Privacy-Preserving Obfuscation of Critical Infrastructure Networks (Extended Version)

Ferdinando Fioretto , **Terrence W.K. Mak** and **Pascal Van Hentenryck**

Georgia Institute of Technology

{fioretto, wmak}@gatech.edu, pvh@isye.gatech.edu

Abstract

The paper studies how to release data about a critical infrastructure network (e.g., a power network or a transportation network) without disclosing sensitive information that can be exploited by malevolent agents, while preserving the realism of the network. It proposes a novel obfuscation mechanism that combines several privacy-preserving building blocks with a bi-level optimization model to significantly improve accuracy. The obfuscation is evaluated for both realism and privacy properties on real energy and transportation networks. Experimental results show the obfuscation mechanism substantially reduces the potential damage of an attack exploiting the released data to harm the real network.

1 Introduction

Critical Infrastructure Networks (CINs), such as electrical power grids and public transportation networks, rely on the tight interaction of cyber and physical components. They play crucial roles in ensuring social and economic stability, and their operations require advanced machine-learning and optimization algorithms. For instance, power network operations perform a power flow computation every few minutes.

Research on CINs is highly dependent on the availability of realistic test cases. However, the release of such datasets is challenging due to privacy and national security concerns. For instance, the electrical load of an industrial customer may indirectly reveal sensitive information on its production levels and strategic investments. Similarly, the ability to link the physical and cyber locations of power generators can be used to launch coordinated attacks on cyber-physical facilities, significantly damaging the targeted network.

To mitigate these concerns, this paper develops an obfuscation mechanism based on *Differential Privacy* (DP) [Dwork *et al.*, 2006], a robust framework that bounds the privacy risks associated with answering sensitive queries or releasing datasets. A DP algorithm introduces carefully calibrated noise to the data to prevent the disclosure of sensitive information. It is immune to *linkage attacks*–attacks in which one exploits auxiliary data to expose sensitive information.

However, DP faces significant challenges when the resulting privacy-preserving datasets are used as inputs to complex
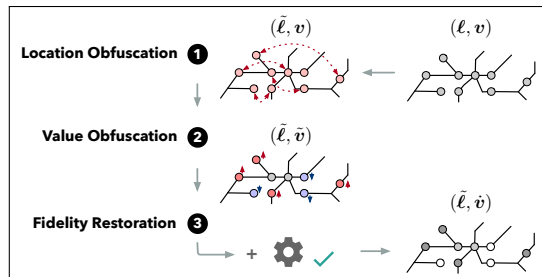

Figure 1: The POCIN Framework

optimization algorithms on CINs. For instance, direct applications of DP may impact the realism of the original dataset or produce networks that do not admit feasible solutions for the problem of interest [Fioretto and Van Hentenryck, 2018].

The paper addresses this gap by introducing a *Privacy-preserving Obfuscation mechanism for CINs* (POCINs), sketched in Figure 1. POCIN takes as input a CIN and executes three phases to (1) obfuscate the *sensitive location* of network elements, (2) hide *sensitive values*, and (3) ensure the satisfaction of important global properties of the CIN and the problem of interest, through the use of optimization.

The paper makes the following contributions: (1) It proposes POCIN, a novel privacy-preserving data release scheme that protects parameters and locations of network elements; (2) POCIN uses a novel bi-level optimization approach to preserve salient properties of the released data and solves it with either exact or approximated methods; and (3) It applies POCIN to two *real* CINs from energy and transportation networks. The paper shows that POCIN ensures strong privacy guarantees and that, on the considered CINs, *the damage inflicted on the real network, when the attacker exploits the obfuscated data, converges to that of a random uninformed attack* as the privacy requirements increase.

2 Preliminaries: Differential Privacy

Differential Privacy (DP) is a framework used to protect the privacy of individuals in a dataset. This notion has emerged as the de-facto standard for privacy protection. A randomized mechanism $\mathcal{M} : \mathscr{D} \to \mathscr{R}$ with domain $\mathscr{D}$ and range $\mathscr{R}$ is ϵ-differential private if, for any output response $O \subseteq \mathscr{R}$ and any two *neighboring* inputs $D, D' \in \mathscr{D}$ differing in at most

one individual (written $D \sim D'$),

$$Pr[\mathcal{M}(D) \in O] \leqslant \exp(\epsilon) \, Pr[\mathcal{M}(D') \in O], \quad (1)$$

where the probability is calculated over the coin tosses of $\mathcal{M}$. The parameter $\epsilon > 0$ is the *privacy loss* of the algorithm, with values close to 0 denoting strong privacy. DP satisfies several important properties, including *composability*, which ensures that a combination of differentially private algorithms preserve differential privacy, and *immunity to post-processing*, which ensures that privacy guarantees are preserved by arbitrary post-processing steps [Dwork and Roth, 2013].

In private data analysis settings, users interact with datasets by issuing queries. A (numeric) *query* is a function from a data set $D \in \mathscr{D}$ to a result set $R \subseteq \mathbb{R}^d$. A query Q can be made differentially private by injecting random noise to its output. The amount of noise depends on the *sensitivity* of the query, denoted by Δ_Q and defined as $\Delta_Q = \max_{D \sim D'} \|Q(D) - Q(D')\|_1$. In other words, the sensitivity of a query is the maximum l_1-distance between the query outputs of any two neighboring dataset D and D'.

While the classical DP notion protects individuals from participating into a dataset, many applications involve components whose *presence* is public information. However, their values and (cyber) locations are highly sensitive, as they may reveal, for instance, how much power a generator is producing and where it is located, or the most congested segments in transportation or logistics network. Therefore the privacy goal of data curators is to protect observed values and locations associated with these components.

The concept of *indistinguishability* was introduced by Andrés *et al.* [2013] to protect user locations in the Euclidean plane and then generalized in [Chatzikokolakis *et al.*, 2013; Koufogiannis *et al.*, 2015] to arbitrary metric spaces. Consider a dataset D to which n individuals contribute their data x_i and $\alpha > 0$. An adjacency relation that captures the data variation of a single individual is defined as:

$$D \sim_\alpha D' \iff \exists i : d(x_i, x_i') \leqslant \alpha \wedge \forall j \neq i : d(x_j, x_j') = 0$$

where d is a distance function on $\mathscr{D}$. Such adjacency definition is useful to hide individual participation up to some quantity α, or a location within a radius α. Given $\epsilon > 0, \alpha > 0$, a randomized mechanism $\mathcal{M} : \mathscr{D} \to \mathscr{R}$ with domain $\mathscr{D}$ and range $\mathscr{R}$ is α-*indistinguishable ϵ-DP* ((ϵ, α)-indistinguishable for short) if, for any event $O \subseteq \mathscr{R}$ and any pair $D \sim_\alpha D'$, $(D, D' \in \mathscr{D})$, Equation (1) holds.

3 The CIN Obfuscation Problem

Problem Setting Consider a dataset $(\mathcal{N}, \mathcal{E})$ describing some critical infrastructure network where $\mathcal{N} \in \mathbb{R}^{n \times p}$ and $\mathcal{E} \in \mathbb{R}^{m \times q}$ describe the set of n nodes and m edges of the network, together with their attributes. The dataset $(\mathcal{N}, \mathcal{E})$ is referred to as *CIN description*. An element n_i of $\mathcal{N}$ is a p-dimensional vector describing the attribute values associated with node i and an element e_{ij} of $\mathcal{E}$, is a q-dimensional vector describing the attribute values associated with edge (i, j) between nodes i and j. Furthermore, consider an optimization problem P that takes as input a CIN description, denoted
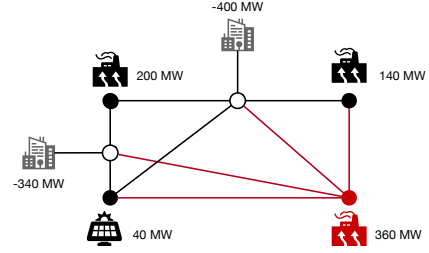

Figure 2: An example CIN with four nodes (power generators). The highlighted node is one being compromised by an attacker.

with $\boldsymbol{G}$ for notational clarity:

$$\mathcal{O}^*(\boldsymbol{G}) = \min_{\boldsymbol{x}} \quad \mathcal{O}(\boldsymbol{x}, \boldsymbol{G}) \\ \text{s.t.} \quad g_i(\boldsymbol{x}, \boldsymbol{G}) \leqslant 0 \ \ \forall i \in [k], \quad (2)$$

where $\boldsymbol{x} \in \mathbb{R}^l$ is a vector of decision variables, $\mathcal{O}$ is the objective function of P, and g_i $(i \in [k])$ are the problem constraints.

To illustrate these concepts, consider the CIN in Figure 2, representing a simplified *power network*. The nodes describe the network *generators* and *loads* (represented as cities in the figure) with values denoting the amount of power being injected into, or withdrawn from, the power grid. The edges represent the transmission lines to carry power from generators to loads. The optimization problem P may be the *optimal power flow* (OPF) that amounts to finding the most cost-effective generator dispatch to serve the load demands while satisfying the physical constraints of the network. The data curator desires to release a CIN description $(\tilde{\mathcal{N}}, \tilde{\mathcal{E}})$ that *obfuscates* the *location* and *value* of some *sensitive* network elements, e.g., the locations and capacities of generators. The curator also wants to ensure that the OPFs on the released and original CINs behave similarly (e.g., are feasible and have similar costs) so that the released data is of high *fidelity*.

For notational simplicity, this paper focuses on obfuscating *sensitive nodes* of a CIN where each node $n_i \in \mathcal{N}$ is a pair $(\ell_i, v_i) \in \mathbb{N} \times \mathbb{R}$, describing its location and salient parameter. The n-dimensional vectors of locations and values are denoted by $\boldsymbol{\ell}$ and $\boldsymbol{v}$ and $\boldsymbol{G}$ is used to denote a CIN description.

Attack Model The paper considers an *attack model* $\mathcal{A}$ in which a malicious agent can disrupt up to b elements (called *attack budget*) to damage the network. Such action produces a new damaged network $\boldsymbol{G}^a = \mathcal{A}(\boldsymbol{G})$ with $\boldsymbol{G}^a \subseteq \boldsymbol{G}$. For instance, Figure 2 highlights in red the generators affected by an attack with a budget $b = 1$, while $\boldsymbol{G}^a$ is shown in black. The paper further assumes that the attacker has full knowledge of the network topology and is capable of solving problem P (e.g., to find the generators with highest dispatch) and assessing the damages resulting from an attack. The *damage* of the attacker is measured using the objective value difference $\mathcal{O}^*(\boldsymbol{G}^a) - \mathcal{O}^*(\boldsymbol{G})$ for problem P. For simplicity, the paper focuses on *ranking attacks*, i.e., attacks that disrupt the b highest-ranked network elements when their values are sorted according to an ordering $\prec$ (e.g., the most dispatched generators), and evaluates the released data against such attacks.

The CIN Obfuscation Problem This paper uses α-indistinguishability to obfuscate the values and locations of

nodes. For a given $\alpha_\ell > 0$, the relation $\sim_\ell \subseteq \mathbb{N}^2$ captures indistinguishability between node positions in the network, using a distance d_ℓ defined as a function of the minimum number of nodes separating two nodes. Similarly, for a given $\alpha_v > 0$ and distance function d_v, relation $\sim_v \subseteq \mathbb{R}^2$ captures the indistinguishability between node values. These neighboring relations can be combined into a new relation $\sim_{\ell v}$ that captures indistinguishability for both values and locations. Two CIN descriptions $\boldsymbol{G} = (\boldsymbol{\ell}, \boldsymbol{v}), \boldsymbol{G}' = (\boldsymbol{\ell}', \boldsymbol{v}') \in \mathbb{N} \times \mathbb{R}$ are *location and value indistinguishable* if and only if

$$\boldsymbol{G} \sim_{\ell v} \boldsymbol{G}' \Leftrightarrow \ell \sim_\ell \ell' \vee \boldsymbol{v} \sim_v \boldsymbol{v}'. \tag{3}$$

A randomized mechanism $\mathcal{M}$ is $(\epsilon, \alpha_\ell, \alpha_v)$-indistinguishable if, for any pair of $\sim_{\ell v}$-adjacent datasets, Equation (1) holds.

The design of obfuscation mechanisms for CIN descriptions should satisfy three desiderata, formalized through the *Privacy-preserving Obfuscation for CIN* (POCIN) problem:

Definition 1 *Given a CIN description $\boldsymbol{G}$, a problem P, and positive real values $\alpha_v, \alpha_\ell, \beta$ and ϵ, the POCIN problem produces an obfuscated CIN description $\tilde{\boldsymbol{G}} = (\tilde{\boldsymbol{\ell}}, \tilde{\boldsymbol{v}})$ such that:*

1. Privacy*: $\tilde{\boldsymbol{G}}$ satisfies $(\epsilon, \alpha_\ell, \alpha_v)$-indistinguishability.*
2. Fidelity*: $\tilde{\boldsymbol{G}}$ admits a candidate solution $\bar{\boldsymbol{x}}$ that (i) satisfies the constraints $g_i(\bar{\boldsymbol{x}}, \tilde{\boldsymbol{G}})$ of P and (ii) is β-faithful to the P's objective, i.e., $\frac{|\mathcal{O}(\bar{\boldsymbol{x}}, \tilde{\boldsymbol{G}}) - \mathcal{O}^*(\boldsymbol{G})|}{\mathcal{O}^*(\boldsymbol{G})} \leqslant \beta$.*
3. Robustness*: $\tilde{\boldsymbol{G}}$ minimizes $|\mathcal{O}^*(\boldsymbol{G}) - \mathcal{O}^*(\mathcal{A}(\tilde{\boldsymbol{G}}))|$, i.e., the damage inflicted by attack $\mathcal{A}$.*

4 The POCIN Mechanism

The POCIN mechanism is divided in three phases, as illustrated in Figure 1:

1. The *location obfuscation* phase produces a new location indistinguishable CIN $\boldsymbol{G}^\ell = (\tilde{\boldsymbol{\ell}}, \boldsymbol{v})$ from the original $\boldsymbol{G}$.
2. The *value obfuscation* phase takes $\boldsymbol{G}^\ell$ as input and produces a new value indistinguishable CIN $\tilde{\boldsymbol{G}} = (\tilde{\boldsymbol{\ell}}, \tilde{\boldsymbol{v}})$.
3. The *fidelity restoration* phase produces a new CIN description $\dot{\boldsymbol{G}} = (\tilde{\boldsymbol{\ell}}, \dot{\boldsymbol{v}})$ that satisfies the problem constraints and is faithful to its objective.

Phase 1: Location Obfuscation
The first phase provides location indistinguishability. The idea is to shuffle the node locations using an instance of the *Exponential Mechanism* [McSherry and Talwar, 2007]. Such mechanism releases a privacy-preserving answer to a query by sampling from its output discrete space O. The sampling probability is determined by a *utility function* $u : (\mathscr{D} \times O) \to \mathbb{R}$ that assigns a score to each output $o \in O$.

Theorem 1 (Exponential Mechanism) *Let $u : (\mathscr{D} \times O) \to \mathbb{R}$ with sensitivity $\Delta_u = \max_{o \in O} \max_{D \sim D'} |u(D, o) - u(D', o)|$. The exponential mechanism that outputs o with probability $\Pr[o$ is selected $] \propto \exp\left(\epsilon u(D, o)/2\Delta_u\right)$ satisfies $\epsilon - DP$.*

POCIN uses a privacy-preserving *shuffling function* that maps each node i with location ℓ_i to a new location $\tilde{\ell}_i$ as:

$$\tilde{\ell}_i \leftarrow \ell_j \text{ with } \Pr \propto \exp\left(\frac{\epsilon d_\ell(\ell_i, \ell_j)}{2\alpha_\ell}\right) \tag{4}$$
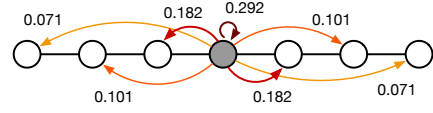

Figure 3: Probabilities of location obfuscations ($\alpha_\ell = \epsilon = 1.0$).

where ℓ_j is the location of a node j. This is an application of the Exponential mechanism with Δ_u being the sensitivity of distance function d_ℓ. Since two different locations may be mapped to the same location, POCIN solves an *assignment problem* so that each node i is mapped to a unique location.

Corollary 1 *Given $\alpha_\ell > 0$, Phase 1 of POCIN achieves α_ℓ-location-indistinguishability.*

Figure 3 illustrates the intuition underlying the mechanism: It displays the probabilities of moving the central node to a specific location. For a sufficiently large α_ℓ, an attacker is provided with no useful location information. Of course, there is a tradeoff between α_ℓ and the fidelity of the network, which is addressed in Phase 3.

Phase 2: Value Obfuscation
The second phase of POCIN takes as input $\boldsymbol{G}^\ell = (\tilde{\boldsymbol{\ell}}, \boldsymbol{v})$ and constructs a privacy-preserving answer $\tilde{\boldsymbol{v}}$ for the node values $\boldsymbol{v}$ using the Laplace mechanism.

Theorem 2 (Laplace Mechanism) *Let Q be a numeric query that maps datasets to $\mathbb{R}^d$. The Laplace mechanism that outputs $Q(D) + z$, where $z \in \mathbb{R}^d$ is drawn from the Laplace distribution $Lap(\Delta_Q/\epsilon)^d$, achieves ϵ-DP.*

In the above, $\text{Lap}(\lambda)$ denotes the Laplace distribution with 0 mean and scale λ, and $\text{Lap}(\lambda)^d$ denotes the i.i.d. Laplace distribution over d dimensions with parameter λ. The Laplace mechanism with parameter $\lambda = \alpha/\epsilon$ satisfies α-indistinguishability [Chatzikokolakis *et al.*, 2013]. As a result, the privacy-preserving values $\tilde{\boldsymbol{v}}$ of the CIN description are obtained as follows:

$$\tilde{\boldsymbol{v}} = \boldsymbol{v} + \text{Lap}(\alpha_v/\epsilon)^n. \tag{5}$$

The Laplace mechanism has been shown to be *optimal*: it minimizes the mean-squared error for identity queries with respect to the L1-norm [Koufogiannis *et al.*, 2015].

Corollary 2 *Given $\alpha_v > 0$, Phase 2 of POCIN achieves α_v-value-indistinguishability.*

Phase 2 generates a CIN description $\tilde{\boldsymbol{G}} = (\tilde{\boldsymbol{\ell}}, \tilde{\boldsymbol{v}})$ obfuscating locations and values of the original network $\boldsymbol{G}$. As a result, $\boldsymbol{G}^v$ satisfies *condition (1)* of the CIN obfuscation problem.

Phase 3: Fidelity Restoration
The noise introduced by Phases 1 and 2 may significantly alter the structure of the data and the solutions to Problem P. To restore the fidelity of the network, POCIN leverages the post-processing immunity of DP and uses a bi-level optimization problem to redistribute the noise introduced in the previous phases. The primary decision variables of the problem are the vector $\dot{\boldsymbol{v}} = (\dot{v}_1, \ldots, \dot{v}_n)$ that represents the post-processed node values after the noise redistribution.

$$P_{\text{BL}} = \min_{(\dot{\boldsymbol{v}}, \boldsymbol{x})} \|\dot{\boldsymbol{v}} - \tilde{\boldsymbol{v}}\|_2 \qquad \text{(b1)}$$

$$\text{s.t.:} \ |\mathcal{O}(\boldsymbol{x}^*, \dot{\boldsymbol{v}}) - \mathcal{O}^*| \leqslant \beta \qquad \text{(b2)}$$

$$\boldsymbol{x}^* = \operatorname{argmin}_{\boldsymbol{x}} \ \mathcal{O}(\boldsymbol{x}, \dot{\boldsymbol{v}}) \atop \text{s.t.} \ g_i(\boldsymbol{x}, \dot{\boldsymbol{v}}) \leqslant 0 \ \forall i \in [k] \qquad \text{(b3)}$$

Figure 4: The bi-level optimization of the Fidelity Restoration.

The problem, called P_{BL}, is shown in Figure 4. It searches for a vector $\dot{\boldsymbol{v}}$ for which problem P has an optimal solution $\boldsymbol{x}^*$ and whose objective value is close to the original optimum $\mathcal{O}^*$ (assumed to be public information). Moreover, the vector $\dot{\boldsymbol{v}}$ must be as close as possible to obfuscated vector $\tilde{\boldsymbol{v}}$, which is ensured by objective (b1). Constraint (b3) ensures that $\boldsymbol{x}^*$ is an optimal solution to problem $\mathcal{O}^*(\dot{\boldsymbol{G}} = (\tilde{\boldsymbol{\ell}}, \dot{\boldsymbol{v}}))$, and Constraint (b2) ensures the fidelity of the objective.

Theorem 3 *POCIN is $(\epsilon, \alpha_p, \alpha_v)$-indistinguishable.*

Theorem 4 *The error induced by POCIN on the CIN node values is bounded by the inequality: $\|\dot{\boldsymbol{v}} - \boldsymbol{v}\|_2 \leqslant 2\|\tilde{\boldsymbol{v}} - \boldsymbol{v}\|_2$.*

Note that POCIN does not explicitly minimize the damages inflicted by an attack on the critical infrastructure network. Indeed, simply evaluating the damage of an attack requires access to the real network $\boldsymbol{G}$, which would violate privacy. However, the shuffling mechanism indirectly addresses this issue. Intuitively, for a sufficiently large α_ℓ in the location obfuscation, the probability of choosing the most dispatched generator with a single attack will be close to $\frac{1}{n}$.

5 Solving the Fidelity Restoration Model

P_{BL} is a *bi-level problem* and bi-level programming is known to be strongly NP-hard [Hansen *et al.*, 1992]: Even evaluating a solution for optimality is NP-hard [Vicente *et al.*, 1994]. This paper explores two avenues to solve or approximate P_{BL}.

When the subproblem (b3) is convex, it can be replaced by its *Karush-Kuhn-Tucker* (KKT) conditions, producing a single-level model. Moreover, when the subproblem is linear, the resulting subproblem is a MIP model, which is the case in our transportation case study. The subproblem can also be replaced by the *relaxation* of P_{BL} shown in Figure 5. This relaxation ensures that $\dot{\boldsymbol{G}}$ has a feasible solution whose value is close to the original objective. However, it does not constrain the optimal solution which will be a lower bound to $\mathcal{O}(\boldsymbol{x}^*, \dot{\boldsymbol{v}})$. This relaxation is used in the power network case study and also satisfies Theorem 4. When the subproblem is convex, it is possible to strengthen Theorem 4 using the existence of a solution (vector $\boldsymbol{v}$), the optimality of $\dot{\boldsymbol{v}}$, and the angle property of a projection on a convex set.

Theorem 5 *When P_{CL} is used for fidelity restoration and is convex, the error induced by POCIN on the CIN node values is bounded by the inequality: $\|\dot{\boldsymbol{v}} - \boldsymbol{v}\|_2 \leqslant \|\tilde{\boldsymbol{v}} - \boldsymbol{v}\|_2$.*

Other methods to solve bi-level problems include *descent methods* [Kolstad and Lasdon, 1990], *penalty function methods* [Aiyoshi and Shimizu, 1984], and *evolutionary approaches* [Mathieu *et al.*, 1994]. The reader can consult [Sinha *et al.*, 2018] for an extensive review on the topic.

$$P_{\text{CL}} = \min_{(\dot{\boldsymbol{v}}, \boldsymbol{x})} \|\dot{\boldsymbol{v}} - \tilde{\boldsymbol{v}}\|_2 \qquad \text{(n1)}$$

$$\text{s.t.:} \ |\mathcal{O}(\boldsymbol{x}, \dot{\boldsymbol{v}}) - \mathcal{O}^*| \leqslant \beta \qquad \text{(n2)}$$

$$g_i(\boldsymbol{x}, \dot{\boldsymbol{v}}) \leqslant 0 \ \forall i \in [k] \qquad \text{(n3)}$$

Figure 5: The Relaxation of P_{BL}.

Network instance	α				
	1	3	5	7	10
nesta_case14_ieee	6	6	6	6	6
nesta_case30_ieee	0	0	0	0	0
nesta_case57_ieee	12	10	6	12	12
nesta_case118_ieee	0	0	0	0	0

Table 1: Feasibility Statistics before Phase 3 (in percentage).

6 Experimental Results

This section presents experimental results on two case studies: power network and traffic obfuscation problems.

6.1 Power Network Obfuscation Problem

This scenario considers a transmission operator who would like to release a description of its network to stimulate research but is worried that malicious actors could use it to design a cyber-attack targeting the generators that would cause maximum damage. The operator seeks to obfuscate the locations and maximum capacities of its generators while preserving the realism of the network description. The experiments were performed on a variety of benchmarks from the NESTA library [Coffrin *et al.*, 2014]. The results analyze the dispatch values produced by POCIN and determine how well the obfuscated networks sustain an attack. For brevity, the results are highlighted on the IEEE 118 bus test case. The experiments use a privacy loss ϵ of 1.0, and vary the *indistinguishability levels* α_l from 1% to 10% of the network diameter $d(\boldsymbol{G})$, and the *faithfulness level* β in $\{10^{-2}, 10^{-1}\}$, while α_v is fixed to 0.1 p.u. $\approx$ 10 MW. The model was implemented using the Julia package PowerModels.jl with IPOPT [Coffrin *et al.*, 2018] for solving the nonlinear AC power flow models. POCIN takes less than 1 minute to produce the obfuscated networks.

Dispatch & Cost Analysis The section studies the privacy and realism of the obfuscated networks produced by POCIN. Figure 6a shows the active dispatch values of all generators in the obfuscated network and compares them with their associated values in the original network. The left plot compares the dispatch for each generator, while the right plot compares the generation on each bus. Note that POCIN only swaps generator locations and do not assign generators to arbitrary buses. Not surprisingly, the dispatch differences are more pronounced as the indistinguishability level increases. *More importantly, the dispatch values can no longer be used to distinguish generators and/or buses.*

Table 1 reports the success rate of obtaining a feasible AC power flow after Phase 2 on various benchmarks over 50 runs. Without Phase 3, the optimization problem P is infeasible in most or all runs for all test cases. *These results highlight*
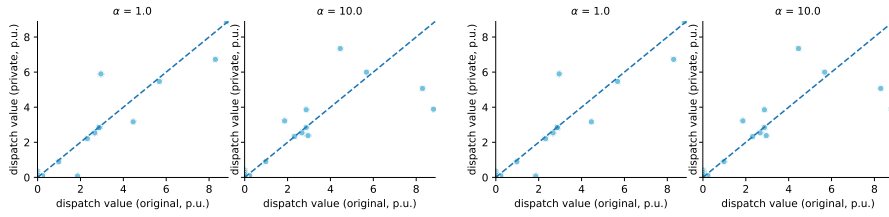
(a) Active dispatch between original and obfuscated networks with $\epsilon = 1.0$, $\beta = 0.01$, $\alpha_l = \alpha\% \times d(\boldsymbol{G})$, $\alpha_v = 0.1$. Left: Generator active generation. Right: Bus active injection.
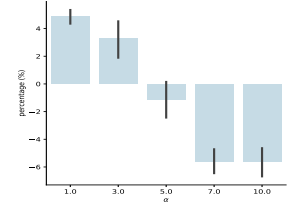
(b) AC optimal dispatch costs differences (in %) between original and obfuscated networks, with $\beta = 0.1$

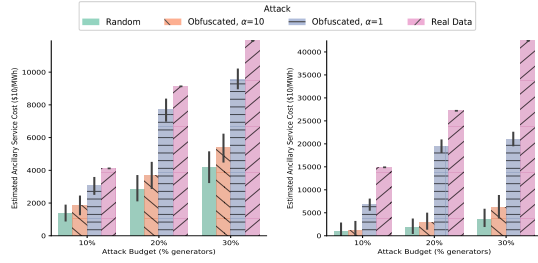Figure 6: (IEEE 118 bus) Dispatch & Cost Analysis

Figure 7: Ancillary service costs for the attack type on the IEEE-57 (left) and IEEE-118 (right) Networks ($\beta = 0.1$).

the critical role of Phase 3 of POCIN. Figure 6b shows the difference, in percentage, between the dispatch costs of the obfuscated and original networks for increasing the indistinguishability levels $\alpha_l : \alpha_l = \alpha\% \times d(\boldsymbol{G})$. It reports the mean and standard deviation (shown with black, solid, lines) over 50 runs. *The results indicate that obfuscated networks are largely within the faithfulness requirement and provide the necessary fidelity for Problem P on the obfuscated network.*

Attack Simulation It remains to study how an attacker may leverage the obfuscated network to damage the original CIN. The attacker is given an *attack budget b* denoting the percentage of generators that can be damaged. To assess the benefits of POCIN, three type of attacks are compared (where k denote the number of generators that makes up $b\%$).

1. *Random Attack*: k generators are randomly selected.
2. *Obfuscated Attack*: The attacker chooses the k generators with the largest dispatches in $\dot{\boldsymbol{G}}$.
3. *Fully-Informed Attack*: The attacker chooses the k generators with the largest dispatches in $\boldsymbol{G}$.

The experiments assume estimated ancillary service costs of $10/MWh to serve the load that cannot be provided by the remaining generators. Figure 7 shows the costs for each attack type at varying of the attack budget $b \in \{10, 20, 30\}$ and the indistinguishability value $\alpha \in \{1.0, 10.0\}, \alpha_l = \alpha\% \times d(\boldsymbol{G})$ on the IEEE-57 and IEEE-118 benchmarks with faithfulness value β and α_v both set to 0.1. The results average 50 simulations for each combination of parameters. The *random attacks* are used as a baseline to assess the damage inflicted by an uninformed attacker. Not surprisingly, they result in the lowest costs in each setting. In contrast, *fully-informed attacks* produce the largest damages on the networks with in-

creasing ancillary costs as the attack budget increases. The results for the *obfuscated attacks* are particularly interesting: The obfuscation significantly reduces the power of an attacker. *Remarkably, as the location indistinguishability values increase, the inflicted damages decrease and converge to that of random attacks.* Larger indistinguishability implies more noise, and thus a higher chance for an attacker to damage less important generators. Note however that *the mechanism still preserves the desired network fidelity.*

6.2 Traffic Network Obfuscation Problem

The scenario involves a city that would like to release its traffic data, but is concerned about a cyber-attack on its traffic control system. The city aims at releasing an obfuscated version of the data that would preserve the trip durations of its commuters and would minimize the damage of a cyber-attack on its traffic controllers regulating traffic flows on the road segments. The case study involves the traffic network of the city of Ann Arbor, MI and 8,000 *real trips* from an origin O to a destination D (O-D pairs). For simplicity, the travel times on an edge e are given by a linear combination $d_e + \gamma t_e$ of the distance d_e and the traffic t_e, but the results generalize naturally to more complex models. POCIN obfuscates the location of the edges and their transit data, the distance and the network being public information. Problem P is characterized by solving a shortest path for each commuter from her origin to her destination, using historical traffic data. Since the shortest path problem is totally unimodular, the bi-level program P_{BL} of POCIN can be reformulated as a MIP, and thus the problem can be solved exactly. The experiments use a privacy loss of $\epsilon = 1.0$, vary the *indistinguishability level* α_l from 1% to 25% of the number of nodes and select the *faithfulness level* β of 0.1. The implementation takes less than 20 minutes to generate the obfuscated networks.

Traffic Weights & Travel Costs The first experiments concern the privacy and realism of the obfuscated network produced by POCIN. Figure 8 shows the travel times on all the roads in the obfuscated and original networks. *As the indistinguishability levels increase, the networks become significantly different.* Figure 9 shows the shortest paths cost of 15 O-D pairs in the obfuscated and original networks. Yellow/'+' dots are travel costs before Phase 3 and blue dots are the travel costs of POCIN. *The results show that POCIN preserves the shortest paths costs with high fidelity and that the fidelity restoration step brings significant benefits.*
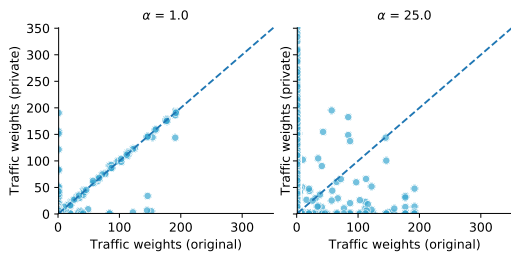
Figure 8: Traffic weights on all the roads between original and obfuscated networks with faithfulness parameters $\beta = 0.1$.
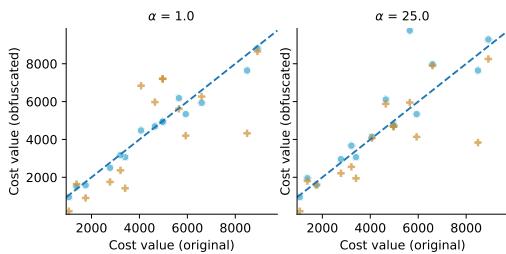

Figure 9: Travel costs of 15 trips between original and obfuscated networks with faithfulness parameters $\beta = 0.1$.

Attack Simulation The second set of results concerns the obfuscation capability to mitigate the damage of an attack. A malicious agent targets the traffic controllers of various road segments to increase the total travel costs of all O-D pairs. Figure 10 shows the average increases (in %) of the travel costs for each attack strategy, at various attack budgets consisting of $k \in \{10, 20, 50\}$ roads, indistinguishability value α, and with faithfulness value $\beta = 0.1$. Larger values of β attain similar results. Results are averages over 50 simulations for each combination of parameters. The results show that random attacks are completely ineffective, while fully-informed attacks increase travel times substantially. In contrast, the obfuscation dramatically decreases the damage. When the indistinguishability level increases, the obfuscation eliminates the damage for attack budgets targeting 10 roads and only increases the travel times by less than 20% on larger attacks.

7 Related Work

The release of differentially private datasets that protects the *value* of the data is a challenging task that has been studied by several authors. Often the released data is generated from a data synopsis in the form of a noisy histogram [Li *et al.*, 2014; Qardaji *et al.*, 2014; Fioretto *et al.*, 2018]. Protecting locations privacy typically uses the framework of *geo-indistinguishability* Andrés *et al.* [2013], in which locations are made indistinguishable in the Euclidean plane.

The release of differentially private networks has also been studied in the literature: It primarily focuses on social networks [Proserpio *et al.*, 2014; Blocki *et al.*, 2013; Kasiviswanathan *et al.*, 2013]. These methods focus on either protecting the privacy of the *node* or the *edge* and ensure that the output distribution does not change significantly when a
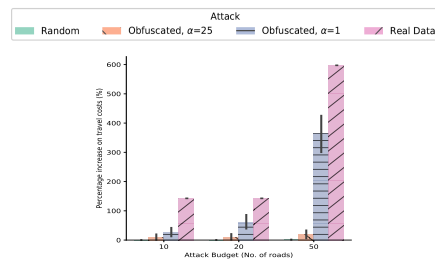

Figure 10: Average increase on travel times (in %) of all O-D pairs with faithfulness parameters $\beta = 0.1$.

node (resp. edge) and all its adjacent edges (resp. nodes) are added to, or removed from, the graph. The goal of these methods is not to preserve the original network topology. In contrast to these approaches, POCIN focuses on the release of privacy-preserving network data that protects locations and values of the network components while preserving the network topology and fidelity. As shown by the experimental results, this places additional requirements on the mechanism, including the need for POCIN to use optimization to redistribute the noise and ensure that the data release admits a realistic solution to the optimization problem.

The evaluation of privacy-preserving techniques with respect to malicious attacks is also a novel contribution of this paper. Traditional research on attack detection and prevention on cyber-physical systems is not concerned with privacy and data release. For instance Ghafouri *et al.* [2018] uses regression to detect anomalous sensor readings and Junejo and Goh [2016] focuses on predicting attacks observing real and simulated data. Finally, bi-level problems have been used in security applications of AI [Pita *et al.*, 2009; Jain *et al.*, 2011; Nguyen *et al.*, 2016; Zhao *et al.*, 2017].

8 Conclusions

This paper presented a privacy-preserving scheme for the release of *Critical Infrastructure Networks* (CINs). The proposed *Privacy-preserving Obfuscation mechanism for CIN* (POCIN) obfuscates values and locations of sensitive network elements combining several differential privacy mechanisms with a bi-level optimization problem. It does so without altering the CIN topology and ensuring that the released obfuscated network preserves the fundamental properties of an optimization problem of interest.

The paper proposes exact and relaxation solutions for solving the bi-level optimization problems. The POCIN mechanism was tested on realistic test cases for the power grid and transportation system. The results show that POCIN is effective in obfuscating values and locations of the network parameters. Importantly, the result also illustrates the effectiveness of POCIN to deceive malicious agents that exploit the data release to produce as much damage as possible to the CIN.

References

[Aiyoshi and Shimizu, 1984] Eitaro Aiyoshi and Kiyotaka Shimizu. A solution method for the static constrained stackelberg problem via penalty method. *AC*, 29(12):1111–1114, 1984.

[Andrés *et al.*, 2013] Miguel Andrés, Nicolás Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. In *CCS*, pages 901–914, 2013.

[Blocki *et al.*, 2013] Jeremiah Blocki, Avrim Blum, Anupam Datta, and Or Sheffet. Differentially private data analysis of social networks via restricted sensitivity. In *ITCS*, pages 87–96, 2013.

[Chatzikokolakis *et al.*, 2013] Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. Broadening the scope of differential privacy using metrics. In *PETS*, pages 82–102, 2013.

[Coffrin *et al.*, 2014] Carleton Coffrin, Dan Gordon, and Paul Scott. Nesta, the NICTA energy system test case archive. *CoRR*, abs/1411.0359, 2014.

[Coffrin *et al.*, 2018] Carleton Coffrin, Russell Bent, Kaarthik Sundar, Yeesian Ng, and Miles Lubin. Powermodels.jl: An open-source framework for exploring power flow formulations. In *PSCC*, pages 1–8, June 2018.

[Dwork and Roth, 2013] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Theoretical Computer Science*, 9(3-4):211–407, 2013.

[Dwork *et al.*, 2006] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, volume 3876, pages 265–284, 2006.

[Fioretto and Van Hentenryck, 2018] Ferdinando Fioretto and Pascal Van Hentenryck. Constrained-based differential privacy: Releasing optimal power flow benchmarks privately. In *CPAIOR*, pages 215–231, 2018.

[Fioretto *et al.*, 2018] Ferdinando Fioretto, Chansoo Lee, and Pascal Van Hentenryck. Constrained-based differential privacy for private mobility. In *AAMAS*, pages 1405–1413, 2018.

[Ghafouri *et al.*, 2018] Amin Ghafouri, Yevgeniy Vorobeychik, and Xenofon D. Koutsoukos. Adversarial regression for detecting attacks in cyber-physical systems. In *IJCAI*, pages 3769–3775, 2018.

[Hansen *et al.*, 1992] Pierre Hansen, Brigitte Jaumard, and Gilles Savard. New branch-and-bound rules for linear bilevel programming. *SIAM J. Sci. and Stat. Comput.*, 13(5):1194–1217, 1992.

[Jain *et al.*, 2011] Manish Jain, Dmytro Korzhyk, Ondřej Vaněk, Vincent Conitzer, Michal Pěchouček, and Milind Tambe. A double oracle algorithm for zero-sum security games on graphs. In *AAMAS*, pages 327–334, 2011.

[Junejo and Goh, 2016] Khurum Nazir Junejo and Jonathan Goh. Behaviour-based attack detection and classification in cyber physical systems using machine learning. In *CPS-Sec*, pages 34–43, 2016.

[Kasiviswanathan *et al.*, 2013] Shiva Prasad Kasiviswanathan, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Analyzing graphs with node differential privacy. In *Theory of Cryptography*, pages 457–476. Springer, 2013.

[Kolstad and Lasdon, 1990] Charles D Kolstad and Leon S Lasdon. Derivative evaluation and computational experience with large bilevel mathematical programs. *J. of optimization theory and applications*, 65(3):485–499, 1990.

[Koufogiannis *et al.*, 2015] Fragkiskos Koufogiannis, Shuo Han, and George J Pappas. Optimality of the laplace mechanism in differential privacy. *arXiv:1504.00065*, 2015.

[Li *et al.*, 2014] Chao Li, Michael Hay, Gerome Miklau, and Yue Wang. A data-and workload-aware algorithm for range queries under differential privacy. *VLDB*, 7(5):341–352, 2014.

[Mathieu *et al.*, 1994] R Mathieu, L Pittard, and G Anandalingam. Genetic algorithm based approach to bi-level linear programming. *RAIRO*, 28(1):1–21, 1994.

[McSherry and Talwar, 2007] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *FOCS*, pages 94–103, 2007.

[Nguyen *et al.*, 2016] Thanh Hong Nguyen, Arunesh Sinha, and Milind Tambe. Conquering adversary behavioral uncertainty in security games: An efficient modeling robust based algorithm. In *AAAI*, 2016.

[Pita *et al.*, 2009] James Pita, Manish Jain, Fernando Ordóñez, Christopher Portway, Milind Tambe, Craig Western, Praveen Paruchuri, and Sarit Kraus. Using game theory for los angeles airport security. *AI magazine*, 30(1):43, 2009.

[Proserpio *et al.*, 2014] Davide Proserpio, Sharon Goldberg, and Frank McSherry. Calibrating data to sensitivity in private data analysis: a platform for differentially-private analysis of weighted datasets. *VLDB*, 7(8):637–648, 2014.

[Qardaji *et al.*, 2014] Wahbeh Qardaji, Weining Yang, and Ninghui Li. Priview: practical differentially private release of marginal contingency tables. In *ICMD*, pages 1435–1446, 2014.

[Sinha *et al.*, 2018] Ankur Sinha, Pekka Malo, and Kalyanmoy Deb. A review on bilevel optimization: from classical to evolutionary approaches and applications. *IEEE Trans. on Evol. Comp.*, 22(2):276–295, 2018.

[Vicente *et al.*, 1994] Luis Vicente, Gilles Savard, and Joaquim Júdice. Descent approaches for quadratic bilevel programming. *J. of optimization theory and applications*, 81(2):379–399, 1994.

[Zhao *et al.*, 2017] Mengchen Zhao, Bo An, Wei Gao, and Teng Zhang. Efficient label contamination attacks against black-box learning models. In *IJCAI*, pages 3945–3951, 2017.

A Missing Proofs

First, we review some important properties of DP.

Composability ensures that a combination of differentially private algorithms preserve differential privacy.

Theorem 6 (Parallel Composition) *Let D_1 and D_2 be disjoint subsets of D and $\mathcal{A}$ be an ϵ-DP algorithm. Computing $\mathcal{A}(D \cap D_1)$ and $\mathcal{A}(D \cap D_2)$ satisfies ϵ-DP.*

Post-processing immunity ensures that privacy guarantees are preserved by arbitrary post-processing steps.

Theorem 7 (Post-Processing Immunity) *Let $\mathcal{A}$ be an ϵ-DP algorithm and g an arbitrary mapping from the set of possible outputs $\mathcal{O}$ to an arbitrary set. Then, $g \circ \mathcal{A}$ is ϵ-DP.*

Corollary 1 *Given $\alpha_\ell > 0$, Phase 1 of POCIN achieves α_ℓ-location-indistinguishability.*

Proof. Observe that the Location Obfuscation phase (Phase 1) combines the exponential mechanism with a post-processing step. Therefore, the result follows by Theorem **??**, parallel composition (Theorem 6), and post-processing immunity of differential privacy (Theorem 7). □

Corollary 2 *Given $\alpha_v > 0$, Phase 2 of POCIN achieves α_v-value-indistinguishability.*

Proof. Observe that the Value Obfuscation phase (Phase 2) is an application of the Laplace Mechanism with scaling factor α_v/ϵ (Equation (4): main article). The result follows from Theorem 2 and by observing that the Laplace Mechanism with parameter α/ϵ achieves (ϵ, α)-indistinguishability [Chatzikokolakis *et al.*, 2013]. □

Theorem 3 *POCIN is $(\epsilon, \alpha_p, \alpha_v)$-indistinguishable.*

Proof. From Corollaries 1 and 2 the application of Phase 1 and Phase 2 of POCIN achieves (ϵ, α_ℓ)-indistinguishability for locations and (ϵ, α_v)-indistinguishability for values, respectively. Therefore, from Equation (3) (main article) the Phase 1 and 2 of POCIN are location and value indistinguishable and achieve $(\epsilon, \alpha_p, \alpha_v)$-indistinguishability. Finally, Phase 3 uses exclusively differentially private outputs and public information. Therefore the result follows from post-processing immunity of differential privacy (Theorem 7). □

The proof of the next theorem is similar to the proof of (Theorem 5) in [Fioretto and Van Hentenryck, 2018].

Theorem 4 *The error induced by POCIN on the CIN node values is bounded by the inequality: $\|\dot{v}^* - v\|_2 \leqslant 2\|\tilde{v} - v\|_2$. where $\dot{v}^*$ is the optimal solution to problem P_{BL}.*

Proof. The results follows from the fact that v is feasible for the lower-level problem, the optimality of $\dot{v}^*$ for P_{BL}, and the triangle inequality on norms. □

Theorem 5 *When P_{CL} is used for fidelity restoration and is convex, the error induced by POCIN on the CIN node values is bounded by the inequality:*

$$\|\dot{v}^* - v\|_2 \leqslant \|\tilde{v} - v\|_2.$$

where $\dot{v}^*$ is the solution to the problem P_{CL}.

Proof. Let $\mathbb{C}_P \in \mathbb{R}^n$ be the feasible set of P_{CL}, i.e., $\mathbb{C}_P = \{v | \exists x \text{ such that } (n2) - (n3) \text{ hold}\}$. By definition, $\mathbb{C}_P$ is closed, convex, and non-empty, since v is a feasible solution. Since v is the original vector, then $v \in \mathbb{C}_P$. The addition of noise to v produces a new vector $\tilde{v} \in \mathbb{R}^n$ that may or may not satisfy the problem constraints.

- Case 1: $\tilde{v} \in \mathbb{C}_P$. That is, $\tilde{v}$ satisfies the constraints of P_{CL}. The minimizer for P_{CL} is thus the vector $\dot{v}^* = \tilde{v}$ with objective value 0. Therefore:

$$\|\dot{v}^* - v\|_2 = \|\tilde{v} - v\|_2.$$

- Case 2: $\tilde{v} \notin \mathbb{C}_P$, i.e., $\tilde{v}$ lies outside the feasible region $\mathbb{C}_P$. Problem P_{CL} can be seen as a *projection problem* onto a convex set, i.e.,

$$\text{Proj}_{\mathbb{C}_p}(x) = \operatorname*{argmin}_{v \in \mathbb{C}_p} \|x - v\|_2.$$

The variational characterization of projection implies that the solution $\dot{v}^*$ satisfies $\langle v - \dot{v}^*, \tilde{v} - \dot{v}^* \rangle \leqslant 0$ and the result follows from the triangle inequality on norms. □