies a saddle-point equilibrium between the two players, and therefore, the optimal strategy for the attacker is the same one that the defender anticipated a priori.

III. Optimal Attack

We are now ready to prove our main results; (1) in this section we find an optimal attack, and (2) in Section IV we an optimal defense against this attack, and any other attack perpetrated by *weaker* adversaries.

Consider is a numerical database $\boldsymbol{X} = \{x_1, x_2, \ldots, x_n\}$ that is used to generate a value Y that is made public (e.g., a trusted aggregator takes the information of the database and publishes the sum $\boldsymbol{Y} = \sum_{j=1}^{n} x_j$). A differential privacy mechanism $\mathcal{M}$ adds zero-mean random noise $\boldsymbol{\eta} \in \Omega$ to the query response $\boldsymbol{Y}$ to guarantee specific levels of privacy (ϵ, or (ϵ, δ)) such that the new published information becomes $\bar{\boldsymbol{Y}} = \boldsymbol{Y} + \boldsymbol{\eta}$. Due to the noise, $\bar{\boldsymbol{Y}}$ follows a probability distribution f_0 with mean $\theta = \boldsymbol{Y}$.

A malicious adversary is able to intercept the published information $\bar{\boldsymbol{Y}}$ and replace it with $\boldsymbol{Y}^a$, which follows a probability distribution f_a (as illustrated in Figure 1).

First Objective of the Attacker
As explained in the previous section, the attacker has two goals: 1) find the probability distribution f_a that maximizes the damage to the system, and 2) remain undetected. For the first goal, the impact of an attack is defined in terms of the amount of bias an attacker can introduce into our computations—this definition is typically used as a metric of the impact of attacks in cyber-phyiscal systems [25], [15], [26]. For this reason, we assume the attacker wants to maximize (or minimize) the mean $E[\boldsymbol{Y}^a] = \int_{r\in\Omega} r f_a(r) dr$ as the difference between the mean of Y^a and the mean of Y is a measure of the damage to the system (i.e., how much can the attacker deviate our computation). Therefore, the payoff of the attacker is $U(\mathcal{A}, \mathcal{D}) = E[Y^a]$.

Second Objective of the Attacker
We assume the adversary knows the DP mechanism and the probability distribution f_0 of the data without attack. For example, the attacker who managed to get the control of the sensor, can observe the actual data generated by the sensor and learn the statistical distribution of the data. To remain undetected according to Definition 2.1, the attacker needs to find a probability distribution f_a that is close enough to f_0 so that, no matter what statistical test the anomaly detection performs to find attacks, the results under f_a will be similar to the results under f_0. Stein's lemma [27], [28] relates the Kullback-Leibler divergence between two probability distributions $D_{KL}(f_a \| f_0)$ to the Neyman-Pearson criterion for the ability of a classifier to be able to correctly identify data with low false alarms and a high true positive rate. As the Kullback-Leibler divergence between the two distributions decreases, the error rates (false alarms and missed detections) increases.

Multi-Criteria Optimization So far we have identified two optimization criteria for the attacker (1) maximize $E[Y^a]$ and (2) minimize $D_{KL}(f_a \| f_0)$. Optimizing two competing objectives falls within the theory of multi-criteria optimization. The most popular way of solving these two competing objectives is to select one objective as the utility to be maximized (or minimized) and the second objective as the constraints of the problem by selecting a fixed parameter γ for the constraint. The solution parameterized by γ can then be used to find the Pareto optimal curve between these competing objectives. Therefore, the attacker's strategy to maximize the chance of remaining undetected while introducing a bias in the data can be found by solving the following optimization problem:

Problem 1

$$\begin{aligned} &\max_{f_a} E[\boldsymbol{Y}^a] \\ &s.t. \\ &D_{KL}(f_a \| f_0) \leq \gamma \\ &f_a \in \mathcal{F} \end{aligned} \tag{3}$$

where $\mathcal{F}$ corresponds to the set of all probability distributions, and $\gamma > 0$ is a constant that indicates how tolerable the adversary is to being detected (or the cost she is willing to pay). For instance, a large γ implies that the adversary does not care to be detected, and small γ will make the adversary distribution hard to distinguish from f_0.

Remark: Functional Optimization
Notice that the optimization problem in equation (3) is not a typical problem with a numerical solution. Instead, in this case the solution is a function (a probability density function) f_a, and the search space corresponds to all the possible probability density functions that satisfy the constraints. Variational methods is one of the ways that can be used to attempt to solve functional optimization problems [29].

We are now ready to present one of the main results of the paper.

Theorem 3.1: For any probability distribution f_0, the optimal strategy f_a^* that solves the optimization problem in equation (3) is given by

$$f_a^*(y) = \frac{f_0(y) e^{\frac{y}{\kappa_1}}}{\int f_0(r) e^{\frac{r}{\kappa_1}} dr}, \tag{4}$$

where κ_1 is the solution to $D_{KL}(f_a^* \| f_0) = \gamma$.

Proof: Maximizing the expected value is the same as minimizing the following equation

$$-E[\boldsymbol{Y}^a] = -\int_{r\in\Omega} r f_a(r) dr. \tag{5}$$

This is the objective function we will use in the Lagrangian (we will see why we changed the objective to a negative function when we find the Lagrange multipliers).

We want to find the probability density that will maximize Equation 5 Subject to the following two constraints: The first constraint is that we want f_a to be stealthy, that is, to follow as close as possible the probability density function of the added DP noise (we model this as being close to f_0 in the Kullback-Leibler sense)

$$D_{KL}(f_a \| f_0) = \int_{r\in\Omega} f_a(r) \ln\left(\frac{f_a(r)}{f_0(r)}\right) dr \leq \gamma. \tag{6}$$

The second constraint for the function f_a is that it needs to be a probability density function:

$$\int_{r\in\Omega} f_a(r) dr = 1. \tag{7}$$

In variational methods, the trick is to find a function (or matrix) by using a perturbation of the optimal function with a parameter α; this trick allows us to optimize a function with a fixed parameter, and that fixed parameter will help us find the shape of the function. Therefore we define the small variation of the optimal function as:

$$q(r,\alpha) = f_a^*(r) + \alpha p(r). \tag{8}$$

where the function we are looking for $f_a^*(r)$ has a small perturbation with parameter α and an unknown function $p(r)$.

By replacing f_a with $q(r,\alpha)$ in the original optimization problem, we can now optimize with respect to α. Because the problem is a constrained optimization problem, we need to find the Lagrangian, which is the objective function in equation (5) and the constraints in equations (6) and (7) multiplied by a Lagrange multiplier.

The Lagrangian of the objective function with $f_a(r)$ replaced with $q(r,\alpha)$ and the constraints is then

$$L(\alpha) = \int_{r\in\Omega} r q(r,\alpha) dr + \kappa_1 \left(\int_{r\in\Omega} q(r,\alpha) \ln \frac{q(r,\alpha)}{f_0(r)} dr - \gamma \right) + \\ + \kappa_2 \left(\int_{r\in\Omega} q(r,\alpha) dr - 1 \right),$$

where κ_j is the j^{th} Lagrange multiplier.

Next, we take the derivative of the Lagrangian with respect to α and set $\alpha = 0$ (the optimal point), for all possible p(r):

$$\frac{dL(\alpha)}{d\alpha}|_{\alpha=0} \text{ for all } p(x)$$

leading to

$$-r + \kappa_1 \ln \frac{f_a^*}{f_0} + \kappa_2 = 0.$$

We can now solve the above optimality condition for f_a^* in terms of the Lagrange multipliers:

$$f_a^* = f_0 e^{\frac{r}{\kappa_1} - \frac{\kappa_2}{\kappa_1}}.$$

We now need to find the values of the Lagrange multipliers, and we do so by solving replacing the above solution in the constraints of the problem. By forcing f_a^* to be a probability distribution, i.e., $\int_{r\in\Omega} f_a^*(r) dr = 1$, we can solve for one of the Lagrange multipliers (κ_2) and we obtain equation (4), which is the answer to our theorem.

In equation (4) we still have one Lagrange multiplier (κ_1) without a fixed value. Unfortunately we cannot get an analytical solution for that remaining Lagrange multiplier, but κ_1 can be found numerically for a given f_0, by solving in terms of κ_1 the following nonlinear equation

$$D_{KL}(f_a^* \| f_0) = \gamma. \tag{9}$$

■

As our first main result we found the optimal attack: **given a DP mechanism which creates a distribution on the data with a probability density function f_0, we can find the optimal attack distribution by using equation (4) and finding the Lagrange multiplier κ_1 by solving numerically equation (9)**.

In the following example we show how to find the optimal attack for the Laplace DP mechanism. We will then show the optimal attack for the Gaussian DP mechanism.

A. *Example: DP with Laplace Mechanism*

Figure 3 illustrates the setup for our problem. Suppose a trusted aggregator takes information from a database and publishes the sum $\sum_{j=1}^{n} x_j$.

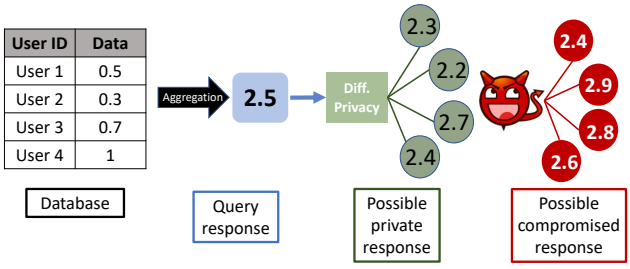

Figure 3. Example to illustrate our main results. The data generated by the DP mechanism follows distribution f_0 (e.g., Laplace mechanism), and the data generated by the attacker follows distribution f_a.

Let us assume that the privacy of the users of the system is protected with a DP mechanism that uses a Laplace distribution with mean zero and variance $2b^2$. Therefore the query response with DP is $\bar{\boldsymbol{Y}} = \sum_{j=1}^{n} x_j + L$, where L is a random variable with the Laplace distribution with zero mean and $2b^2$ variance.

Let $\theta = \sum_{j=1}^{n} x_j$. Then, $\bar{\boldsymbol{Y}}$ is a random variable with Laplace distribution:

$$f_0(y) = \frac{1}{2b} e^{-|y-\theta|/b} \tag{10}$$

with mean θ and variance $2b^2 = \sigma^2$.

Now assume the attacker wants to maximize the bias added to the query, but at the same time, in order to remain undetected, she wants to add noise with a distribution close to f_0 in equation (10). Using our theorem, the attacker now only needs to use equation (10) in equation (4), to get the following attack distribution:

$$f_a^*(y) = \frac{\kappa_1^2 - b^2}{2b\kappa_1^2} e^{-\frac{|y-\theta|}{b} + \frac{(y-\theta)}{\kappa_1}} \tag{11}$$

where κ_1 can be found by solving the following equation $\frac{2b^2}{\kappa_1^2 - b^2} + \ln(1 - \frac{b^2}{\kappa_1^2}) = \gamma$ for $\kappa_1 > b$.

Now that we found the optimal attack distribution against a Laplace DP mechanism, we can also find the average amount of bias introduced in the sampled data (the damage to the utility of the data). We call this the "impact" of the optimal attack, $\mu_a^* = E[\boldsymbol{Y^a}]$, which is given by

$$\mu_a^* = \frac{b^2(\theta - 2\kappa_1) - \theta\kappa_1^2}{b^2 - \kappa_1^2}. \tag{12}$$

Remark 3.1: Notice that, if $\gamma = 0$, then $\kappa_1 \to \infty$, such that $\lim_{\kappa_1 \to \infty} \mu^{a*} = \theta$ and $f_a^* = f_0$.

Figure 4 illustrates the optimal attack distribution f_a^* for different γ when the ϵ-differential privacy mechanism follows a Laplace distribution with $b = 2$ and when the original query response is $\boldsymbol{Y} = 2.5$ (as in Figure 3).

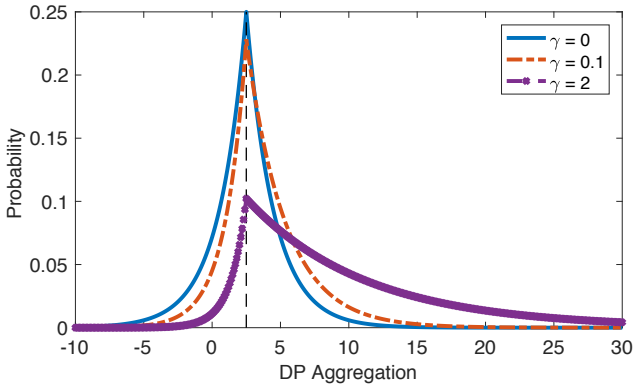

Figure 4. Example of f_a^* with a Laplace differential privacy mechanism for $\epsilon = 0.5$ and $b = \frac{1}{\epsilon}$. Notice that, for $\gamma = 0$, $f_a^* = f_0$. When $\gamma > 0$, f_a^* is not a Laplace distribution, which illustrates the value of our results.

Trade-off between Impact of the Attack and Privacy:

For a fixed γ, we can show the relationship between the privacy loss ϵ that dictates the level of privacy, and the bias introduced by the attacker μ_a^*.

For the example described in Figure 3, let us assume that $b = \frac{1}{\epsilon}$ i.e., the sensitivity is 1. The trade-off between the impact of the attack $\mu_a^* = E[\boldsymbol{Y}^a]$ and the privacy loss ϵ is illustrated in Figure 5. Clearly, more privacy (i.e., small ϵ) leads to more noise, which results in a larger attack impact (i.e., it becomes easier for an adversary to hide in the differential privacy noise).

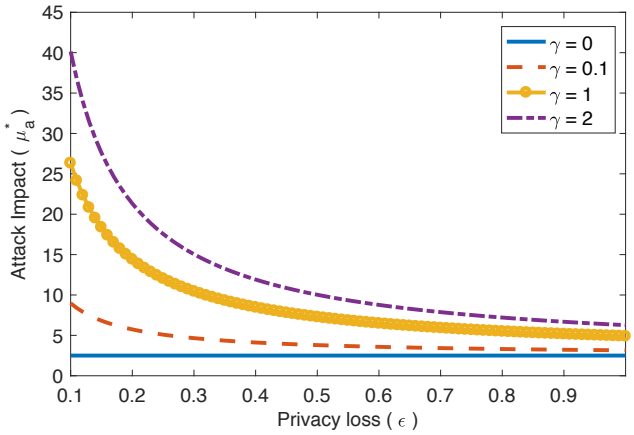

Figure 5. Trade-off between the impact of the attack μ_a^* and the privacy loss (ϵ) for different γ. Clearly, larger ϵ (lower privacy) leads to less noise, which results in lower attack impact.

B. *Use Case 2: Gaussian Mechanism*

In the above example we assumed the DP mechanism was Laplace. We now show how to obtain the optimal attack distribution if the DP mechanism is Gaussian. Let us assume a Gaussian distribution with mean θ_r and variance σ_r^2 of the form

$$f_0(r) = \frac{1}{\sqrt{2\pi}\sigma_r} e^{-\frac{(r-\theta)^2}{2\sigma_r^2}}$$

We can replace f_0 with a Gaussian distribution in equation (4) and then we need to solve for κ_1 in equation (9).

In order to obtain κ_1, we need $\int f_1^* \ln \frac{f_1^*}{f_0} \leq \gamma$. Solving the integral for all the domain of the normal distribution, we have that $\kappa_1^2 = \frac{\sigma_r^2}{2\gamma}$. This result shows that the solution to κ_1 does not need to be numerical all the time, in the case of the Gaussian distribution we find this Lagrange multiplier analytically.

Replacing κ_1 with the above result, we obtain the optimal attack distribution:

$$f_a^*(r) = \frac{1}{\sqrt{2\pi}\sigma_r} e^{-\frac{(r-\theta_r-\sqrt{2\gamma}\sigma_r)^2}{2\sigma_r^2}}. \tag{13}$$

Clearly, the optimal attack f_a^* that maximizes the impact of the attack against a Gaussian DP mechanism is also described by a Gaussian distribution with the same variance of the residuals without the attack but with shifted expected value $\mu^{a*} = \theta_r + \sqrt{2\gamma}\sigma_r$.

IV. Countermeasure: Designing an Optimal Defense

We have shown how to design optimal attacks that take advantage of the differential privacy mechanism to hide malicious data perturbations. Now, we assume that a defender takes the differentially-private query response and analyzes it in order to determine if there was an attack.

The defender's goal is to decide whether a random variable y belongs to hypothesis H_0 (normal behavior) or H_1 (anomalous behavior). This is a classification problem that can be generally solved using a variety of machine learning

methods. Machine learning is good whenever we do not know a priori the exact distribution of data under a normal situation; however, in the particular case of DP, we do know the exact distribution of the data f_0.

One of the most well-known results in hypothesis testing is that if we know both the exact distribution f_0 of the values under H_0 and the exact distribution of the data f_1 under the alternative hypothesis H_1, then the classifier that finds the optimal trade-off curve (the ROC curve) between the probability of a false alarm and the probability of a true detection is the log-likelihood ratio test:

$$\mathcal{D}(y) := \Lambda(y) = \ln \frac{f_1(y)}{f_0(y)},$$

where an alarm is triggered if $\Lambda(y) > \tau$.

The big challenge of using the test above for adversarial situations, is the fact that the attacker can tailor the attack so that y follows an arbitrary distribution f_a, but the defender in general does not know that distribution, i.e., $f_1 \neq f_a$, because the defender acts first. To describe this defender-attacker interaction, we can define a sequential game, where the defender plays first and selects the detection $\mathcal{D}$ which depends on f_0 and f_1 (where f_1 is what the defender assumes the attacker is going to use), and then the attacker chooses an attack distribution f_a which may or may not be equal to f_1.

We define the pair (f_a, f_1) as a strategy of the game, where the defender move is to assume f_1 and the attacker's move is to select f_a. We define the payoff $U(f_a, \mathcal{D})$ as the utility gained by each player when playing the strategy (f_a, f_1). They payoff in this case is the likelihood that $\mathcal{D}(y)$ will not raise an alarm.

Let $E_a[\Lambda]$ denote the expected value of the log-likelihood ratio test when the adversary chooses f_a, where

$$E_a[\Lambda] = \int_{y\in\Omega} f_a(y)\Lambda(y)dy. \tag{14}$$

Since the defender acts first, we assume the attacker knows f_1, and therefore can design an attack f_a in such a way that $E_a[\Lambda] \leq \widetilde{\gamma}$. In other words, a stealthy attacker wants to guarantee that the expectation of the log-likelihood ratio test remains below the threshold $\widetilde{\gamma}$, which dictates the cost that the adversary is willing to pay for not being detected. For instance, if the adversary wants to decrease the possibility of being detected, she can select $\widetilde{\gamma} < \tau$.

Notice that when the defender chooses f_a^* and the attacker chooses the same distribution, then $E_a[\Lambda]$ becomes the KL divergence we studied in the last section.

The goal of the adversary is to maximize the payoff $U(f_a, \mathcal{D}) = E[\boldsymbol{Y^a}]$ while minimizing the probability of being detected. The goal of the defender is to anticipate the worst-possible probability distribution the attacker can select to damage the system, and then try to minimize the negative impact of this attack. In other words, the game-theoretic problem we would like to solve is:

$$\min_{f_1\in\mathcal{F}} \max_{f_a\in\mathcal{F}_a} U(f_a, \mathcal{D}) \tag{15}$$

where $\mathcal{F}$ is the set of all valid pdfs, and $\mathcal{F}_a \subset \mathcal{F}$ is the set of pdfs such that $E_a[\Lambda] \leq \widetilde{\gamma}$.

A. Solution of the min-max Problem

Solving a min-max problem is not easy and it typically requires the use of numerical tools. However, in this paper, we show a method to prove analytically that the max-min solution is the same as the min-max solution.

In particular, to solve the min-max problem in (15), we follow these steps:

1: Formulate a max-min problem assuming that the attacker plays first, such that the detection strategy is always $f_1 = f_a$.

2: Find the optimal solution f_a^*.

3: Show that the solution of the max-min problem f_a^* and $f_1 = f_a^*$ corresponds to a saddle point (also known as a Nash equilibrium) and that it is also a solution of the original min-max problem in equation (15).

Max-min Problem: For the first step we assume that the attacker plays first, and then the defender selects the strategy $f_1 = f_a$. In this case, $E_a[\Lambda] = D_{KL}(f_a \| f_0)$, such that the max-min game is equivalent to Problem 1. Thus, the solution of the max-min game $f_a^* = f_1^*$ corresponds to the solution of Problem 1 in equation (4).

For the third step, we can prove that the solution in equation (4) is also the solution of the min-max problem in equation (15), by showing that it is a saddle point of the function U, according to the following definition.

Definition 4.1: A pair $(f_a^*, \mathcal{D}^*)$ is called a saddle point of the function U if

$$U(f_a, \mathcal{D}^*) \leq U(f_a^*, \mathcal{D}^*) \leq U(f_a^*, \mathcal{D}) \tag{16}$$

The connection with our min-max problem in equation (15) lies in the fact that the saddle point $(f_a^*, \mathcal{D}^*)$ also solves the min-max problem [30].

We are now ready to prove the second main result of the paper.

Theorem 4.1: Let $\mathcal{D}$ be the defender strategy with pdfs f_0 and f_1. Suppose the adversary launches an attack with distribution f_a. The solution of Theorem 3.1, when $f_1 = f_a$, corresponds to a saddle point $(\mathcal{D}^*, f_a^*)$ of U, such that the defender or the attacker do not have any incentive to chose anything other than $f_1^* = f_a^* = f^*$.

Proof: The right-hand inequality in equation (16) is the result of the optimality of the log-likelihood ratio if we know the probability distributions of both classes.

In order to show that the left-hand inequality in (16) holds, we can assume that the adversary selects any pdf $f_a = f_2 \in \mathcal{F}_a$ with mean μ_2, and the defender chooses the solution of Problem 1, $f_1 = f^*$ according to (4) with mean μ^*. Therefore, from (14) we have that the expected value of the log-likelihood ratio test with respect to f_2, $E_2[\Lambda^*(y)]$ is

$$\begin{aligned}\int f_2(r) \ln \frac{f_a^*(r)}{f_0(r)} dr &= \int f_2(r) \ln \frac{e^{r/\kappa_1}}{\int f_0 e^{s/\kappa_1} ds} dr \\ &= \int \frac{f_2(r) r}{\kappa_1} dr - \ln \int f_0 e^{s/\kappa_1} ds \\ &= \frac{\mu_2}{\kappa_1} - \ln \int f_0 e^{s/\kappa_1} ds. \end{aligned} \tag{17}$$

On the other hand, notice that if we had selected f_a^* instead of f_2, the log-likelihood ratio test becomes

$$\int f_a^*(r) \ln \frac{f_a^*(r)}{f_0(r)} dr = \frac{\mu^*}{\kappa_1} - \ln \int f_0 e^{s/\kappa_1} ds = \gamma.$$

Therefore, this equation shows us how the parameter for stealthiness γ is related to the mean value $\mu*$, which is what we had maximized in the previous section. If we decide to be stealthier, then the mean value must decrease and vice versa. The Lagrangian κ_1 will change signs depending on whether we are maximizing our bias or minimizing it.

Summary of main contributions: The solution to problem 1, which is a functional optimization problem with an analytic solution, shows the optimal way for an attacker to imitate the noise added by DP to remain as close as possible to the statistical properties of DP, while adding the maximum amount of deviation to the data. The solution to our defense shows that the optimal way to detect this attack is to use a log-likelihood ratio test with the DP noise as the distribution under normal conditions, and the solution of problem 1 as the distribution under attack conditions. The Saddle point equilibrium proof of the solution of problem 2 shows that if the attacker deviates from this strategy, its payoff will be smaller. **That is if we assume a weaker attacker (with less knowledge of the system), then our countermeasure will work even better.**

V. Extension to Time Series Systems

Our previous results were specified for static databases; however, our results can easily be extended to time-series analysis when using a sequential anomaly detection strategy and given a differential privacy mechanism. We are interested in particular in anomaly detection algorithms used in cyber-physical systems such as in transportation systems or the power grid.

A. Differential Privacy for Time-Series

Let $\boldsymbol{X} = \{\boldsymbol{x}(0)^\top, \ldots, \boldsymbol{x}(T-1)^\top\}$ for $\boldsymbol{X} \in \mathbb{R}^{nT}$ be a sequence (which we will call state trajectory as it denotes the "state" of a system or sensor) over a time window of size T that we want to keep private.

Definition 5.1: Two vectors $\boldsymbol{x}(k)$ and $\boldsymbol{x'}(k)$ are adjacent at a time instant k if there exists a j such that $x_j(k) \neq x'_j(k)$ and $x_i(k) = x_i(k)$ for all $i \neq j$. In other words, the pair $\boldsymbol{x}(k), \boldsymbol{x'}(k)$ differ only in one element.
Two trajectories $\boldsymbol{X}, \boldsymbol{X'}$ are adjacent if the pair $\boldsymbol{x}(k), \boldsymbol{x'}(k)$ is adjacent for all time instants $k = 0, 1, \ldots, T-1$.

We can extend the standard differential privacy definition for state trajectories during a window of time T as follows

Definition 5.2 ((ϵ, δ)-Differential Privacy [31]): A randomized mechanism $\mathcal{M} : \mathbb{R}^{nT} \times \Omega \to \mathfrak{M}$ preserves (ϵ, δ)-differential privacy for the state trajectories of a physical process if for all adjacent sequences $\boldsymbol{X}, \boldsymbol{X'}$ and for all subsets of possible observations $\mathcal{O} \subseteq \mathfrak{M}$

$$\Pr[\mathcal{M}(\boldsymbol{X}) \in \mathcal{O}] \leq e^\epsilon \Pr[\mathcal{M}(\boldsymbol{X'}) \in \mathcal{O}] + \delta.$$

If $\delta = 0$ we say the mechanism preserves ϵ-differential privacy.

Remark 5.1: If the privacy mechanism ensures (ϵ, δ)-DP up to time T, then it also guarantees (ϵ, δ)-DP for all $k \leq T$.

In particular, we are interested in mechanisms that add a random noise $\eta_i(k) \in \Omega$ to sensor readings, so that the differentially private sensor $\bar{y}_i(k) = y_i(k) + \eta_i(k)$. Let $\boldsymbol{Y} = \{\boldsymbol{y}(0), \boldsymbol{y}(1), \ldots, \boldsymbol{y}(T-1)\}$ be the sequence of sensor readings that depend on the state trajectories, i.e., for any set of trajectories $\boldsymbol{X}$, there is a sequence of sensor readings $\boldsymbol{Y}$. Similarly, we can define $\boldsymbol{Y'}$ as the readings obtained from an adjacent trajectory $\boldsymbol{X'}$. The global sensitivity for two adjacent traces $\boldsymbol{X}, \boldsymbol{X'}$ is then

$$\Delta_{y,q} = \max_{\boldsymbol{X}, \boldsymbol{X'}} \|\boldsymbol{Y} - \boldsymbol{Y'}\|_q \tag{18}$$

where q indicates the ℓ_q-norm.

The composition properties of the sequence of states is inherent in the calculation of the Sensitivity by assuming that our dataset is not given by the states $\boldsymbol{x}(k)$ at each time instant, but by a sequences of states of size T, $\boldsymbol{X}$. In other words, we have an extended dataset formed by $\{\boldsymbol{x}(0), \boldsymbol{x}(1), \ldots, \boldsymbol{x}(T-1)\}$. Therefore, any DP that guarantees privacy for this extended dataset, also guarantees DP for any $k < T$. In this case, we ensure that for the first T samples, (ϵ, δ) differential privacy is guaranteed. For instance, if we calculate the 1-sensitivity for a pair of adjacent states $\boldsymbol{x}(k), \boldsymbol{x}(k)'$ and compare it with the 1-sensitivity of two adjacent sequences $\boldsymbol{X}, \boldsymbol{X'}$, we can observe that the latter is T times larger, which leads to larger noise that already takes into consideration the composition theorem. However, selecting T is challenging because in most cases, it implies that very large T induces large noises that may lead to a significant deterioration of the system estimation and control.

Two common randomized mechanisms that guarantee (ϵ, δ)-differential privacy are described in the following lemmas:

Lemma 5.1: **Laplace Mechanism [32]:** For a dataset $\boldsymbol{X}$ and an output $\boldsymbol{Y}$, a Laplace mechanism preserves $(\epsilon, 0)$-differential privacy if $\eta(k)$ is drawn from a zero-mean Laplace distribution with parameter $b_\eta \geq \Delta_{y,1}/\epsilon$.

Lemma 5.2: **Gaussian Mechanism[32]:** For a dataset $\boldsymbol{X}$ and an output $\boldsymbol{Y}$, a Gaussian mechanism preserves (ϵ, δ)-differential privacy if at each $\eta_i(k)$ is drawn from a zero-mean Gaussian distribution with $\sigma_{\eta,i} \geq \sqrt{2\ln(1.25/\delta)}\Delta_{y,2}/\epsilon$.

We will first show what are the state-of-the-art algorithms used for Bad Data Detection in cyber-physical systems, then show how vulnerable they are to our attacks, and then show the design of a new anomaly detection algorithm based in our optimal defenses.

B. Vanilla Bad Data Detection (BDD) in Cyber-Physical Systems

Anomaly detection algorithms used in cyber-physical systems leverage temporal and geographical correlations between sensors to validate what is expected (or physically possible) with what is received from the sensors. These algorithms are usually referred to as Bad Data Detection (BDD) algorithms. The first step is usually to generate a prediction or *estimation* of the sensor readings $\hat{\boldsymbol{y}}(k)$ and comparing it with the reading

reported by the sensors $\boldsymbol{y}(k)$. If they are significantly different, then the BDD algorithm raises an alarm.

We can define the residuals $r_i(k)$ generated from comparing the differentially private sensor measurement $\bar{y}_i(k)$ with a prediction $\hat{y}_i(k)$ as follows

$$r_i(k) = \bar{y}_i(k) - \hat{y}_i(k). \tag{19}$$

Because of the DP mechanism, $r_i(k)$ is a random variable with mean $\theta_{r,i}$ and variance $\sigma^2_{r,i}$.

There is a strong type of attacks that have been introduced in [33], [34], whose aim to force each residual to follow a specific random distribution $f_{a,i}$. The attack is as follows:

$$\bar{y}^a_i(k) = \hat{y}_i(k) + \eta^a_i(k), \tag{20}$$

where $\eta^a_i(k)$ is the value the attacker controls, which is drawn from a $f_{a,i}$. Replacing (20) in (19), the residuals under attack are given by

$$r^a_i(k) = \bar{y}^a_i(k) - \hat{y}_i(k) = \eta^a_i(k), \tag{21}$$

Then, classical BDD algorithms can be represented as follows:

$$\mathcal{D}_i(\bar{y}_i) := \begin{cases} 1, & \text{if } |r_i(k)| > \bar{\tau}_i \\ 0, & otherwise \end{cases}, \tag{22}$$

for all $i = 1, \ldots, p$.

C. Optimal Defense: DP-BDD

We now adapt our results to the problem of sequential hypothesis testing. Our problem is to decide whether a sequence of observations $\boldsymbol{r} = \{r(1), r(2), \ldots\}$ belongs to hypothesis H_0 (normal behavior) or H_1 (anomalous behavior). From sequential decision theory, we know that the test that minimizes the time to detection subject to fixed bounds on the number of false alarms and missed detections is the sequential probability ratio test [35]:

$$\lambda(k+1) = \lambda(k) + \Lambda(k),$$

for $\lambda(0) = 0$, $k \in \mathbb{Z}_+$, and $\Lambda(k) = \ln \frac{f_1(r_k)}{f_0(r_k)}$ (where $f_0(\cdot)$ is the pdf of $\boldsymbol{r}$ under H_0 and $f_1(\cdot)$ the pdf of $\boldsymbol{r}$ under H_1). The detection algorithm then raises alerts after N iterations if the anomaly detection tool is either above or below a given threshold; otherwise it just keeps collecting data:

$$\mathcal{D} = \begin{cases} 1, & \text{if } \lambda(N) > \tau_U \\ 0, & \text{if } \lambda(N) < \tau_L \\ undecided, & \text{if } \tau_L \leq \lambda(k) \leq \tau_U, \text{for k=0,1,}\ldots \end{cases}$$

Notice that this algorithm requires the defender to know the distribution of the residuals subject to attack. However, similar to our formulation for the static case, the defender plays first and needs to anticipate the worst possible attack. Meanwhile, the attacker also adapts its attack according to the defender strategy. This interaction can be modeled as a minmax game, similar to the one introduced in (15), where f_1 is the distribution of the residuals under attack assumed by the defender and f_a is the distribution of the residuals caused by the adversary.

Wald's identity [36] decouples the probability of detection ($\widetilde{P}^d = \Pr_1[\mathcal{D} = 1]$) and false alarms ($\widetilde{P}^{fa} = \Pr_0[\mathcal{D} = 1]$) from the time it takes for the detector $\mathcal{D}$ to make a decision. In other words, $\widetilde{P}^d$ and $\widetilde{P}^a$ are fixed parameters, and the only thing the defender and attacker can modify for $\mathcal{D}$ is the time the test remains undecided (i.e., the time $\lambda(k)$ remains above τ_L and below τ_U) in the presence of an attack. Therefore, *the attacker will try to maximize the time $\mathcal{D}$ remains undecided to remain stealthy for as long as possible* while maximizing its impact with respect to the expected deviation of the residuals $E[r^a]$. Using Wald's identity, the expected value of this time is given by:

$$E_a[N] = \frac{\widetilde{P}^d_i \tau_U + (1 - \widetilde{P}^d_i)\tau_L}{\int f_a(r) \ln \frac{f_1(r)}{f_0(r)} dx}. \tag{23}$$

Therefore, the interaction between the attacker and the defender can be described by the following minimax optimization problem

$$\min_{f_1 \in \mathcal{F}} \max_{f_a \in \mathcal{F}_a} U(f_a, \mathcal{D}) \tag{24}$$

where $\mathcal{F}$ is the set of all valid pdfs, and $\mathcal{F}_a \subset \mathcal{F}$ is the set of pdfs such that $E_a[N] \geq T_d$, for $T_d > 0$ the minimum expected time that the detection test remains undetected.

Theorem 5.1: Let us consider the optimal sequential defense with fixed τ_L, τ_U chosen to guarantee desired probabilities of detection and false alarms, $\widetilde{P}^d, \widetilde{P}^{fa}$. Let f_0 be the distribution of the residuals without attack. The solution of the minimax problem in (24) when $U(f_a, \mathcal{D}) = E[r^a]$ is given by (4) where κ_1 is such that $D_{KL}(f_a || f_0) \leq \widetilde{\gamma}$ and

$$\widetilde{\gamma} = \frac{\widetilde{P}^d_i \tau_U + (1 - \widetilde{P}^d_i)\tau_L}{T_d}.$$

Proof: Notice that if the attacker wants $E_a[N]$ to satisfy a constraint (e.g. force the average time for $\mathcal{D}$ to make a decision greater than some specific value), then, according to (23)), it will be equivalent as forcing a constraint on the denominator (as the numerator is a fixed value). Recall that the denominator of (23) is the expectation of the log-likelihood test with respect to f_a,

$$E_a[\Lambda(r)] = \int f^*_a \ln \frac{f_1}{f_0(r)} dr. \tag{25}$$

Therefore, the constraint $E_a[N] > T_d$ can be rewritten as

$$E_a[\Lambda] \leq \frac{\widetilde{P}^d_i \tau_U + (1 - \widetilde{P}^d_i)\tau_L}{T_d}.$$

As a consequence the minimax problem in (24) is equivalent to the minimax problem in (15)**, such that the solution introduced in** (4) **is the one that maximizes the impact of the attack while guaranteeing that the time the sequential detector remains undecided is lower bounded by T_d. Clearly, since the solution is also a saddle point (as proven in Section 4.1), the attacker and defender do not have any incentive to select a different strategy**. ■

Remark 5.2: Notice that our previous results for the non-time-series formulation still hold because the optimization problems (for the attacker and for the defender) become the same thanks to equation (23). The saddle point proof is also the same one.

Remark 5.3: One of the main advantages of our saddle-point result (as in the case of a static database considered in

the previous section), is that, if the attacker is weaker (e.g., does not have all the knowledge necessary to launch optimal attacks), then our optimal attack detection will perform even better.

We now *illustrate the practicality of our optimal attacks and optimal defenses by implementing these solutions in a real-world traffic estimation system, and a smart metering case*.

VI. First Case Study: Traffic Density Estimation

We look at the problem of traffic density estimation for a variety of reasons, (1) location data is highly sensitive as it can reveal a variety of personal habits [37], [38], (2) loop detector data can be used to re-identify vehicles in roads [39], [40], [41], and (3) differential privacy algorithms have already been proposed for addressing privacy concerns in traffic density estimation problem [42].

In Section VI-A, we illustrate how traffic density is estimated by a transportation management center, in Section VI-B we describe a previously proposed differential privacy mechanism for traffic density estimation, and we then finalize the section studying our optimal attacks and optimal defenses.

A. Traffic Density Estimation Model

Transportation Management Centers (TMCs) use **traffic density**–the number of vehicles per mile per lane–for a variety of services, including intersection control (coordinating red lights) [43], ramp metering (stop lights before allowing you to enter a freeway) [44], and pricing in managed lanes (adjusting the toll rate upward and downward based on congestion) [45]. The TMC processes and fuses measurements from sensors in the road, with other operational and control data, and then it releases traffic information that can be used by the media and the public.

To estimate traffic density, TMCs use *loop detector* sensors. A loop detector consists of an induction electromagnetic device placed at a fixed location along a road segment that changes its current when a vehicle passes by. The information from inductive loops is collected in the traffic control cabinet, which transmits the data to the TMC. To leverage loop detector measurements, a road is usually divided in *cells*. A cell i is characterized by its length L_i, and its number of lanes l_i. This setup is illustrated in Figure 6.

For each cell i and lane j, a loop detector provides two main measurements (at periodic time intervals $\mathcal{T}$) [46]: 1) *flow estimate* $y_{\phi,i}(k)$ *of each cell*, based on the vehicle count $c_i^j(k)$, i.e., the number of vehicles that crossed the sensor during the last $\mathcal{T}$ seconds at each lane j, and 2) *occupancy* $o_i^j(k)$, which is the fraction of the time interval during which the cross-section is occupied by a vehicle. In particular, the flow estimate (vehicles per hour per lane) is computed as follows:

$$y_{\phi,i}(k) = \frac{1}{l_i \mathcal{T}} \sum_{j=1}^{l_i} c_j^i(k). \tag{26}$$

Besides from the loop detector measurements, sections of the road are characterized using historical data in order to fit a fundamental diagram, which is a state-of-the-art procedure that helps to relate the traffic flow with the traffic density, depending on the conditions of the road, i.e., if the road is

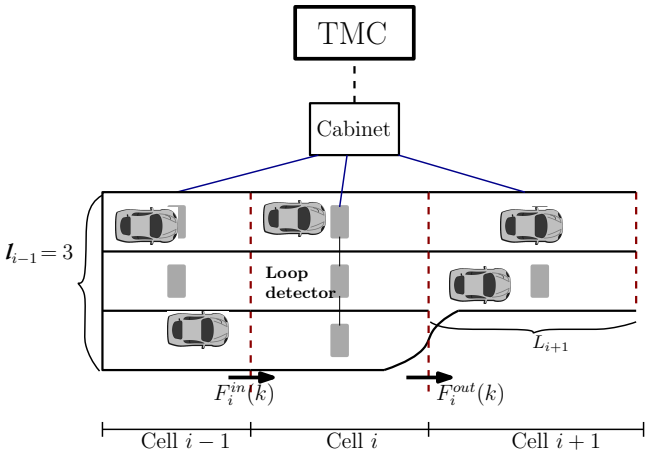

Figure 6. A road is divided into I cells, each one of length L_i and number of lanes l_i. The information from loop detectors is collected in the cabinet, and transmitted to the TMC.

congested or in freeflow mode (see [47] for the details about the fundamental diagram). The notation used in this section is summarized in Table I.

Given the loop detector measurements and the fundamental diagram parameters, traffic density estimation is computed by the steps shown in Figure 7. Notice that the loop detectors information is sent to the TMC, which first computes a raw density value y, and then sanitizes it through a density estimate, as follows.

Table I. Notation of the traffic density estimation

Symbol	Description
l_i	Number of lanes
$y_i, \hat{y}_i$	Traffic density / density estimation.
$y_{\phi,i}$	Traffic flow approximation
m_i	Mode estimation (F or C)
L_i	Length of the cell.
c_j^i	Vehicle count of lane j in cell i.
o_j^i	Vehicle occupancy of lane j in cell i.
$\mathcal{T}$	Loop detector sampling period
$w, v_f, q^{max}, \rho_C, \rho_J$	Fundamental diagram parameters

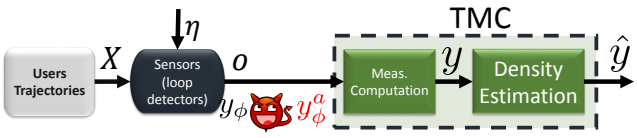

Figure 7. General description of the traffic estimation process with the differential privacy mechanism. o is the occupancy and $\bar{y}_\phi$ is the differentially private flow reading that is sent to the TMC. An adversary can falsify that information.

Mode estimate m**:** Highways can have two modes of operation: *freeflow* or *congested*. The characterization of these modes is given by the *fundamental diagram* of the lane, which is a triangular-shaped diagram that relates traffic density to traffic flow given the mode of operation—freeflow represents the part of the graph where the flow increases as the density increases, and congestion represents the part of the graph where the actual flow of cars decreases when the car density increases [47]. The loop detector can help us decide in which interval we currently are. Let $m_i(k) \in \{C, F\}$ be the current traffic mode in cell i where $C = congested$ and $F = freeflow$.

If the sensed occupancy value o_j^i produced by the sensor is larger than a predefined threshold, then the cell is assumed to be in *congested mode* (i.e., $m_i(k) = C$), and if the sensed value is lower than the threshold then the cell is assumed to be in *freeflow mode* (i.e., $m_i(k) = C$).

Density measurement computation y: Given the estimated mode m, we know in which part of the fundamental diagram we are, and we can then use the approximated flow y_ϕ to compute a traffic density measurement:

$$y_i(k) = \begin{cases} \frac{y_{\phi,i}(k)}{v_{f,i}}, & \text{if } m_i(k) = F \\ \rho_J - \frac{y_{\phi,i}(k)}{w_i}, & \text{if } m_i(k) = C \end{cases} \tag{27}$$

Predicted density estimate $\hat{y}$: Raw sensor data is generally noisy and sensitive to measurement errors. Similar to how power systems use **state estimation** to reconcile Kirchhoff's current and voltage physical laws with the sensed data received [48] (and to eliminate bad data), traffic algorithms also use **state estimation** to reconcile traffic flow equations (modeling the expected physical behavior of vehicles in a highway) with the values computed from raw measurements. The foundation of traffic flow models are the hydrodynamic flow-density equations describing the conservation of vehicles in the road, and they basically describe the notion that the traffic flow $y_i(k+1)$ in a cell i, is equal to the current traffic flow $y_i(k)$ plus the flow of cars into the cell F_i^{in} and minus the flow of cars out of the cell F_i^{out} (normalized by some parameters). The discrete-time version of these equations is called the Cell Transmission Model (CTM) [49]:

$$\begin{aligned} \hat{y}_i(k+1) =& \hat{y}_i(k) + \frac{\mathcal{T}}{l_i}\left(\frac{l_{i-1}}{l_i}F_i^{in}(k) - F_i^{out}(k)\right) \\ &+ Q_i(y_i(k) - \hat{y}_i(k)) \end{aligned} \tag{28}$$

where $Q_i \in \mathbb{R}$ is the estimator gain (e.g., from a Kalman filter).

Notice how the estimator takes advantage of the traffic dynamics to obtain a sanitized density estimation while minimizing the distance between the raw estimate $y_i(k)$ and the sanitized estimate $\hat{y}_i(k)$. For example, if $y_i(k)$ is greater than $\hat{y}_i(k)$, the term $Q_i(y_i(k) - \hat{y}_i(k))$ will cause our future estimate $\hat{y}_i(k+1)$ to increase. One particular property of this kind of estimators is that, for a proper selection of Q_i, the estimator acts as a filter. In our case, since $y_i(k)$ is noisy, the estimator provides smooth and accurate density values. Details about the hydro-dynamic model can be found in [47].

B. Differential Privacy for Traffic Density

We now focus on a differentially-private estimator for traffic density [42]. We begin by considering the discrete-time trajectories of users (vehicles), where each element of the trajectory at the k^{th} instant $x_l(k)$ for each user l, can be considered as a position in the road or GPS coordinates. For n vehicles, $\boldsymbol{x}(k) = [x_1(k), \ldots, x_n(k)]^\top$ is the vector of states and the dataset consists of set of trajectories $\boldsymbol{X} = \{x(0)^\top, x(1)^\top, \ldots\}$.

Now, let us consider the flow approximation from raw sensor data $y_{\phi,i}(k)$ for $i = 1, \ldots, M$ based on the count information provided by the loop detectors described in (26). The output vector is then $\boldsymbol{y_\phi}(k) = [y_{\phi,1}, \ldots, y_{\phi_M}]^\top$. The count of vehicles at each loop detector can be considered as the response of a count query, where the dataset is the location $\mathbf{X}$ and the query output is $\boldsymbol{Y_\phi} = \{y_\phi(1), y_{\phi,2}(2), \ldots\}$. According to Definition 5.1, two sets $\mathbf{X}, \mathbf{X}'$ are adjacent if they only differ by at most a single trace. Therefore, based on (18), the sensitivity is described by

$$\Delta_{\phi,p} = \max_{Adj\{\mathbf{X},\mathbf{X}'\}} \|\boldsymbol{Y_\phi} - \boldsymbol{Y_\phi'}\|_p$$

Previous work [42] have proposed the use of a Gaussian mechanism, and they obtained the 2-norm sensitivity given by:

$$\Delta_{\phi,2} = \|\boldsymbol{Y_\phi} - \boldsymbol{Y_\phi'}\|_2 \leq \frac{\sqrt{2}}{\mathcal{T}}\sqrt{\sum_{i=1}^{M}\frac{1}{l_i^2}}. \tag{29}$$

On the other hand, for a Laplace mechanism, it is easy to verify that the sensitivity is given by

$$\Delta_{\phi,1} = \|\boldsymbol{Y_\phi} - \boldsymbol{Y_\phi}'\|_1 \leq \frac{2}{\mathcal{T}}\sum_{i=1}^{M}\frac{1}{l_i}.$$

Notice that in this particular case, the sensitivity is independent of T because each vehicle only affects a sensor for a single time instant (because the road does not have loops), such that there is no need for composition and we can consider $T \to \infty$. Recall that the sensitivity is calculated as $\Delta_{\phi,2} = \|Y_\phi - Y_\phi'\|_2$. Then,

$$\|Y_\phi - Y_\phi'\|_2^2 = \sum_{k=0}^{T}\sum_{i=1}^{p}|y_{\phi,i}(k) - y_{\phi,i}(k)'|^2$$

where p is the number of sensors. Now, for each sensor i we have that

$$\sum_{k=0}^{T-1}|y_{\phi,i}(k) - y_{\phi,i}(k)'|^2 = \frac{1}{\mathcal{T}^2 l_i^2}\sum_{k=0}^{T-1}\left|\sum_{j=1}^{l_i}(c_j^i(k) - c_j^i(k)')\right|^2$$

An adjacent trace can affect sensor i in only two scenarios: at a different time than the original trace (i.e., the adjacent trace considers a vehicle that crosses the sensor at a different time) and/or by passing through a different lane. Notice that once a vehicle goes through sensor i, it would never affect future counts of the same sensor. For this reason, even if $T \to \infty$,

$$\sum_{k=0}^{T-1}|y_{\phi,i}(k) - y_{\phi,i}(k)'|^2 \leq \frac{2}{\mathcal{T}^2 l_i^2}.$$

The reasoning is the same for all p sensors, such that

$$\|Y_\phi - Y_\phi'\|_2 \leq \frac{\sqrt{2}}{\mathcal{T}}\sqrt{\sum_{i=1}^{p}\frac{1}{l_i^2}}$$

Therefore, the sensitivity is independent of T.

The objective of the differential privacy mechanism is to preserve privacy of the vehicle traces by adding a zero-mean random noise η_i to each flow measurement according to Lemmas 5.2 or 5.1. Thus, $\bar{y}_{\phi,i}(k) = y_{\phi,i}(k) + \eta_i(k)$ for each instant k and the new traffic density measurement becomes

$$\bar{y}_i(k) = \begin{cases} \frac{y_{\phi,i}(k)+\eta_i(k)}{v_{f,i}}, & \text{if } m_i(k) = F \\ \rho_J - \frac{y_{\phi,i}(k)+\eta_i(k)}{w_i}, & \text{if } m_i(k) = C \end{cases}.$$

The goal is to keep the count of cars during all the time in all loop detectors indistinguishable from any adjacent count.

Remark 6.1: Since $\bar{y}$ preserves (ϵ, δ)-differential privacy, the estimation $\hat{y}$ also guarantees the same level of privacy due to the post processing property of differential privacy mechanisms, according to [32].

C. Optimal Stealthy Attack

Recall that the residuals are defined as $r_i(k) = \bar{y}_i(k) - \hat{y}_i(k)$. Using historical data, it is possible to obtain an empirical cumulative function of the residuals [50].

Note that without attack, $r_i(k)$ includes two sources of noise or uncertainty, one caused by the variable vehicle count and the change of mode, and one from the differential privacy mechanism. Therefore, let $\sigma_{r,i}, \theta_{r,i}$ be the standard deviation of and mean $r_i(k)$ in normal conditions, respectively, that can be obtained by analyzing historical data.

For this problem, since the objective of the TMC is to obtain an accurate estimation of the traffic density, we define the payoff U as the deviation of the density estimation $\hat{y}$ from a specific operational or desired state $\boldsymbol{y}^{ref}$ during the entire duration of the attack, i.e., from an attack starting at $k = k_a$ and finishing at $k = k_f$.

$$U = \frac{1}{|\mathcal{V}^a|(k_f - k_a)} \sqrt{\sum_{k=k_a+1}^{k_f+1} \sum_{i \in \mathcal{V}^a} (E[(\hat{y}_i(k) - y_i^{ref})]^2} \tag{30}$$

An adversary with enough historical data about $\hat{y}^{ref}(k)$ can design an attack that solves Problems 1 or 2.

From our formulation in Section V, we can define the stealthy attack adapted from (20) as follows

$$y^a_{\phi,i}(k) = \begin{cases} v_{f,i}(\hat{y}_i(k) + \eta_i^a(k)), & \text{if } m_i(k) = F \\ \rho_J w_i - w_i(\hat{y}_i(k) + \eta_i^a(k)), & \text{if } m_i(k) = C \end{cases}, \tag{31}$$

where $\eta_i^a(k)$ is the attacker's random variable with $\mu_i^a = E[\eta_i^a]$. Replacing the attack in (31) into (27) leads to $y_i^a(k) = \hat{y}_i(k) + \eta_i^a(k)$.

Applying the results introduced in Sections III and V, the adversary can design $\eta_i^a(k)$ that maximizes her payoff while remaining stealthy for our optimal sequential defense strategy DP-BDD. We will compare our propose defense with the typical bad data detection (BDD) algorithm as well.

D. Experiments

We use the loop detectors data available from the Mobile Century experiment database [51]. This data consists of counts and occupancy measurements from single loop detectors for each lane of the freeway I-880 CA in California from post-mile[1] 16.5 to 27.7. 27 loop detectors (M=27) are located as illustrated in Figure 8, and their information is transmitted to the TMC with a sampling period of $T = 30$ seconds. The fundamental diagrams of each section are the same, and have the following parameters: $v_f = 65\ mph$, $w = 11.6\ mph$, $\rho_J = 193\ vehicles/mile/lane$, $\rho_C = 30\ vehicles/mile/lane$. In particular, between Tennyson Rd. and CA92 (i.e., postmile 26 to 27), the northbound (NB) direction presents a *recurrent and severe bottleneck* during the afternoon. Besides, on the day of the experiment, there was an accident during the morning, which caused a non-recurrent bottleneck at the same location.

[1]Postmile refers to a location marker in miles, used by the California Department of Transportation

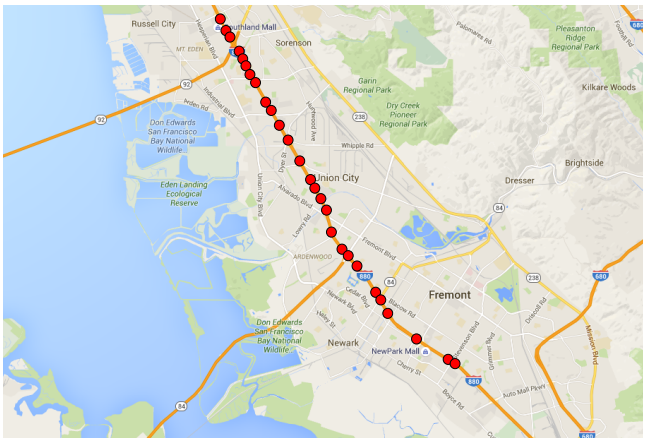

Figure 8. Stretch of the I-880 freeway used for the Mobile Century Experiment. Red circles indicate the location of loop detectors. The road is divided to 27 cells of non-uniform length (i.e., $M = 27$).

We design and launch stealthy attacks with and without the differential privacy mechanism. The objective of the adversary is to disrupt the density estimate when the traffic is in freeflow condition by alerting operators of a fake congestion when in reality there is none.

1) Attacks Against a System Without DP: We initially focus our attention on the segment between postmiles 17.5 to 23.5 (loop detectors 1 to 15), which are not affected by the accident (we later consider the accident).

We select the threshold τ_i such that the residuals of the anomaly detector maintains a probability of false alarms lower than $P_i^{fa} = 2\%$ for all i.

We assume the attacker compromises all measurements in the segment (15 loop detectors out of 27). Because the attacker knows the threshold τ_i and the model estimation parameters, she is able to design an attack according to Section VI-C.

First, let us consider the case *without differential privacy*. Using historical data, we estimate the standard deviation $\sigma_{r,i}$ of the residuals without attack in freeflow conditions. We then design an optimal stealthy attack by adding Normal noise with $\sigma_{r,i}$. Figure 9 illustrates the effects of the attack in the density map. The attack causes an increment on the estimated density, and in particular, the attacker can add 2.37 $veh/mile/lane$ per cell, and per unit of time.

2) Attacks Against a System Using DP: Now let us study a particular case where a differential privacy mechanism is included. We consider a Gaussian mechanism with parameters $\epsilon = 0.4, \delta = 3.5 \times 10^{-5}$. Notice that with these parameters (e.g., large δ value), *the level of privacy provided by DP is minimal, and yet, as we show, this gives an attacker a significant ability to manipulate the system.* The sensitivity obtained according to (29) is $\Delta_{\phi,2} = 196$, so the mechanism adds a zero-mean Gaussian noise with $\sigma_{\eta,i} = 1406$. Adding differential privacy affects the density maps by making them more noisy, as illustrated in Figure 10 (Top). However, for the selected DP mechanism, the map still provides a good estimation of the state of the freeway. When an optimal stealthy

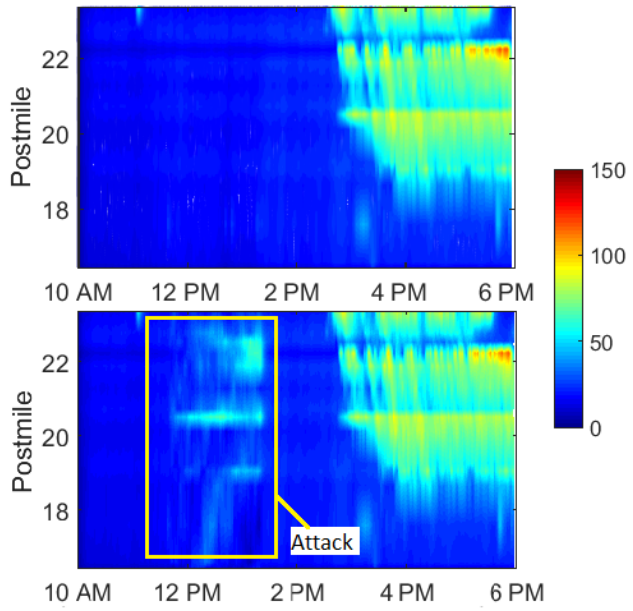

Figure 9. Density map without differential privacy mechanism between postmiles 17.5 to 23.5 from 10 AM to 6 PM. At 2:30 PM the effect of the recurrent bottleneck can be observed (Top). A stealthy attack is deployed from 11:50 AM to 1:50 PM for $P^{fa} = 2\%$ and $\xi = 1\%$ (bottom).

attack is launched, however, it causes a large increase in the estimated density during the time of the attack (Figure 10 bottom) with an impact of $S = 38.3\ veh/mile/lane$. This in return causes TMC to observe a *fake traffic jam*. However, for

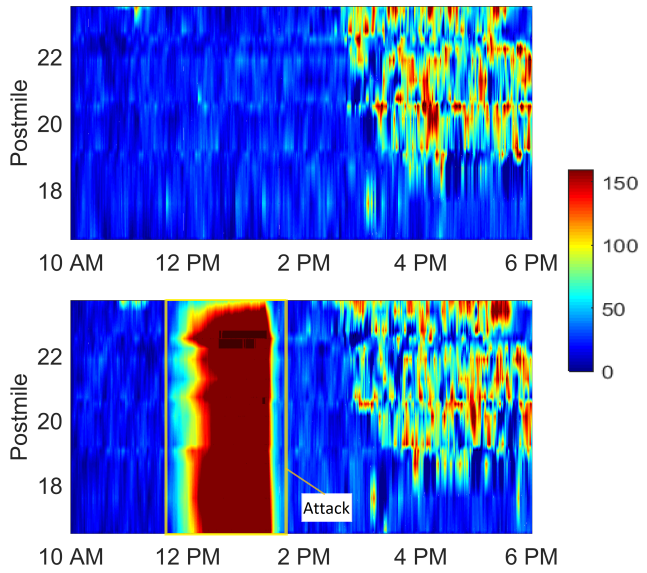

Figure 10. Density map between postmiles 17.5 to 23.5 with $(0.4, 3.5 \times 10^{-5})$-Differential private mechanism. The map under normal conditions (Top) and when a stealthy attack is deployed in 15 sensors from 11:50 AM to 1:50 PM (bottom) when BDD is present. The attacks is designed for $P^{fa} = 2\%$ and $\xi = 1\%$

the same level of privacy, it is possible to decrease the impact of the attack by using the DP-BDD detection mechanism, as depicted in Figure 11. Even though the adversary is still able to deviate the normal estimation, the deviation is not as large as with the BDD.

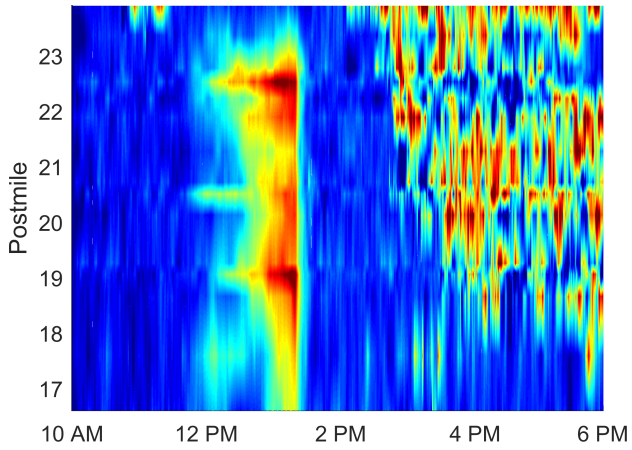

Figure 11. Density map between postmiles 17.5 to 23.5 with $(0.4, 3.5 \times 10^{-5})$-Differential private mechanism. The stealthy attack is deployed in 15 sensors from 11:50 AM to 1:50 PM when DP-BDD is implemented. The attacks is designed for $P^{fa} = 2\%$ and $\xi = 1\%$

Figure 12 illustrates the traffic density estimation in sensor 11 for both cases, with and without DP. Note that when the attack is launched at 11:50 AM, the traffic density estimation increases, as desired by the adversary. Without DP the rate of increase is relatively low, and it does not reach congestion. The optimal stealthy attack against a system that uses DP and BDD has a considerable effect, causing the illusion of a traffic jam in less than an hour. On the other hand, when the detection strategy uses DP-BDD, the rate of increase is lower than with BDD with the same level of DP, causing the illusion of slight traffic congestion, but not a traffic jam. Figure 13 show the

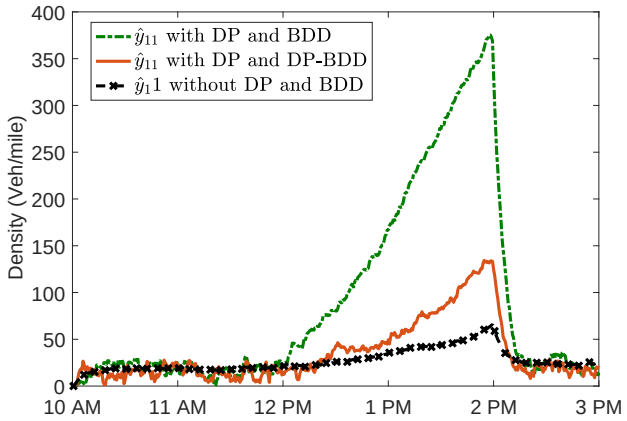

Figure 12. Density estimation in sensor 11 for the cases with and without the $(0.4, 3.5x10^{-5})$-Differential private mechanism. The optimal attacks are designed in both cases with $P^{fa} = 2\%$ and $\xi = 1\%$. Clearly, including DP allows an adversary to launch worse stealthy attacks; however using DP-BDD limits considerably the impact of the attack.

anomaly detection statistics (and thresholds) with and without attack for both detection strategies, BDD and DP-BDD. Notice that the anomaly detection statistics without attack and with attack look very similar in both cases (even though they cause significant deviations to our traffic flow estimates). This is the

behavior we were expecting given the constraints we added to our optimization problems.

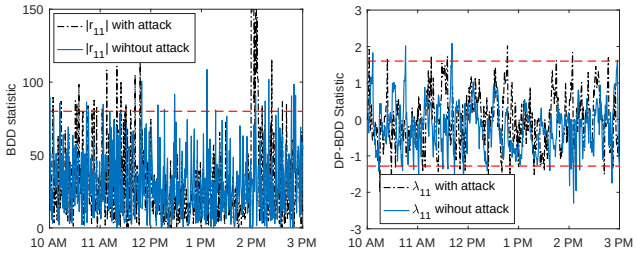

Figure 13. Distance measure of the BDD (left) and DP-BDD (right) in sensor 11 for the case without attack and with the optimal attack for $P^{fa} = 2\%$ and $\xi = 1\%$, and a $(0.4, 3.5x10^{-5})$-Differential private mechanism. The red dashed lines indicate the decision thresholds.

Now we want to analyze the trade-off between security and privacy for the two detection mechanisms proposed above. We assume that the objective of the adversary is to maintain the probability of being detected bounded by $\xi = 1\%$, while attacking a *subset* of 15 loop detectors for 2 hours. At the same time, an (ϵ, δ)-differential privacy mechanism is implemented, which preserves a desired level of privacy.

First, for a specific level of privacy, the detection mechanism selects proper thresholds that satisfy the desired probability of false alarms. Based on the selected threshold, we are able to calculate optimal false-data injection attacks that maximize the impact on the system while preserving the statistical properties that make the attack stealthy. Using the Mobile

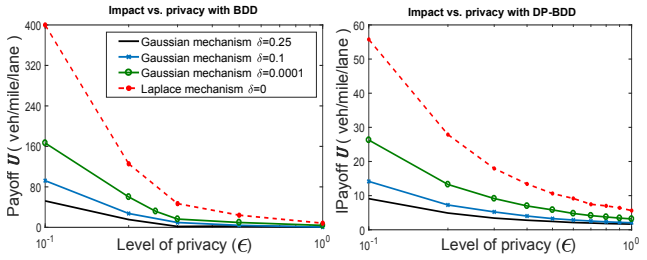

Figure 14. Maximum impact achieved by the optimal attack for several differential privacy mechanisms with BDD (left) and DP-BDD (right). Increasing ϵ or δ (i.e., decreasing the level of privacy) reduces the impact an adversary can cause because the added noise is reduced.

Century experiment data, we launch several optimal attacks for different levels of privacy for both detection strategies, and we compute the maximum impact it causes in the traffic density estimation, as illustrated in Figure 14. As expected, higher levels of privacy allow the attacker to launch more and more damaging attacks; however, according to Figure 14, the maximum impact the attacker can cause when we use the DP-BDD solution is significantly lower when compared with the impact the attacker can launch against BDD algorithms.

Our trade-off results can be used by the TMC to select an appropriate differential privacy mechanism that offers good privacy properties and low attack impact depending on the anomaly detection strategy implemented. From Figure 14, for each δ, there exists a Pareto optimal solution that maximizes both, privacy and security. For instance, assume we fix $\delta = 0$, (i.e., the Laplace mechanism); in this case $\epsilon = 0.4$, has 18% increase in privacy when compared to $\epsilon = 0.1$ (i.e., $e^{0.1} = 1.102$ and $e^{0.2} = 1.284$); however, it reduces the maximum impact by approximately 75%. Therefore, by reducing the given privacy protections, we can significantly reduce the negative impact of the adversary.

On the other hand, the selection of δ is strongly related to the number of sensors (size of the query output). For the Laplace mechanism, increasing the size of the query results in adding large amounts of noise, such that the quality of each answer deteriorates with the sum of the sensitivities of the queries (i.e., $\Delta_{\phi,2}$ increases considerably with the number of sensors) [31]. This can be mitigated to some extent using Gaussian noise instead of Laplacian. As a result, less noise is needed and the room for injecting false information is reduced. In our case, since our query output is of size $M = 27$, Gaussian mechanisms with $\delta \neq 0$ are perhaps a more adequate choice. As illustrated, in Figure 14, increasing δ minimizes drastically the damage caused by the adversary. Therefore, our analysis also facilitates the selection of an adequate δ that preserves a certain level of privacy while minimizing the attack impact.

VII. Second Case Study: Real-time Pricing for the Smart Grid

To show the generality of our results, we now briefly outline how to apply them to a different problem.

The model considers an electricity market with n consumers of electricity, a set of suppliers of electricity, and a third party entity—a demand response operator—with the goal of matching supply and demand by setting the retail market price for electricity.

The general assumption is that the ISO determines, at each time instant $k \in \mathbb{N}_+$, a clearing price $u(k)$ valid for the period of time $[k \cdot T, (k+1) \cdot T]$ (this is called an *ex-ante* market) every T hours (e.g., $T = 0.5h$) and announces it to the suppliers and consumers.

The electricity demand of each user is characterized by two components: a baseline electricity consumption $b_{c,i}(k)$ that captures the electricity consumption that is independent of the pricing mechanism (i.e., the necessary power to satisfy the main consumer needs at each instant k such as refrigerator, cooking devices, light bulbs), and a price-responsive demand $w_i(u(k))$, which captures the amount of electricity consumption that can be controlled by the pricing signal $u(k)$. For instance, doing laundry when the price is low, or turning off the lights of rooms that are not being used.

The demand of consumer i at each time instant k is $x_i^c(k) = w_i(u(k)) + b_i(k)$, for $x_i^c(k) \in [0, x_{max}^c \ kW]$. The total power consumed can be defined as the aggregated demand $x_T^c(k) = \sum_{i=1}^n x_i^c(k) = w(u(k)) + b_T(k)$.

Similarly, for the supply of electricity, Tan et al. [52], propose a linear regression between supply and cost, a model they validated from the Australian Energy Market Operator and the electricity market in California. Under these assumptions the total supply of electricity can be modeled by the following equation:

$$x_T^s(\lambda(k)) = au(k) + f, \tag{32}$$

where a and f are parameters estimated by the historical market data from the area of study.

The **control objective** of the ISO is to send price signals $u(k)$ to the users to keep the error between supply and demand of electric power $\mathcal{E}(k) = x_T^c(k) - x_T^s(k)$ close to zero for every time instant k in order to guarantee a stable operation of the grid (if the mismatch is too large, it will cause frequency deviations which can damage equipment and trip relays).

This can be seen as a control problem in which the system to be controlled is the outcome of a market, the control variable is the price signal $u(k)$ and measured variable is the error between supply and demand $y_{\mathcal{E}}^m(k) = x_T^c(k) - x_T^s(k)$.

The price signal $u(k)$ must be carefully designed because a direct feedback of the wholesale prices to the users might cause oscillations or even instability [52], [53].

Tan et al. [52] propose the following price-setting algorithm:

$$u(k) = u(k-1) - K y_{\mathcal{E}}^m(k-1), \tag{33}$$

where K is the control parameter selected in such a way that the overall feedback system remains stable (i.e., the supply-demand mismatch remains bounded).

A. Differential Privacy for RTP

Smart meters allow the utility provider to monitor consumption in order to update prices or adjust electricity generation. However, due to the accuracy and granularity of the data, it could be possible for a curious party to infer behavior profiles for each consumer. For example, in the field of non-intrusive load monitoring, it is possible to extract information about the type of appliance that is being used [54], [55]. We consider the differential privacy mechanism proposed by Ács and Castellucia [56], where each smart meter adds noise from Gamma distributions, such that the aggregation results in a Laplace noise $\eta(k)$, such that $\bar{y}_{\mathcal{E}}^m = y_{\mathcal{E}}^m + \eta(k)$.

Two set of adjacent traces differ only in the consumption of one user. Since the power consumption of each individual is bounded by x_{max}^c and since there is only one query, the sensitivity is then $\Delta_1 = |T x_{max}^c|$. Therefore, the Laplace noise has mean zero and parameter $b_\eta = \Delta_1/\epsilon = T x_{max}^c/\epsilon$.

B. Attack Detection

We assume that an adversary can modify the data from a subset of smart meters or tamper directly with the information disclosed to the control center. An anomaly detection mechanism can take the measurement with DP mechanism $\bar{y}_{\mathcal{E}}^m(k) = y_{\mathcal{E}}^m + \eta(k)$ and compare it with a prediction $\hat{y}_{\mathcal{E}}(k)$. For our experiments, we use the same CEO model from [52], but to predict the expected behavior (so our expected measurement is different from the real value), we assume a linearized model of the form $w(k) = \beta u(k) + w_0$, where β is a constant selected such that the model fits the price-responsive demand. The predicted error is obtained by the use of a classic Kalman filter [57], with the linear models of the total supply in (32) and the consumption as follows:

$$\hat{y}_{\mathcal{E}}(k+1) = (a-\beta)u(k) + Q(\bar{y}_{\mathcal{E}}^m(k) - \hat{y}_{\mathcal{E}}(k)) + f - w_0 - b_T.$$

Q is the Kalman filter gain that minimizes the estimation error. The residuals are $r(k) = \bar{y}_{\mathcal{E}}^m(k) - \hat{y}_{\mathcal{E}}(k)$.

Therefore in this case-study we need to find the worst type of attacks based on the results in Sections III and IV.

C. Experiments

We use a distribution feeder specification [58], which covers a moderately populated urban area composed by 1405 households. For our experiments with real-time pricing and its control, we use the same parameters from Tan et al. [52].

The detection threshold τ is selected such that $P^{fa} = 0.05$. The attacker launches an attack after 8 $days$ of simulation, and we evaluate it for different levels of privacy.

Figure 15 shows the performance of the system (ideally the mismatch between supply and demand should be zero) when it is subject to worst-possible attacks (fake price signal) against a system that does not use differential privacy but uses BDD for anomaly detection (in black) a system that uses differential privacy (with differential privacy for $\epsilon = 0.1$) and a BDD anomaly detection (in red) and a system that uses differential privacy and the proposed DP-BDD for attack-detection.

Note that when there is no DP noise, the maximum attack that an adversary can launch has little effect on the system; however, due to the addition of noise, an adversary has enough room to launch a stealthy attack that causes a significant deviation. The impact of the attack with the DP-BDD is overall smaller.

Figure 16 illustrates the impact in expectation for the proposed stealthy attack. For each ϵ, we launch the optimal stealthy attack against BDD and the DP-BDD. Note that there is a trade-off between privacy and impact. Increasing privacy (low ϵ) allows the attacker to inject larger stealthy attacks, and as a result, produce larger deviations from the desired state.

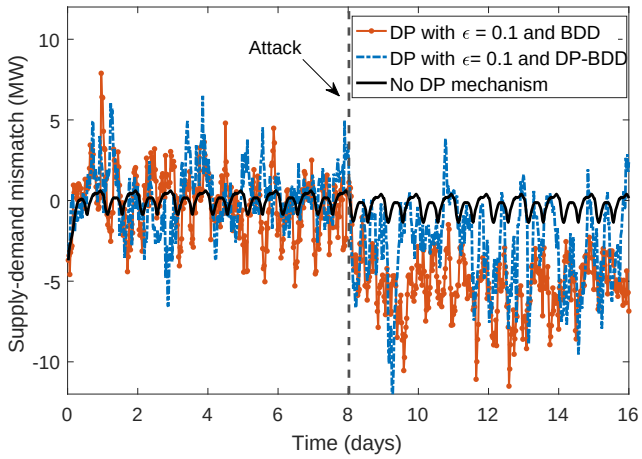

Figure 15. Supply-demand mismatch with and without Differential Privacy for a stealthy attack launched starting day 8.

VIII. Related Work

The main contribution of our paper when compared to previous work on (1) adversarial classification (which studies security vs. utility) and on (2) differential privacy (which

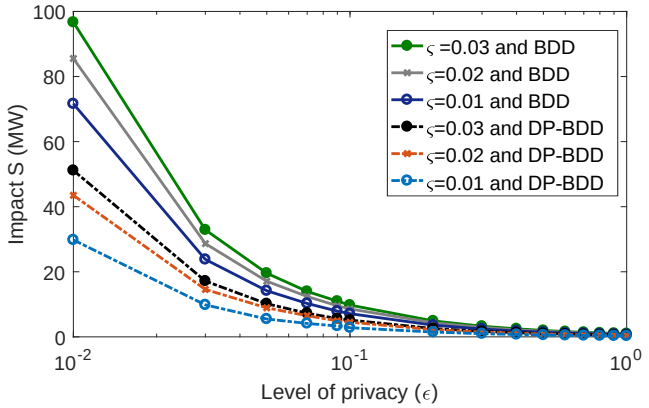

Figure 16. Maximum impact achieved by the optimal attack for different levels of privacy. Clearly, increasing ϵ implies less added noise, which leads to less impact.

studies privacy vs. utility), is to start the discussion of the trade-offs between security and privacy, as illustrated in Figure 17.

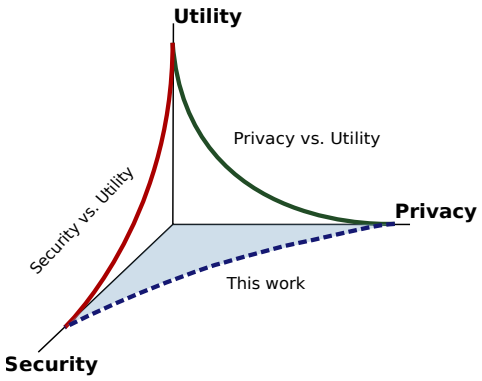

Figure 17. The literature on adversarial classification focuses on the Security vs. Utility trade-off and the literature on DP focuses on the Utility and Privacy trade-off. Some systems will need to provide both; attack detection algorithms, and DP, and we need to identify the trade-offs between security, privacy and utility, this paper is the first step in this direction.

The closest work to our own is the use of DP as a *defense* for adversarial machine learning [59]—instead of using DP as a way to attack a system (like we do in this paper), they use DP as a way to defend it. Their proposal falls into the topic of robustness evaluation of machine learning classifiers to adversarial attacks [60], which focus on finding the minimum accuracy a machine learning classifier can have when they have been exposed to adversarial training examples. The goal of robustness evaluation is to achieve prediction robustness, which means that adversarial examples cannot change the expected prediction of the classifier.

Most of the literature on robustness evaluation focuses on heuristics, and provable robustness approaches have seen limited progress. Lecuyer et al. [59] proposed a provable robust method by using DP at prediction time. The key idea is that adversarial examples consist of small inputs that can change the output significantly, and the solution is to make prediction models DP. In particular, their framework is called *randomized smoothing* (by [61]), which inherits theoretical results from the differential privacy community, allowing them to evaluate the level of accuracy under attack of their method. Follow up work to Lecuyer et al. include [62] and [61].

In contrast to using DP as a defense mechanism, in this paper we look at the converse problem; i.e., how DP can be exploited by attackers to improve their attacks, and how to design a classifier that minimizes the impact of these new DP-enhanced attacks.

In addition to work on adversarial machine learning, our work is related to the literature studying the security of Cyber-Physical Systems (CPS) against integrity attacks (also known as false data injection attacks). These attacks happen when sensors are not trusted or they can be compromised by an attacker, so traditional message authentication codes cannot prevent false data injected by an attacker. For example, attacks to traffic estimation in transportation networks have been studied before [7], [5], [6], where false information is used to modify the state estimation of traffic conditions. Similarly, attacks in the power industry have also received significant attention [14], [52]. The transportation community has also developed algorithms to detect failures of loop detector sensors in highways [63].

Work in Bad Data Detection (BDD) [14], [63], [15], [16], [17] has focused on identifying these false data injection attacks. The typical trade-off considered by the literature on BDD is one between the safe operation of the system (utility) and one of security. For example a fully secure system can be easily designed by raising an alert at every instance (this way we are guaranteed to have a 100% true detection rate). However this system will also generate several false alarms, and therefore (a system that detects even the most subtle attacks) can have a negative impact on the performance and safety of the system (when the system is not under attack). As far as we are aware, none of these papers has considered how to detect sensor failures (or false-data injection attacks) in systems that use differential privacy.

Another related line of work to our own results is the focus on privacy for CPS. For example, privacy has been considered before in transportation network applications [64], [65], and location privacy [38]. Differential Privacy in particular has been applied to a variety of location and traffic estimation problems [66], [67], [42], [68]. Privacy in the smart grid has also received significant attention [55]. None of these papers has considered how to develop BDD algorithms in systems with DP, or how to develop attack-resilient BDD algorithms against adaptive adversaries that exploit DP to hide their attacks.

Our motivation for formulating the problem of adversarial classification as a least-favorable probability distribution in a functional space comes from our previous work in the study of MAC-layer misbehavior in Wi-Fi networks [13], where we identified the optimal attack distribution to obtain unfair access to the transmission medium, while minimizing the deviation from the probability distribution of the Wi-Fi standard.

As far as we are aware we are the first to (1) consider the problem of *bad data detection* in a system protected by differential privacy, (2) formulate *optimal bad data injection attacks* tailored to maximize the dissemination of false information while remaining undetected, and (3) formulate *optimal attack-detection defenses* to minimize the negative impact of this attacker.

IX. Conclusions

We study of how attackers can affect the utility and integrity of a system by leveraging DP noise. While research in DP considers a curious adversary, our new attacker is not interested in the privacy of the data, instead, it takes advantage of the fact that DP adds noise and then strategically uses this noise to hide false data that affects the utility of the system.

We showed how attacks for systems providing DP have a negative impact that is orders of magnitude higher than attacks against systems without DP. We then proposed a better detection strategy based on solving a game-theory problem in the functional space of all possible distributions. Our experiments show the value of our defenses by comparing our DP-BDD solution to state of the art BDD tools currently used in practice for traffic estimation.

Our main contribution, summarized by Eq. 4, shows how the optimal attack (characterized by the probability density function f_a^*) is related to the probability distribution that DP uses to inject noise to the data (Laplace or Gaussian noise). We also show how our optimal DP-BDD algorithm is a Nash equilibrium between an attacker trying to find an optimal attack distribution, and the defender trying to design an optimal bad data detection algorithm.

As far as we are aware, our results are the first to consider adversarial evasion in the context of differential privacy, and we believe they open a new research area considering the trade-offs between utility, security, and privacy.

Acknowledgments

This work is supported by the CPS program of NSF through collaborative NSF awards CNS #1837517 and CNS #1929410. This work is also partially supported by ARO grant W911NF-17-1-0356 and NSF CNS-1931573.

References

[1] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography conference*. Springer, 2006, pp. 265–284.

[2] B. Ghena, W. Beyer, A. Hillaker, J. Pevarnek, and J. A. Halderman, "Green lights forever: analyzing the security of traffic infrastructure," in *8th USENIX Workshop on Offensive Technologies (WOOT 14)*, 2014.

[3] (2014, October 28) Sensys networks traffic sensor vulnerabilities (update a). https://ics-cert.us-cert.gov/advisories/ICSA-14-247-01A.

[4] G. Darroch, "Guess what's 'easily hacked'? yes, that's right: Smart city transport infrastructure," *The Register*. [Online]. Available: https://www.theregister.co.uk/2016/04/22/smart_transport_hackable/

[5] G. Wang, B. Wang, T. Wang, A. Nika, H. Zheng, and B. Y. Zhao, "Defending against sybil devices in crowdsourced mapping services," in *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 2016, pp. 179–191.

[6] M. T. Garip, M. E. Gursoy, P. Reiher, and M. Gerla, "Congestion attacks to autonomous cars using vehicular botnets," in *NDSS Workshop on Security of Emerging Networking Technologies (SENT), San Diego, CA*, 2015.

[7] M. B. Sinai, N. Partush, S. Yadid, and E. Yahav, "Exploiting social navigation," *arXiv preprint arXiv:1410.0151*, 2014.

[8] S. Hendrix, "Traffic-weary homeowners and Waze are at war, again. guess who's winning?" *The Washington Post*. [Online]. Available: https://www.washingtonpost.com/local/traffic-weary-homeowners-and-waze-are-at-war-again-guess-whos-winning/2016/06/05/c466df46-299d-11e6-b989-4e5479715b54_story.html

[9] A. A. Cárdenas and J. S. Baras, "Evaluation of classifiers: Practical considerations for security applications," in *Proc. AAAI Workshop Evaluation Methods for Machine Learning*, 2006, pp. 777–780.

[10] T. H. Ptacek and T. N. Newsham, "Insertion, evasion, and denial of service: Eluding network intrusion detection," DTIC Document, Tech. Rep., 1998.

[11] D. Wagner and P. Soto, "Mimicry attacks on host-based intrusion detection systems," in *Proceedings of the 9th ACM Conference on Computer and Communications Security*. ACM, 2002, pp. 255–264.

[12] A. A. Cárdenas, J. S. Baras, and K. Seamon, "A framework for the evaluation of intrusion detection systems," in *2006 IEEE Symposium on Security and Privacy (S&P'06)*. IEEE, 2006, pp. 15–pp.

[13] A. A. Cárdenas, S. Radosavac, and J. S. Baras, "Evaluation of detection algorithms for mac layer misbehavior: Theory and experiments," *IEEE/ACM Transactions on Networking*, vol. 17, no. 2, pp. 605–617, 2008.

[14] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," in *Proceedings of the 16th ACM Conference on Computer and Communications Security*, ser. CCS '09. New York, NY, USA: ACM, 2009, pp. 21–32.

[15] 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*, 2016, pp. 1092–1105.

[16] I. Sajjad, D. D. Dunn, R. Sharma, and R. Gerdes, "Attack mitigation in adversarial platooning using detection-based sliding mode control," in *Proceedings of the First ACM Workshop on Cyber-Physical Systems-Security and/or PrivaCy*. ACM, 2015, pp. 43–53.

[17] D. Mashima and A. A. Cárdenas, "Evaluating electricity theft detectors in smart grid networks," in *International Workshop on Recent Advances in Intrusion Detection*. Springer, 2012, pp. 210–229.

[18] C. Dwork, "Differential privacy," in *33rd International Colloquium on Automata, Languages and Programming- ICALP 2006*, 2006, pp. 1–12.

[19] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar, "Privacy, accuracy, and consistency too: a holistic solution to contingency table release," in *PODS '07: Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. New York, NY, USA: ACM, 2007, pp. 273–282.

[20] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee, "Toward privacy in public databases," in *Theory of Cryptography Conference*, Cambridge, MA, Feb. 9-12 2005. [Online]. Available: http://research.microsoft.com/research/sv/DatabasePrivacy/public.ps

[21] A. Inan, M. Kantarcioglu, G. Ghinita, and E. Bertino, "Private record matching using differential privacy," in *EDBT 2010, 13th International Conference on Extending Database Technology, Lausanne, Switzerland, March 22-26, 2010, Proceedings*, ser. ACM International Conference Proceeding Series, I. Manolescu, S. Spaccapietra, J. Teubner, M. Kitsuregawa, A. Léger, F. Naumann, A. Ailamaki, and F. Özcan, Eds., vol. 426. ACM, 2010, pp. 123–134. [Online]. Available: http://doi.acm.org/10.1145/1739041.1739059

[22] G. Jagannathan, K. Pillaipakkamnatt, and R. N. Wright, "A practical differentially private random decision tree classifier," *Transactions on Data Privacy*, vol. 5, no. 1, pp. 273–295, 2012. [Online]. Available: http://www.tdp.cat/issues11/abs.a082a11.php

[23] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[24] D. P. Bertsekas and A. Scientific, *Convex optimization algorithms*. Athena Scientific Belmont, 2015.

[25] R. Lanotte, M. Merro, and S. Tini, "Towards a formal notion of impact metric for cyber-physical attacks," in *International Conference on Integrated Formal Methods*. Springer, 2018, pp. 296–315.

[26] R. Mitchell, I.-R. Chen *et al.*, "Effect of intrusion detection and response on reliability of cyber physical systems," *IEEE Transactions on Reliability*, vol. 62, no. 1, pp. 199–210, 2013.

[27] J.-F. Chamberland and V. V. Veeravalli, "Decentralized detection in sensor networks," *IEEE Transactions on Signal Processing*, vol. 51, no. 2, pp. 407–416, 2003.

[28] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[29] M. Struwe and M. Struwe, *Variational methods*. Springer, 1990, vol. 31999.

[30] D. P. Bertsekas, A. Nedi, A. E. Ozdaglar *et al.*, *Convex analysis and optimization*. Athena Scientific, 2003.

[31] C. Dwork, "Differential privacy: A survey of results," in *Theory and applications of models of computation*. Springer, 2008, pp. 1–19.

[32] 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.

[33] 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*, ser. ASIACCS '11. New York, NY, USA: ACM, 2011, pp. 355–366.

[34] C. Murguia and J. Ruths, "Cusum and chi-squared attack detection of compromised sensors," in *2016 IEEE Conference on Control Applications (CCA)*, Sep. 2016, pp. 474–480.

[35] A. Wald and J. Wolfowitz, "Optimum character of the sequential probability ratio test," *The Annals of Mathematical Statistics*, pp. 326–339, 1948.

[36] A. Wald, *Sequential analysis*. Courier Corporation, 1973.

[37] A.-M. Olteanu, K. Huguenin, R. Shokri, M. Humbert, and J.-P. Hubaux, "Quantifying interdependent privacy risks with location data," *IEEE Transactions on Mobile Computing*, 2016.

[38] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux, "Quantifying location privacy," in *2011 IEEE Symposium on Security and Privacy*. IEEE, 2011, pp. 247–262.

[39] B. Coifman, "Vehicle re-identification and travel time measurement in real-time on freeways using existing loop detector infrastructure," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1643, pp. 181–191, 1998.

[40] B. Coifman and M. Cassidy, "Vehicle reidentification and travel time measurement on congested freeways," *Transportation Research Part A: Policy and Practice*, vol. 36, no. 10, pp. 899–917, 2002.

[41] K. Kwong, R. Kavaler, R. Rajagopal, and P. Varaiya, "Arterial travel time estimation based on vehicle re-identification using wireless magnetic sensors," *Transportation Research Part C: Emerging Technologies*, vol. 17, no. 6, pp. 586–606, 2009.

[42] J. Le Ny, A. Touati, and G. J. Pappas, "Real-time privacy-preserving model-based estimation of traffic flows," in *ICCPS'14: ACM/IEEE 5th International Conference on Cyber-Physical Systems (with CPS Week 2014)*. IEEE Computer Society, 2014, pp. 92–102.

[43] S. Datta, J. Carroll, V. Petit, B. Sathyamangalam, and M. G. R Hebner, "Attributes of direct measurement of inductance in a loop detector for traffic control," *CEM Publications*, 2015.

[44] J. Lee, R. Jiang, and E. Chung, "Traffic queue estimation for metered motorway on-ramps through use of loop detector time occupancies," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2396, pp. 45–53, 2013.

[45] E. Sullivan and M. Burris, "Benefit-cost analysis of variable pricing projects: Sr-91 express lanes," *Journal of transportation engineering*, vol. 132, no. 3, pp. 191–198, 2006.

[46] B. Coifman and S. Neelisetty, "Improved speed estimation from single-loop detectors with high truck flow," *Journal of Intelligent Transportation Systems*, vol. 18, no. 2, pp. 138–148, 2014.

[47] M. Treiber and A. Kesting, "Traffic flow dynamics," *Traffic Flow Dynamics: Data, Models and Simulation, Springer-Verlag Berlin Heidelberg*, 2013.

[48] 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.

[49] C. F. Daganzo, "The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory," *Transportation Research Part B: Methodological*, vol. 28, no. 4, pp. 269–287, 1994.

[50] N. Henze, "Empirical-distribution-function goodness-of-fit tests for discrete models," *Canadian Journal of Statistics*, vol. 24, no. 1, pp. 81–93, 1996.

[51] J. C. Herrera, D. B. Work, R. Herring, X. J. Ban, Q. Jacobson, and A. M. Bayen, "Evaluation of traffic data obtained via gps-enabled mobile phones: The mobile century field experiment," *Transportation Research Part C: Emerging Technologies*, vol. 18, no. 4, pp. 568–583, 2010.

[52] R. Tan, V. Badrinath Krishna, D. K. Yau, and Z. Kalbarczyk, "Impact of integrity attacks on real-time pricing in smart grids," in *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. ACM, 2013, pp. 439–450.

[53] M. Roozbehani, M. Rinehart, M. Dahleh, S. Mitter, D. Obradovic, and H. Mangesius, "Analysis of competitive electricity markets under a new model of real-time retail pricing," in *Energy Market (EEM), 2011 8th International Conference on the European*, may 2011, pp. 250–255.

[54] A. Molina-Markham, P. Shenoy, K. Fu, E. Cecchet, and D. Irwin, "Private memoirs of a smart meter," in *Proceedings of the 2nd ACM workshop on embedded sensing systems for energy-efficiency in building*. ACM, 2010, pp. 61–66.

[55] M. Jawurek, F. Kerschbaum, and G. Danezis, "Privacy technologies for smart grids - a survey of options," Tech. Rep. MSR-TR-2012-119, November 2012.

[56] G. Ács and C. Castelluccia, "I have a dream!(differentially private smart metering)," in *Information Hiding*. Springer, 2011, pp. 118–132.

[57] T. K. Moon and W. C. Stirling, *Mathematical methods and algorithms for signal processing*. Prentice hall, 2000.

[58] K. P. Schneider, Y. Chen, D. P. Chassin, R. Pratt, D. Engel, and S. Thompson, "Modern grid initiative distribution taxonomy final report," *PNNL-18035, Pacific Northwest National Laboratory, Richland, Washington*, 2008.

[59] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, "Certified robustness to adversarial examples with differential privacy," in *40th IEEE Symposium on Security and Privacy*, 2019.

[60] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin, "On evaluating adversarial robustness," *arXiv preprint arXiv:1902.06705*, 2019.

[61] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. Long Beach, California, USA: PMLR, 09–15 Jun 2019, pp. 1310–1320. [Online]. Available: http://proceedings.mlr.press/v97/cohen19c.html

[62] R. Pinot, L. Meunier, A. Araujo, H. Kashima, F. Yger, C. Gouy-Pailler, and J. Atif, "Theoretical evidence for adversarial robustness through randomization: the case of the exponential family," *arXiv preprint arXiv:1902.01148*, 2019.

[63] C. Chen, J. Kwon, J. Rice, A. Skabardonis, and P. Varaiya, "Detecting errors and imputing missing data for single-loop surveillance systems," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1855, pp. 160–167, 2003.

[64] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady, "Preserving privacy in gps traces via uncertainty-aware path cloaking," in *Proceedings of the 14th ACM conference on Computer and communications security*. ACM, 2007, pp. 161–171.

[65] B. Hoh, M. Gruteser, R. Herring, J. Ban, D. B. Work, J. C. Herrera, A. M. Bayen, M. Annavaram, and Q. Jacobson, "Virtual trip lines for distributed privacy-preserving traffic monitoring," in *Proceedings of the 6th International Conference on Mobile Systems, Applications, and Services (MobiSys 2008), Breckenridge, CO, USA, June 17-20, 2008*, 2008, pp. 15–28.

[66] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," in *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. ACM, 2013, pp. 901–914.

[67] D. J. Mir, S. Isaacman, R. Caceres, M. Martonosi, and R. N. Wright, "Dp-where: Differentially private modeling of human mobility," in *Big Data, 2013 IEEE International Conference on*. IEEE, 2013, pp. 580–588.

[68] R. Chen, G. Acs, and C. Castelluccia, "Differentially private sequential data publication via variable-length n-grams," in *Proceedings of the 2012 ACM conference on Computer and communications security*. ACM, 2012, pp. 638–649.