Cyber-physical risk modeling with imperfect cyber-attackers

Efthymios Karangelos and Louis Wehenkel,
Department of EE & CS - Montefiore Institute,
University of Liège, Belgium.
{e.karangelos;l.wehenkel}@uliege.be

***Abstract*—We model the risk posed by a malicious cyber-attacker seeking to induce grid insecurity by means of a *load redistribution* attack, while explicitly acknowledging that such an actor would plausibly base its decision strategy on imperfect information. More specifically, we introduce a novel formulation for the cyber-attacker's decision-making problem and analyze the distribution of decisions taken with randomly inaccurate data on the grid branch admittances or capacities, and the distribution of their respective impact. Our findings indicate that inaccurate admittance values most often lead to suboptimal cyber-attacks that still compromise the grid security, while inaccurate capacity values result in notably less effective attacks. We also find common attacked cyber-assets and common affected physical-assets between all (random) imperfect cyber-attacks, which could be exploited in a preventive and/or corrective sense for effective cyber-physical risk management.**

***Index Terms*—Cyber-physical system, power system security, risk modeling.**

I. Introduction

The digitalization of electric power system control & communications is reshaping the scope for security management. In addition to *physical* threats (*e.g.*, failures of the physical infrastructure, forecasting errors, *etc.*), securing the system against *cyber* threats has also become essential [1]. Such threats notably include the adversary actions of malicious external agents, seeking to exploit weaknesses in the security of the system cyberspace so as to disrupt the physical supply of electricity. Going from physical to cyber-physical security management therefore requires modeling not only the interdependencies between cyber and physical infastructures, but also the interaction between *cyber-attackers* and *grid-operators*.

A. Related literature

The framework of multi-level optimization is most commonly used in the state of the art to model the cyber-attacker *vs* grid-operator interaction [2], [3]. The foundations of this literature lie in [4], showing that an attacker can successfully introduce false data without being detected in the case of linear state estimation. Accordingly, bilevel formulations including a malicious attacker as the upper agent and a grid-operator solving the linear *DC Optimal Power Flow* (DC-OPF) problem have been proposed to model alternative cyber-physical attack scenarios. Yuan *et al.* introduced in [5] the concept of *load redistribution* while modeling an attacker seeking to falsify load measurements so as to provoke out-of-merit generation dispatching by the grid-operator. The same concept was exploited in [6] by an attacker seeking to maximize the loading of a transmission line. Zhang and Sankar developed in [7] an elaborate bilevel formulation for an attacker using load redistribution in order to hide a physical change in the grid topology, while a model of an attacker seeking to maximize the number of overloads induced through load redistribution has been presented in [8].

We underline the shared assumption of a cyber-attacker relying on *perfect* information (*i.e.*, the correct model of the power system and of its operator) to determine its attack strategy. However, Rahman and Mohsenian-Rad [9] stress that realistically *imperfect* cyber-attackers would have to rely on an inaccurate grid model, as they cannot be plausibly assumed to observe in real-time the status of every circuit breaker, tap-changer *etc.*. These authors modeled an attacker with incorrect line admittance data and found that the probability of detecting a load redistribution attack designed with such imperfect information remains rather low. Reference [10] provides further evidence for challenging the perfect information assumption by means of a sensitivity analysis with respect to the attacker's knowledge of the occurrence of a single line outage, showing that the incorrect grid topology undermines the evaluation of the cyber-physical attack. Sanjab and Saad studied in [11] the interaction between a defender taking preventive actions and potential realistic cyber-attackers with limited system knowledge and found that the Nash-equilibrium strategy against the assumed perfect, fully rational attacker is not the best defense of the grid.

B. Paper scope & contributions

In this paper we focus on cyber-physical risk modeling while explicitly acknowledging that a realistic cyber-attacker would plausibly base her strategy on imperfect information. We investigate the effect of such imperfect information in terms of: i) the impact of a cyber-physical attack on the electricity transmission grid and ii) the various attack vectors that may be launched by a malicious cyber-attacker. The former relates primarily to cyber-physical risk assessment applications, while the latter allows to draw conclusions for

This work has been prepared with the support of the Belgian Energy Transition Fund, project CYPRESS (https://cypress-project.be/).

cyber-physical risk management. Our analysis is based on the well-studied scenario of load redistribution attacks wherein a malicious attacker seeks to deceive the grid-operator with the final purpose of rendering the grid insecure. On top of imperfect information on the network branch admittances (as in [9]), we posit that a realistic cyber-attacker may also rely on inaccurate branch capacity data. Indeed, the ability of transmission branches to securely sustain loading relies on ambient conditions and so does in practice the tolerance of grid-operators for higher loading levels when these ambient conditions are favourable.

To perform our investigation we use the *Monte Carlo* framework while modeling sequentially the decisions of a (random) imperfect cyber-attacker, the corresponding reaction of the grid-operator and finally the resulting state of the grid. The core component of our framework is a bilevel "$\max\min$" optimization model of a cyber-attacker, anticipating the reaction of the grid-operator to a load redistribution attack. In addition to the standard constraints included in the state of the art of such models, we formalize here novel constraint expressions to reflect a malicious cyber-attacker with the intention of inducing a challenging insecure state, potentially triggering a cascading failure event.

Our analysis showcases that (even minor) informational imperfections imply a broad spectrum of potential cyber-attacks and of respective physical impacts on the electricity system. Moreover, the spectrum of potential cyber-attacks clearly features groups of common attacked assets in the cyber sub-system and common affected assets in the physical sub-system. The implication is that protecting the cyber sub-system to avoid/detect intrusion of such common attacked assets and/or the physical sub-system to withstand the possible failure of such common affected assets could be effective cyber-physical risk management strategies.

C. Paper organization

Section II introduces the model of a malicious cyber-attacker seeking to maximize grid insecurity on the basis of the data she perceives to be true. Section III presents the proposed methodology and metrics for acknowledging the informational imperfections of such an actor in the context of risk assessment and risk management. Section IV discusses the application of such methodology on the single area version IEEE-RTS96 benchmark [12] while conclusions are drawn in Section V.

II. Cyber-attacker decision-making problem

We model a malicious cyber-attacker seeking to maximize the grid physical insecurity through a load redistribution attack. More specifically, we consider an attacker falsifying bus load measurements so as to mislead the grid-operator into perceiving the grid as insecure and implementing unnecessarely generation redispatch actions. The cyber-attacker's objective is to maximize the total magnitude of branch overloads caused by the injection of false measurements and the resulting generation redispatching of the mislead grid-operator. Invoking the DC-power flow approximation, we cast the cyber-attacker's decision making problem as the following bi-level *Mixed Integer Linear Programming* (MILP) problem:

$$\max \sum_{\ell \in \mathcal{L}} r_\ell \tag{1}$$

$$\sum_{\ell \in \mathcal{L}} \left(u_\ell^+ + u_\ell^-\right) \geq U \tag{2}$$

$$\sum_{n \in \mathcal{N}} a_n \leq A \tag{3}$$

$$\sum_{n \in \mathcal{N}} e_n = 0 \tag{4}$$

for all nodes $n \in \mathcal{N}$:

$$-a_n \cdot \epsilon \cdot d_n \leq e_n \leq a_n \cdot \epsilon \cdot d_n \tag{5}$$

$$a_n \in \{0,1\} \tag{6}$$

$$\sum_{g \in \mathcal{G}} \gamma_{g,n} \left(p_{g0} + p_g^\star\right) - \sum_{\ell \in \mathcal{L}} \lambda_{\ell,n} \cdot f_\ell^{ca} = d_n \tag{7}$$

for all branches $\ell \in \mathcal{L}$:

$$f_\ell^{ca} = (1/X_\ell) \cdot \sum_{n \in \mathcal{N}} \lambda_{\ell,n} \cdot \theta_n^{ca} \tag{8}$$

$$u_\ell^+ + u_\ell^- + u_\ell^0 = 1 \tag{9}$$

$$f_\ell^{ca} - \rho_\ell \cdot \overline{f}_\ell \leq u_\ell^+ \cdot M \tag{10}$$

$$f_\ell^{ca} - \rho_\ell \cdot \overline{f}_\ell \geq (u_\ell^+ - 1) \cdot M \tag{11}$$

$$-f_\ell^{ca} - \rho_\ell \cdot \overline{f}_\ell \leq u_\ell^- \cdot M \tag{12}$$

$$f_\ell^{ca} + \rho_\ell \cdot \overline{f}_\ell \geq (1 - u_\ell^-) \cdot M \tag{13}$$

$$r_\ell \leq (1 - u_\ell^0) \cdot M \tag{14}$$

$$(u_\ell^+ - 1) \cdot M + (f_\ell^{ca} - \overline{f}_\ell) \leq r_\ell \tag{15}$$

$$r_\ell \leq (1 - u_\ell^+) \cdot M + (f_\ell^{ca} - \overline{f}_\ell) \tag{16}$$

$$(u_\ell^- - 1) \cdot M - (f_\ell^{ca} + \overline{f}_\ell) \leq r_\ell \tag{17}$$

$$r_\ell \leq (1 - u_\ell^-) \cdot M - (f_\ell^{ca} + \overline{f}_\ell) \tag{18}$$

$$u_\ell^+, u_\ell^-, u_\ell^0 \in \{0,1\} \tag{19}$$

subject to the model of the mislead grid-operator:

$$\min \sum_{g \in \mathcal{G}} c_g \cdot \pi_g \tag{20}$$

for all generators $g \in \mathcal{G}$:

$$\pi_g \geq 0 \tag{21}$$

$$\pi_g \geq p_g^\star \tag{22}$$

$$(\underline{p}_g - p_{g0}) \leq p_g^\star \leq (\overline{p}_g - p_{g0}) \tag{23}$$

for all nodes $n \in \mathcal{N}$:

$$\sum_{g \in \mathcal{G}} \gamma_{g,n} \left(p_{g0} + p_g^\star\right) - \sum_{\ell \in \mathcal{L}} \lambda_{\ell,n} f_\ell^{go} = d_n + e_n \tag{24}$$

for all branches $\ell \in \mathcal{L}$:

$$f_\ell^{go} = (1/X_\ell) \cdot \sum_{n \in \mathcal{N}} \lambda_{\ell,n} \cdot \theta_n^{go} \tag{25}$$

$$-\overline{f}_\ell \leq f_\ell^{go} \leq \overline{f}_\ell. \tag{26}$$

$\mathcal{G}$ set of generating units;

$\mathcal{L}$ set of transmission branches;

$\mathcal{N}$ set of nodes;

r_{ℓ} upper-level continuous variable, measuring the magnitude of the branch overloads induced by the attack;

$u_{\ell}^{\cdot}$ upper-level binary variable, indicating the overload status of branch ℓ, with superscripts $(+/-)$ for an overloaded branch in the positive/negative flow direction or 0 for no overload;

U parameter, modeling the minimum number of overloaded branches targeted by the attacker;

a_n upper-level binary variable, indicating the injection of false data at the load measurement of node n;

A parameter, modeling the attacker's available budget for attacking the grid load meters;

e_n upper-level continuous variable, modeling the false active power demand measurement data injected by the attacker at node n;

ϵ parameter, modeling the maximum relative amount of false load measurement data that can be injected by the attacker;

d_n parameter, modeling the active power demand at node n;

$\gamma_{g,n}$ parameter, modeling the connectivity of generator g with node n;

p_{g0} parameter, modeling the dispatch of generator g;

p_g lower-level continuous variable, modeling the active power redispatch of generator g by the grid-operator;

$\lambda_{\ell,n}$ parameter, modeling the connectivity of branch ℓ with node n and the assumed flow direction;

f_{ℓ}^{ca} upper-level continuous variable, modeling the cyber-attacker's perceived active power flow value through branch ℓ;

X_{ℓ} parameter, modeling the reactance of branch ℓ;

θ_n^{ca} upper-level continuous variable, modeling the cyber-attacker's perceived voltage angle value at node n;

ρ_{ℓ} parameter, modeling the minimum threshold of overloaded flow per branch targeted by the attacker;

$\overline{f}_{\ell}$ parameter, modeling the capacity of branch ℓ;

M a large constant parameter;

c_g parameter, modeling the non-negative upward redispatch marginal cost of generator g;

π_g lower-level continuous variable, modeling the upward redispatch of generator g;

$\underline{p}_g$ parameter, modeling the minimum stable output of generator g;

$\overline{p}_g$ parameter, modeling the capacity of generator g;

f_{ℓ}^{go} lower-level continuous variable, modeling the grid-operator's perceived active power flow value through branch ℓ;

θ_n^{go} upper-level continuous variable, modeling the grid-operator's voltage angle value at node n.

Objective function (1) seeks to maximize the total magnitude of the branch overloads induced by the cyber-attack. We introduce inequality constraint (2) to model that a malicious cyber-attacker may strategically prefer to overload at least a minimum number of branches ($U \geq 2$) in order to create an overwhelming grid insecurity instance outside the *"comfort zone"* of N-1 security.

Expression (3) imposes a limit on the maximum number of load meters that can be manipulated by the attacker, while (4) enforces that the false load measurement data injection is balanced across the grid and (5) sets the maximum relative amount of false data that can be injected by the attacker at any node[1]. Equalities (7,8) model the power flow of the grid as perceived by the cyber-attacker only. Notice that the power balance constraint (7) includes the optimal values of the generation redispatch variables $(p_g^{\star})$ as decided in the grid-operator's lower-level problem (20 – 26). We adopt here the so-called *optimistic* bilevel optimization framework [14], implying that if the cyber-attacker's strategy yields multiple optimal solutions for (20 – 26), the optimistic cyber-attacker believes that the choice of the grid-operator $(p_g^{\star})$ will be the one most suiting objective (1).

The group of inequalities (10 – 13) is used to flag overloaded branches either in the positive or in the negative flow direction, while (14 – 18) are used to measure the magnitude of the branch overloads caused by the cyber attack. Here we originally introduce a parameter $(\rho_{\ell} \geq 1)$ to model that a malicious cyber-attacker may strategically prefer to cause an overloaded flow larger than a threshold on every overloaded branch in order to create an overwhelming grid insecurity instance. Indeed, by way of (10 – 18), only overloads above such threshold contribute in the right-hand-side of constraint (2) and objective function (1). To the best of our knowledge, the consolidation of (1 – 2, 9 – 19) to establish the number and minimum magnitude of branch overloads sought by a cyber-attacker constitutes a new formulation for the load redistribution attack problem.

The lower-level problem (20 – 26) is a standard DC-OPF problem modeling the reaction of the grid-operator to the injection of the false data by the attacker, seeking to minimize the cost of upward generation redispatching so as to maintain all perceived (*i.e.*, false) branch flow values within the respective capacity ratings. The cyber-attacker's decision strategy appears as the false data injection variable (e_n) in the right-hand-side of the power balance constraint (24), while supersrcipt $({}^{go})$ denotes the (false) branch flow and voltage angle values perceived by the grid-operator.

III. Models & metrics for cyber-attacks with imperfect information

We follow the Monte Carlo approach while sampling random error terms for the grid parameters to reflect that a cyber-attacker with imperfect information would base her decisions on randomly inaccurate grid parameter values. More specifically, we assume that the branch admittances or transmission capacities may be imperfectly known by the attacker and

[1] As discussed in [13] constraints (4,5) are the standard *proxy* constraints for the undetectability of a load redistribution attack in the DC model.

form a simulation sample by drawing a unique, uniformly distributed, error term per branch. For each simulation sample, we model the sequence of a cyber-attack as detailed in III-A. To analyze the resulting distribution of cyber-attacks in the context of cyber-physical risk assessment and risk management we introduce novel metrics in sections III-B and III-C respectively.

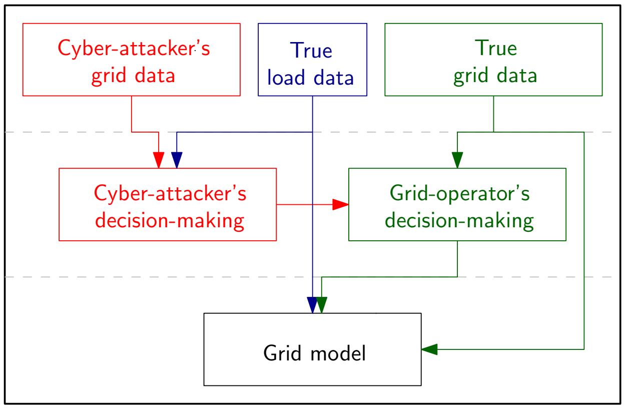
Cyber-attacker's grid data
True load data
True grid data
Cyber-attacker's decision-making
Grid-operator's decision-making
Grid model

Fig. 1. Cyber-attack with imperfect information modeling

A. Cyber-attack sequence modeling

Fig. 1 presents the proposed flowchart for modeling a cyber-attack with imperfect information.

At the top layer of this figure we distinguish between two different datasets for the grid, the so-called *"cyber-attacker's"* and *"true"* grid data. The former includes the (possibly inaccurate) data on the electricity grid that a cyber-attacker would exploit to decide her cyber-attack strategy. The latter includes the correct values for all the parameters of the electricity grid and is only available to the grid-operator. Notice that the load demand data is not included in any of these two datasets as it is assumed to be a "true" set of data only known to the cyber-attacker.

The middle layer of Fig. 1 models the interaction between the cyber-attacker and the grid-operator by means of two distinctive decision-making models, which are solved in sequence. First, the red box *"cyber-attacker's decision-making"* corresponds to bilevel model (1 – 26) discussed in section II, used by the cyber-attacker to identify her attack vector for load redistribution (horizontal red arrow). The green box *"grid-operator's decision-making"* models the reaction of the grid-operator to the cyber-attack. We must stress here that, even though a model for the grid-operator's reaction is embedded in the cyber-attacker's problem (1 – 26), the true reaction of the grid-operator to the attack vector will be based on the true grid data she has access to. Therefore, to model such reaction, we solve here the lower-level optimization problem (20 – 26) only, given the optimal attack vector from the solution of (1 – 26) and all parameters from the *"true"* grid dataset[2].

The lower layer of Fig. 1 illustrates a model of the physical impact of the cyber-attack on the electricity transmission grid, which is solved by combining: i) the true load data, ii) the true grid data and iii) the redispatching decisions the grid-operator would take given the load redistribution attack and her knowledge of the true grid data. Seeking to isolate the effect of imperfect information, in our implementation we combine such inputs through the same physical model as in the cyber-attacker's decision-making problem (*i.e.*, the DC power flow equations) to measure grid insecurity in terms of the number and magnitude of overloaded branches[3].

B. Metrics related to cyber-physical risk assessment

Cyber-physical security assessment serves to quantify the threat posed by a malicious cyber-attacker. The paradigm of the perfect information cyber-attack (*i.e.*, a cyber-ataker having access to the *"true"* grid data) is commonly employed in assessment applications, to anticipate the *worst-case* physical impact on the electricity system. Acknowledging a cyber-attacker's imperfect information yields a set of random cyber-attack samples and respective impact indicators. Beyond the expected value and distribution of the impact indicators over the Monte Carlo samples, we propose to analyze the *risk* of a cyber-attack with imperfect information by means of the following exclusive categories.

- *Perfect*: all samples wherein the attack vector of an imperfect cyber-attack matches the vector from the perfect information cyber-attack.
- *Success*: all other samples wherein an imperfect cyber-attack would still achieve the cyber-attacker's goals in terms of minimum number of overloaded branches with a flow above the respective threshold.
- *Partial success*: all other samples wherein an imperfect cyber-attack results in overloading at least one transmission branch with a flow above the respective threshold.
- *Failure*: all samples wherein an imperfect cyber-attack would cause no branch overload.
- *No attempt*: all samples wherein the cyber-attacker, given her imperfect information, fails to identify a feasible cyber-attack on the grid.

The share of samples in the first category shows the relevance of the *worst-case* perfect information cyber-attack, or alternatively the relevance of acknowledging a cyber-attacker's informational imperfections. Note that this category does not only include instances wherein the cyber-attacker's grid data randomly turn out to be perfectly accurate, but also instances wherein the cyber-attacker's informational imperfections have

[2] We should also acknowledge that restricting the physical models and equations in the grid-operator's decision-making model to match those of the cyber-attacker is not necessary by default. We made such choice here so as to isolate the impact of inaccurate data, and refer the reader to [15] for a study of the impact of simplifications in the cyber-attacker's modeling of a grid-operator's decision-making.

[3] An alternative physical model, more detailed than the one used by the cyber-attacker, may well be relevant for generally assessing the system vulnerabilities as shown in [6].

no effect on her strategy. A larger share of samples in the last two categories indicates that the cyber-physical electricity system is inherently more secure, either by way of “absorbing” the physical impact of imperfect cyber-attacks or by way of appearing more robust to the cyber-attacker.

C. *Metrics related to cyber-physical risk management*

Cyber-physical risk management serves to efficiently protect the grid from the threat of a malicious cyber-attacker. The perfect information cyber-attacker is commonly used in respective applications to anticipate a *worst-case* atttack vector against which resources should be deployed in advance and/or prepared to be deployed. Facing a distribution of imperfect attackers translates into a distribution of attack vectors which we will classify by way of i) the assets in the cyber sub-system potentially targeted to launch (imperfect) cyber-attacks, and ii) the assets of the physical sub-system that would undergo the physical impact of cyber-attacks.

The first classification is more relevant for the deployment of preventive countermeasures on the cyber sub-system, in order to impede a successful cyber-physical attack. In the considered load redistribution attack mode, we propose to rank the system loads in terms of the share of attack vectors wherein their respective measurements are tampered with. In other words, rank the system loads in order of attack likelihood so as to efficiently select which load measurements to protect from falsifying. Noting that protecting (a sub-set of) the measurements under attack may suffice to render a load redistribution attack detectable, we will further count the share of imperfect attack vector samples that target an increasing sub-set (*i.e.*, from at least one to all) of measurements in common with the perfect information cyber-attack.

We finally propose to rank groups of transmission grid branches in terms of the share of instances wherein all branches in a group would undergo an overload following a cyber-attack. Such ranking can be used to design effective emergency control strategies for the physical sub-system, so as to alleviate overloads in a timely manner before triggering cascading failure events.

IV. Case Studies

A. *Test case setup*

We adopt the single-area version (24 bus) of the IEEE-RTS96 benchmark[4]. Following the practice of relevant studies (*e.g.*, [5], [6], [8]) we simulate a stressed operational condition by reducing all branch transmission capacities to 65% of the original values. We further model a malicious cyber-attacker seeking to overload at least $U = 2$ transmission elements to at least $\rho_\ell = 5\%$ of the respective capacities. We set the cyber-attacker's resource constraint to falsifying at most $A = 10$ distinct load measurements and the maximum relative amount of false data per measurement to $\epsilon = 20\%$.

B. *Perfect information load redistribution attack*

Under the assumed conditions a cyber-attacker with perfect information, solving model (1 – 26) with the correct values for all grid parameters, would indeed be able to induce 2 overloads in the grid by more than 5% of the respective branch capacities. More specifically, the cyber-attacker would provoke erroneous redispatch by the grid-operator eventually overloading branch 12 to 109.1% of its capacity and branch 23 to 118.6% of its capacity. The total magnitude of measurable overloads (*i.e.*, above the 5% threshold) would amount to 48.8 MW. Figure 2 illustrates the optimal attack vector, with the x-axis showing the index of the affected bus load meter and the y-axis the percentage of change in the falsified load data.

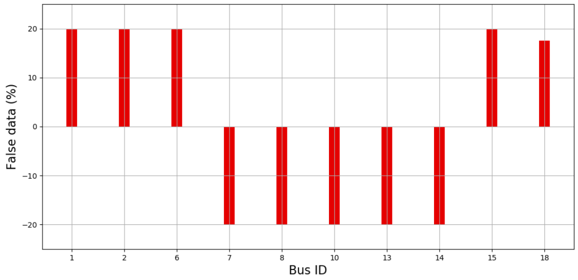

Fig. 2. Perfect information optimal attack vector

C. *Cyber-attacks with imperfect information on the grid admittances only*

We start by considering that a cyber-attacker may rely on inaccurate data considering the grid admittances only. To do so, we derive 10000 inaccurate grid samples, by applying a distinct error term to the admittance value of each branch, which is uniformly distributed in the range ±10%. Performing the respective simulations, we found that such (moderate) inaccuracy translates into 2677 (out of 10000) unique load redistribution attack vectors, with an average impact (*i.e.*, total measurable overload) of 28.36 MW. The histogram in Fig. 3 shows the distribution of the impact of such potential attacks, which as anticipated ranges from 0 (for the case of not attempted or failed attacks) to the upper-bound set by the perfect information cyber-attack.

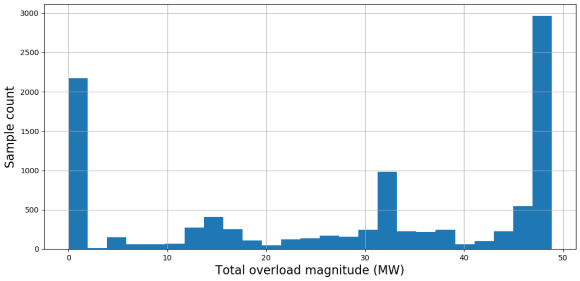

Fig. 3. Impact distribution of cyber-attacks with imperfect admittance data

[4] All system data can be found at https://matpower.org/docs/ref/matpower5.0/case24_ieee_rts.html.

We further assess the risk posed by the imperfect cyber-attacker by means of the categories introduced in section III-B through the pie-chart in Fig. 4. As shown in this chart, due to the assumed informational imperfections the cyber-attacker would only be able to correctly identify the optimal perfect information attack vector from Fig. 2 on 23.4% of the simulated instances. Conversely, on 15% of the sampled instances the cyber-attacker would falsely believe that it would be fruitless to launch any load redistribution attack while on 6.5% of the instances, she would launch an attack that would not be harmful to the grid. Observing that on 40.9% of the instances a cyber-attack with imperfect information would cause an overflow on at least two grid branches, while on 78.5% it would cause an overflow on at least one branch, we may infer from Fig. 4 that this system is insecure[5].

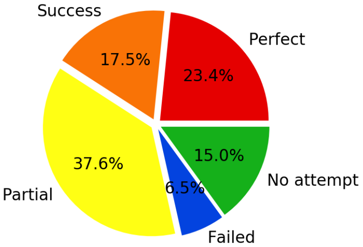

Fig. 4. Classification of cyber-attacks with imperfect admittance data

Pursuing the analysis from a risk management perspective, Fig. 5.a shows the relative frequency of attacking each distinct bus load meter amongst the 10000 sampled cyber-attacks. The bars in blue correspond to the perfect information optimal attack vector from Fig. 2 and it is notable that these are the meters ranked first in order of decreasing frequency. Specifically, the least-frequently attacked meter from those in the perfect information optimal vector has been selected in 58% of imperfect attacks while the most-frequently attacked meter from those not in the perfect information optimal vector has only been selected in 41% of the imperfect attacks. Further, as illustrated further in Fig. 5.b, 97.5% of the imperfect cyber-attacks share at least 7 common attacked asset(s) with the perfect information cyber attack while all 10 meters from Fig. 2 have been attacked in 39.5% of the sampled instances. The important take-away here is that protecting the meters that would have been attacked in the perfect information case may well be sufficient to detect and prevent with very high probability the cyber-attacks under imperfect information from physically harming the system.

Finally, Fig.6 demonstrates which transmission branches would be overloaded due to the imperfect cyber-attacks. Adopting the color-coding of Fig. 4, we show that for a large share of the samples the imperfect cyber-attack results in overloading the same branches as the perfect information attack, albeit to a smaller degree. The take-away here is that taking physical preventive/corrective measures for the possible joint outage of these branches could also be an effective strategy for managing cyber-physical risk. Notice the small frequency of imperfect cyber-attacks overloading three branches, which are suboptimal in terms of total overload magnitude.

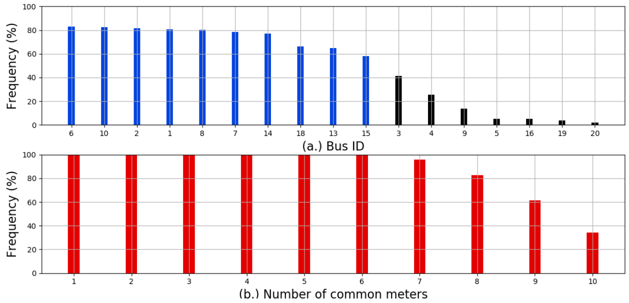

Fig. 5. Frequency of attacks (a.) per meter and (b.) sharing common meters with the perfect information attack

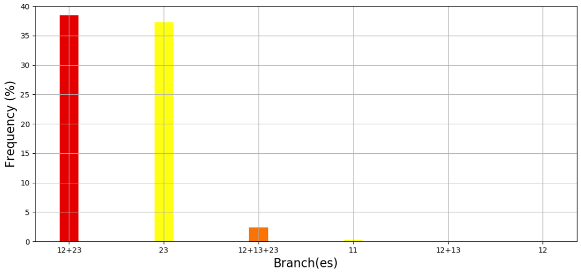

Fig. 6. Physical impact of of cyber-attacks with imperfect admittance data

D. Sensitivity analysis with respect to the admittance error range

To validate the aforementioned observations we perform a sensitivity analysis by drawing two additional samples of 10000 inaccurate grid instances while assuming that the imperfect cyber-attacker's error in admittance values is uniformly distributed in the ±5% and ±15% ranges. As anticipated, in the former case the average impact of the imperfect cyber-attacks increases to 35.6 MW (with 1428 unique attack vectors) while in the latter it reduces slightly to 26.72 MW (with 4044 unique attack vectors). It is noteworthy that in the case of reduced inaccuracy, Fig. 7.a., the percentage of so-called *perfect* attacks more-than doubles to 51.3%. This shows that (the reduced) inaccuracy has a smaller effect on the attack vector of the imperfect cyber-attacker. Conversely, in Fig. 7.b., increased inaccuracy almost halves the percentage of *perfect attacks*, with the most notable increase observed in the *partial* attack class.

In our detailed results we further find that for both cases (*i.e.*, under reduced or increased randomness) the set of meters included in the optimal perfect information attack vector from

[5] One may notice however that informational imperfections are in favor of security, as a perfectly informed attacker would be able to induce insecurity with 100% likelihood.

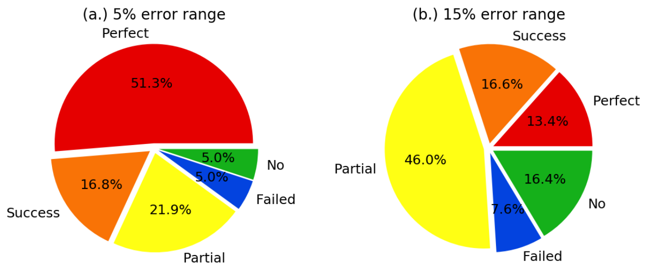

Fig. 7. Classifications for (a.) ±5% and (b.) ±15% admittance error

Fig. 2 remains the set of the 10 most frequently attacked meters, while the frequency of attacking a large subset of these meters remains as high. Specifically, for the ±5% error range 99.7% of the imperfect attacks share at least 7 meters in common with the perfect information attack and for the ±15% error range this percentage only reduces to 95.4%. These findings are well in line with the argument that protecting the meters involved in the perfect information cyber-attack is a good starting point for detecting and preventing any random imperfect cyber-attack vector. Similarly, concerning the branches that may undergo overloads in the aftermath of an imperfect cyber-attack, our sensitivity analysis detailed results qualitatively follow the representation of Fig. 6. That is, most frequently both branches that would be overloaded in the case of the perfect cyber-attack are also affected by the imperfect cyber-attacks.

E. Cyber-attacks with imperfect information on the branch capacities only

We continue the analysis by henceforth considering the case where the cyber-attacker relies on inaccurate data about the branch capacities only. We sample additionally 10000 inaccurate grids, by applying a distinct error term to the capacity value of each branch, which is again uniformly distributed in the range ±10%. With such assumptions, the average cyber-attack impact reduces to 25.31 MW while the number of unique cyber-attacks increases to 6737.

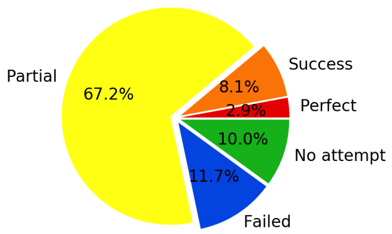

Fig. 8. Classification of cyber-attacks with imperfect capacity data

Fig. 8 presents the classification of the random cyber-attacks as per the categories introduced in section III-B. The qualitative difference with respect to imperfect admittance values is striking in comparison to Fig. 4. Indeed, for the same error range: i) the share of *perfect* attacks has collapsed, ii) the share of *success* attacks is (more than) halved, iii) the share of *partial* attacks is considerably increased, and, iv) the share of ineffective attacks is moderately increased. In other words, imperfect information on the branch capacities leads to much less effective cyber-attacks posing a smaller risk to the system cyber-physical security.

We can identify systematic reasons for this finding. Indeed, in case the cyber-attacker undervalues branch capacities, she is prone to overestimating the impact of an attack vector in firstly misleading the grid-operator to redispatch generation to avoid overloads under the load redistribution, and secondly in causing actual overloads by way of the erroneous redispatch. This explains the large shift from *perfect/success* to *partial* attacks. Also, in case the cyber-attacker overvalues branch capacities, she is prone to believing there is no potential for attacking the grid.

Concerning risk management, we once again find that the frequency of attacking a large subset of meters identified in the perfect information attack remains indicative, with 97.8% of the imperfect attacks targeting at least 6 meters from the perfect information optimal vector and 87.3% of the imperfect attacks targeting at least 7 of these meters. As should be anticipated by the dominance of the *partial* attack category, the most frequent overflow in the system now concerns a single transmission branch, Fig. 9. Notice here that the groups of affected branches (x-axis) are all in common with Fig. 6. Both these findings further showcase the relevance of these groups of cyber and physical sub-system assets for preventive and corrective cyber-physical risk management.

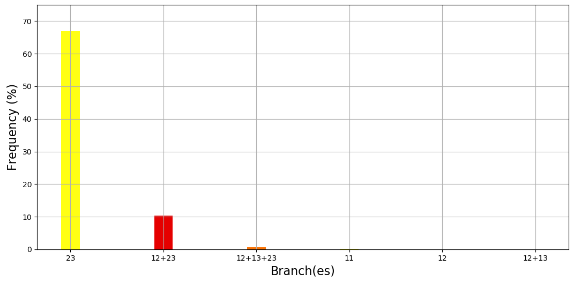

Fig. 9. Physical impact of of cyber-attacks with imperfect capacity data

F. Computational environment

Our implementation of the Monte Carlo simulation framework was developed in Julia [16] using the JuMP modeling language [17] and the PowerModels.jl framework [18]. We solved the bilevel optimization problem (1 – 26) *via* its single level equivalent reformulation, replacing the lower-level inner minimization problem (20 – 26) with its *Karush-Kuhn-Tucker* optimality conditions. We used the *big-M* approach to rewrite the disjunctive inequalities expressing the complementary slackness conditions as mixed-integer linear constraints and eventually solved all instances of the single-level MILP problem with the CPLEX [19] solver.

V. Conclusions

In this paper we have modeled the risk of cyber-physical insecurities of the electricity transmission grid, while explicitly taking into account plausible informational imperfections of a real-world cyber-attacker. We have considered the scenario of a load redistribution attack with the final aim of inducing measurable overloads (*i.e.*, beyond a minimum relative magnitude) to a number of transmission branches. For the purposes of this study, we have introduced novel constraint expressions to reflect a cyber-attacker's intention to create an overwhelming grid insecurity in the standard $\max\min$ load redistribution problem formulation. We have further performed a series of Monte Carlo simulations, modeling potential cyber-attackers with inaccurate data on branch admittances or transmission capacities, and proposed a set of metrics to synthesize the outcome of such simulations in the context of risk assessment and risk management.

From a risk *assessment* perspective, we have found in our case study that inaccurate knowledge of the grid admittance matrix is not a considerable impediment to inducing physical insecurity through the cyber sub-system. Indeed, for an increasing degree of inaccuracy on the branch admittance values between $[5-15]\%$ the frequency of cyber-attacks putting the IEEE-RTS96 benchmark in an insecure state was found in the $[90-76]\%$ range. On the other hand, relying on imperfect information on the branch transmission capacities was found to lead to a quite stronger reduction of the cyber-physical risk as it may lead a cyber-attacker to either i) launch less effective attacks when underestimating some branch capacities, and/or ii) give up the idea of attacking the system when overestimating some of them.

From a risk *management* perspective we observed in our case study that in spite of random inaccuracies the meters identified in the perfect information attack are distinctively the most frequent targets. This implies that monitoring the state of the meters that a perfectly informed attacker (*i.e.*, the "worst-case" from the view-point of the electricity grid end-users) would select could be a very effective preventive detection strategy. Moreover the set of the grid assets undergoing the physical impact of the cyber-attack was in all cases found to be relatively small, opening the possibility for efficient attack mitigation strategies on the physical sub-system.

Notice that while we relied on the specific load redistribution scenario and the specific cyber-attacker model introduced here, our analysis in principle generalizes to alternative cyber-attack instances, provided that the cyber-attacker is indeed optimizing her strategy while presuming perfect knowledge of the grid model and the grid-operator's strategy. Further work will therefore be devoted in modeling alternative cyber-attacker types, for instance an actor potentially launching any attack vector that meets some impact threshold constraints (*i.e.*, any feasible rather than an optimal solution to a bilevel optimization model). Beyond this direction, we will also pursue the question of efficiently taking cyber-physical risk management decisions under uncertainty on the realistic cyber-attacker properties.

References

[1] D. Kirschen and F. Bouffard, "Keeping the lights on and the information flowing," *IEEE Power and Energy Magazine*, vol. 7, no. 1, pp. 50–60, 2009.

[2] H. Zhang, B. Liu, and H. Wu, "Smart grid cyber-physical attack and defense: A review," *IEEE Access*, vol. 9, pp. 29 641–29 659, 2021.

[3] H. He and J. Yan, "Cyber-physical attacks and defences in the smart grid: a survey," *IET Cyber-Physical Systems: Theory & Applications*, vol. 1, pp. 13–27(14), December 2016. [Online]. Available: https://digital-library.theiet.org/content/journals/10.1049/iet-cps.2016.0019

[4] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," vol. 14, no. 1, Jun. 2011. [Online]. Available: https://doi.org/10.1145/1952982.1952995

[5] Y. Yuan, Z. Li, and K. Ren, "Modeling load redistribution attacks in power systems," *IEEE Transactions on Smart Grid*, vol. 2, no. 2, pp. 382–390, 2011.

[6] J. Liang, L. Sankar, and O. Kosut, "Vulnerability analysis and consequences of false data injection attack on power system state estimation," *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 3864–3872, 2015.

[7] J. Zhang and L. Sankar, "Physical system consequences of unobservable state-and-topology cyber-physical attacks," *IEEE Transactions on Smart Grid*, vol. 7, no. 4, 2016.

[8] M. Tian, M. Cui, Z. Dong, X. Wang, S. Yin, and L. Zhao, "Multilevel programming-based coordinated cyber physical attacks and countermeasures in smart grid," *IEEE Access*, vol. 7, pp. 9836–9847, 2019.

[9] M. A. Rahman and H. Mohsenian-Rad, "False data injection attacks with incomplete information against smart power grids," in *2012 IEEE Global Communications Conference (GLOBECOM)*, 2012, pp. 3153–3158.

[10] J. Zhang, Z. Chu, L. Sankar, and O. Kosut, "Can attackers with limited information exploit historical data to mount successful false data injection attacks on power systems?" *IEEE Transactions on Power Systems*, vol. 33, no. 5, pp. 4775–4786, 2018.

[11] A. Sanjab and W. Saad, "On bounded rationality in cyber-physical systems security: Game-theoretic analysis with application to smart grid protection," in *2016 Joint Workshop on Cyber- Physical Security and Resilience in Smart Grids (CPSR-SG)*, 2016, pp. 1–6.

[12] C. Grigg, P. Wong, P. Albrecht, R. Allan, M. Bhavaraju, R. Billinton, Q. Chen, C. Fong, S. Haddad, S. Kuruganty, W. Li, R. Mukerji, D. Patton, N. Rau, D. Reppen, A. Schneider, M. Shahidehpour, and C. Singh, "The IEEE Reliability Test System-1996. A report prepared by the Reliability Test System Task Force of the Application of Probability Methods Subcommittee," *IEEE Transactions on Power Systems*, vol. 14, no. 3, pp. 1010–1020, Aug 1999.

[13] L. Che, X. Liu, Z. Li, and Y. Wen, "False data injection attacks induced sequential outages in power systems," *IEEE Transactions on Power Systems*, vol. 34, no. 2, pp. 1513–1523, 2018.

[14] B. Colson, P. Marcotte, and G. Savard, "An overview of bilevel optimization," *Annals of operations research*, vol. 153, no. 1, pp. 235–256, 2007.

[15] Z. Chu, J. Zhang, O. Kosut, and L. Sankar, "N-1 reliability makes it difficult for false data injection attacks to cause physical consequences," *IEEE Transactions on Power Systems*, vol. 36, no. 5, pp. 3897–3906, 2021.

[16] J. Bezanson, A. Edelman, S. Karpinski, and V. Shah, "Julia: A fresh approach to numerical computing," *SIAM Review*, vol. 59, no. 1, pp. 65–98, 2017. [Online]. Available: https://doi.org/10.1137/141000671

[17] I. Dunning, J. Huchette, and M. Lubin, "JuMP: A modeling language for mathematical optimization," *SIAM Review*, vol. 59, no. 2, pp. 295–320, 2017. [Online]. Available: https://doi.org/10.1137/15M1020575

[18] C. Coffrin, R. Bent, K. Sundar, Y. Ng, and M. Lubin, "PowerModels.jl: An open-source framework for exploring power formulations," in *2018 Power Systems Computation Conference (PSCC)*, June 2018.

[19] I. I. Cplex, "V12. 1: User's manual for CPLEX," *International Business Machines Corporation*, vol. 46, no. 53, p. 157, 2009.