# City Research Online

## City St George's, University of London

**Citation:** Netkachov, O., Popov, P. T. & Salako, K. (2019). Quantitative Evaluation of the Efficacy of Defence-in-Depth in Critical Infrastructures. In: Resilience of Cyber-Physical Systems. (pp. 89-121). Berlin, Germany: Springer International Publishing. ISBN 978-3-319-95597-1 doi: 10.1007/978-3-319-95597-1_5

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

**Permanent repository link:** https://openaccess.city.ac.uk/id/eprint/21559/

**Link to published version:** https://doi.org/10.1007/978-3-319-95597-1_5

**Copyright and Reuse:** Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

# Quantitative Evaluation of the Efficacy of Defence-in-Depth in Critical Infrastructures

Oleksandr Netkachov[1], Peter Popov[1] and Kizito Salako[1]

[1] City, University of London, London, Northampton Square, EC1V 0HB, UK
{ Oleksandr.Netkachov, P.T.Popov, K.O.Salako}@city.ac.uk

**Abstract.** This chapter reports on a model-based approach to assessing cyber-risks in a *cyber-physical system* (CPS), such as power-transmission systems. We demonstrate that quantitative cyber-risk assessment, despite its inherent difficulties, is feasible. In this regard: i) we give experimental evidence (using Monte-Carlo simulation) showing that the losses from a *specific* cyber-attack type can be established accurately using an abstract model of cyber-attacks – a model constructed without taking into account the details of the specific attack used in the study; ii) we establish the benefits from deploying *defence-in-depth* (DiD) against failures and cyber-attacks for two types of attackers: a) an attacker unaware of the nature of DiD, and b) an attacker who knows in detail the DiD they face in a particular deployment, and launches attacks sufficient to defeat DiD. This study provides some insight into the benefits of combining design-diversity – to harden some of the protection devices in a CPS – with periodic "proactive recovery" of protection devices. The results are discussed in the context of making evidence-based decisions about maximising the benefits from DiD in a particular CPS.

**Keywords:** stochastic models, defence-in-depth, power transmission system, adversary model, cyber-attacks, NORDIC-32, IEC 61850

## 1 Introduction

Security of *industrial control systems* (ICS) used to control *critical infrastructure* (CI) has attracted the attention of researchers and practitioners. The evidence is overwhelming that the services offered by CI are somewhat robust with respect to single component failures of the underlying network. The reaction to multiple and cascade failures, however, is much more difficult to understand and to predict, especially when *cyber-attacks* are taken into consideration.

Dealing adequately with cyber-threats requires a credible assessment of the effectiveness of cyber-security controls deployed in a particular system. This is

particularly important if the results from the analysis are used to support *decision making*, e.g. about how to maximize the benefits from a given limited investment. Cyber-security *assessment* has matured over the last decade[1]. Yet, recommendations to deploy specific security controls are often made with *no quantification* of the benefits these are likely to bring to a *particular system*. The assessment results, especially when qualitative assessment techniques are used, are often difficult to reproduce. Decision makers struggle to justifiably answer practical questions such as "How much should I invest in improving cyber-security?", "How much better is my system after spending the available budget on additional cyber-security controls?", and "Have I done enough to secure my system?".

Probabilistic models for assessment are widely used in critical systems, for quantitative reliability assessment as well as highlighting *serious misconceptions, e.g.* the well-known controversy surrounding the quantification of the benefits from design diversity for software reliability [1], [2]. The success of these models motivated the present work – using a similar style of modelling, we develop a method for cost-benefit analysis of defence-in-depth in a CI. *Defence-in-depth* (DiD) – a multi-layered approach to defending against accidental and design faults – has been widely used in safety-critical systems for many decades. The essence of DiD is that a number of defence mechanisms, typically diverse in nature, are deployed to defend a system from threats, such as accidental/design faults or malicious activities (e.g. cyber-attacks). Respectable bodies, e.g. ICS-CERT, have recommended DiD for cyber-security of ICS [3]. While DiD has been demonstrated to bring significant *safety* benefits in safety-critical systems, its benefits with respect to cyber-threats are yet to be demonstrated convincingly. In this chapter we take some steps in this direction.

In this chapter:

- We study how the behaviour of a complex system model (of a power transmission system) is affected by the level of abstraction in modelling the effect of cyber-attacks on smart devices (i.e. those devices containing non-trivial software) deployed in a power transmission network. We compare the behaviour of the same system model using two alternative models for the effect of cyber-attacks on the smart devices: i) a *conceptual (i.e. abstract) model* of the reliability of smart devices deployed in adverse environments and ii) a more detailed model of the effects of successful, specific cyber-attacks described in our previous work [4]. Our results demonstrate how the abstract cyber-attack model *can be tuned by a suitable parameterisation*, so that the system model behaves *comparably* to how it behaves using the more detailed cyber-attack model. This observation suggests that a model-based risk assessment (or a cost-benefit analysis) can be performed, perhaps even for *unknown cyber-threats*, by using an abstract model of cyber-attacks with a suitable parameterisation.
- We apply the abstract model in studying the benefits from deploying a specific form of DiD in a power transmission network and its respective ICS. Here, DiD involves replicating some of the smart devices using *design diversity*, e.g. devices from

[1] A range of standards deal with risk assessment including cyber-attacks on industrial control systems, e.g. IEC 62443, ISO/IEC 15408, ISO 27005, etc.

different vendors are combined, together with a maintenance policy, such as "proactive recovery" [5] or "cleansing" [6].

The rest of the chapter is organized as follows: In section 2 we state the problem of quantifying the benefits from DiD against cyber-attacks in CI. In section 3 we provide a brief description of the case study and the particular models adopted for DiD. Section 4 summarizes our findings, which we discuss in section 5. Related research is highlighted in section 6 and, finally, section 7 concludes the chapter with an outline of directions for future research.

## 2 Problem statement

Investment in improving CI resilience is high on the agenda of many companies' boards. An investment decision is typically taken in the face of a large number of alternatives and uncertainties, thus requiring an evaluation and comparison of the efficacy and associated risks of employing each of these alternatives. An example investment decision, taken from a power systems context, considers whether or not the monitoring of physical network assets should be accomplished using either *Wide Area Measurements Systems* (WAMS) that employ high frequency GPS-synchronized *phasor-measurement units* (PMU), or a more traditional monitoring network comprised of low frequency *remote terminal units* (RTU).

WAMS, being newer technology than RTU-based monitoring solutions, allow operators in control centres to conduct more sophisticated and accurate calculations of a power system's state. However, the adoption of WAMS appears to be largely driven by the "obvious" superiority of the new WAMS technology over more traditional RTU-based state estimators[2]. But there are risks in adopting this new technology, not least its apparent sensitivity to cyber-threats, and such risks do not appear to have been adequately addressed. For instance, take WAMS dependence on GPS. A recognized concern, GPS signal jamming, is a critical failure-mode. Recent studies, e.g. [7, 8], demonstrate that sophisticated *man-in-the-middle* (MiM) attacks on PMU readings may remain unnoticed by WAMS state-estimation algorithms and lead to significant biases in the power system state perceived by operators in control centres. This approach to adopting technological improvements – based on "obvious" benefits and either disregarding or underestimating new risks – seems to follow a well-established pattern in industry[3].

Furthermore, while WAMS may well bring benefits to both vendors and adopters alike in the long-run, short term risks for *early adopters* may be significant;

[2] This came to our attention from private conversations with WAMS vendors.

[3] In a highly regarded book on the theory of "disruptive innovation", [9], Christensen demonstrated that the initial technological inferiority of products and services is typically temporary and is no impediment for adopting disruptive products/technologies, provided it addresses real market needs (e.g. creates new markets, reduces the cost, etc.). In the particular case, the WAMS technology may be inferior in terms of cyber-risks, e.g. GPS jamming is not a problem at all for low-frequency state-estimation, but it *is* a critical failure mode for WAMS.

quantification of these risks seems highly desirable. The decision "to invest now or not" should be based on a sound *cost-benefit analysis*. Some of the costs and benefits are clear: i) technical advantages of WAMS over traditional state-estimators are demonstrable; ii) the upfront costs are known. The costs due to failures, however, are much more difficult to estimate. They depend on the frequency of failures and on the harm these failures cause, how good the new technology is in the face of failure, the particular operational environment, and any *additional controls* used in the particular deployment. For instance, in some installations, the WAMS dependence on GPS may be compensated by deploying *atomic clocks* which allow the PMU to continue to work with accurate timestamps even if the GPS signal becomes unavailable (e.g. due to accidental failure of a GPS receiver or due to jamming). Absence of atomic clocks in a particular installation will make WAMS dependence on GPS a serious risk for this system. Similarly, if controls are in place (e.g. strong encryption) which make MiM attacks unlikely, perturbation of the PMU readings may be assumed unlikely, which in turn will justify ignoring the problems discussed in [7, 8]. Finally, if one is uncertain about the quality of the controls[4], then a more detailed study of how the quality of a particular control impacts system operation may be needed.

Given all of the foregoing, we contend that a sound risk assessment (or a cost-benefit analysis) should be done for a *specific system*, rather than solely relying on the results of pilot studies conducted elsewhere or merely adopting generic "best practices" which may ignore important deficiencies of a specific system. Consider the case when one needs to spend a fixed budget, sensibly, to improve a particular system. A rational approach to solving this problem would be exploring the space of possible system changes, i.e. consider a number of alternative ways of investing in CI resilience and ranking the alternatives according to the benefits each of these brings. It is typically too expensive for more than a few alternatives to be tried for real. But this problem can be overcome by *using high fidelity models* – one per plausible alternative – and conducting a model-based comparison. Provided the models are credible [10], one can establish the losses due to failures under comparable threat scenarios, an essential consideration for making a sound cost-benefit analysis of the planned investment.

Such an approach is feasible – we study the benefits from adopting a specific form of defence-in-depth on a non-trivial CPS such as NORDIC-32, a reference architecture of a power transmission system. Of particular interest is the effect of hardening the instrumentation/control by introducing *design diversity* in the protection devices of power lines, generators and loads. We consider investment in protection devices by replacing legacy devices with fault-tolerant, two-channel protection devices, each of which works as a 1-out-of-2 system. That is, the specific function of the device (e.g. a line protection) only fails if both channels become failed simultaneously. We assume that the channels, although functionally equivalent, are implemented differently. For example, when protection is based on different algorithms (functional diversity) or on different implementations (by different vendors) of the same algorithm. There are two important consequences of such *diversification*: the channels may be less likely to i)

[4] For instance, strong encryption may guarantee the integrity of PMU readings, but i) the use of encryption in practice is not guaranteed, and ii) encryption keys may be compromised.

fail simultaneously due to *the same design fault* (e.g. the same software bug); ii) contain *the same exploitable vulnerabilities* (than if the channels were identical). Thus, repeating the same attack on each of the channels is unlikely to compromise both. Compromising both channels is still possible, but may require different attacks be carried out either simultaneously (or in quick succession, but with a very short duration between each attack) or at different times.

# 3 The case study

We use a non-trivial case study of a power transmission network, NORDIC 32, to demonstrate our approach. The system model was developed by the FP7 EU project AFTER – the NORDIC 32 network was enhanced with an industrial distributed control system (IDCS), compliant with the international standard IEC 61850 "Communication networks and subsystems in sub-stations". A detailed description of the system model is beyond the scope of this chapter, but a short summary is provided below.

## 3.1 The cyber-physical system under study

The transmission network (Figure 1) consists of a large number of transmission lines, which connect 19 power generators and 19 loads. All of the connections of the lines, generators and links are done in 32 substations.

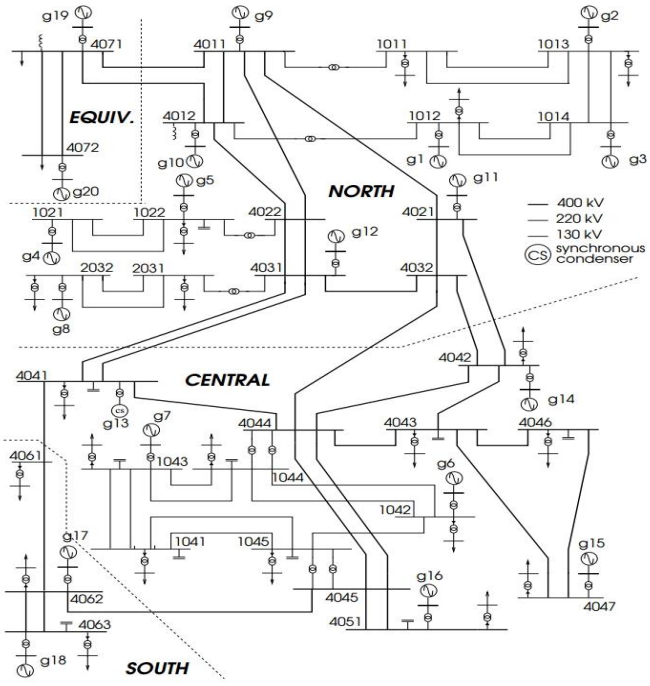


*Figure 1. NORDIC 32 power system topology. This topology is well documented in the technical literature, e.g. [11] or the more easily accessible [12].*

Each substation is arranged in a number of bays. Each bay is responsible for connecting a single element – a line, a generator or a load – to the transmission network. The substations are assumed compliant with IEC 61850. Figure 2 shows an example of one such substation (substation 4011). The other substations have similar architecture, but they may differ in their number and types of *bays*. Some substations have generators and/or loads, and all sub-stations contain Line-bays connecting transmission lines to the bus-bar of the particular substation.

Each sub-station has a *Local Area Network* (LAN), allowing local devices to communicate with one another. The LAN is protected from the rest of the world by a *firewall*. Legitimate traffic in and out the sub-station is allowed, of course.

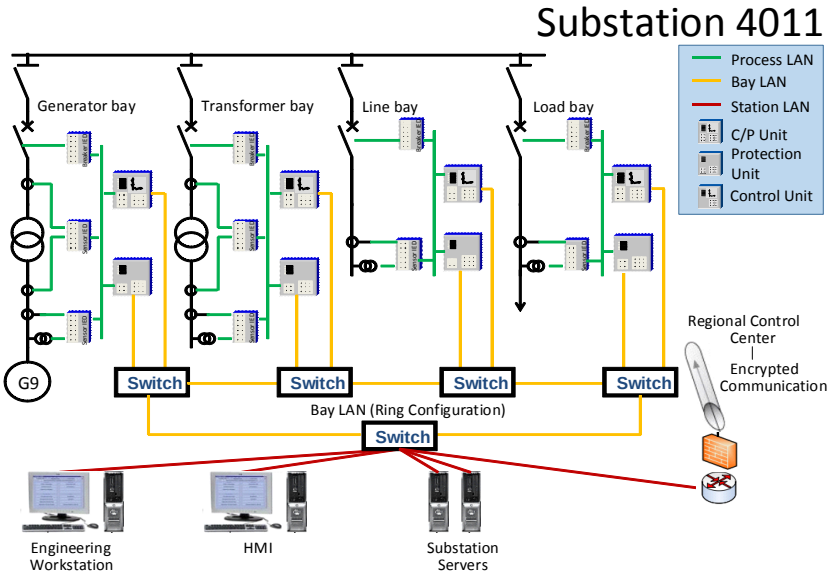


*Figure 2. Substation topology (IEC 61850 compliant).*

The substations are connected via a sophisticated communication infrastructure (Figure 3), which includes a number of control centres, communication channels and data centres.

During system operation, each protection/control function (with respect to the individual bays in substations) that is needed to maintain system integrity is either available when needed or unavailable. Availability is determined by the state (operational or not) of the equipment required for enacting the control function. For instance, shedding off a specific load to balance the power between the available generators and loads in the system, can only be achieved if the respective components – relays, communication from control centres to the respective substations, etc. - are all operational. So, in a model of the system, availability of a control function is determined by a *predicate* on the minimal cut set for the function (measurement, protection or control). Only when the predicate evaluates to "true" is the respective function available; if the predicate evaluates to "false" instead, the respective function becomes unavailable. The function will only deliver the expected outcome when it is available.

If an unavailable function is called upon to execute, it will fail to achieve the required outcome. For instance, if the function to shed some load is called upon when it is unavailable, the load will not be disconnected from the power network.

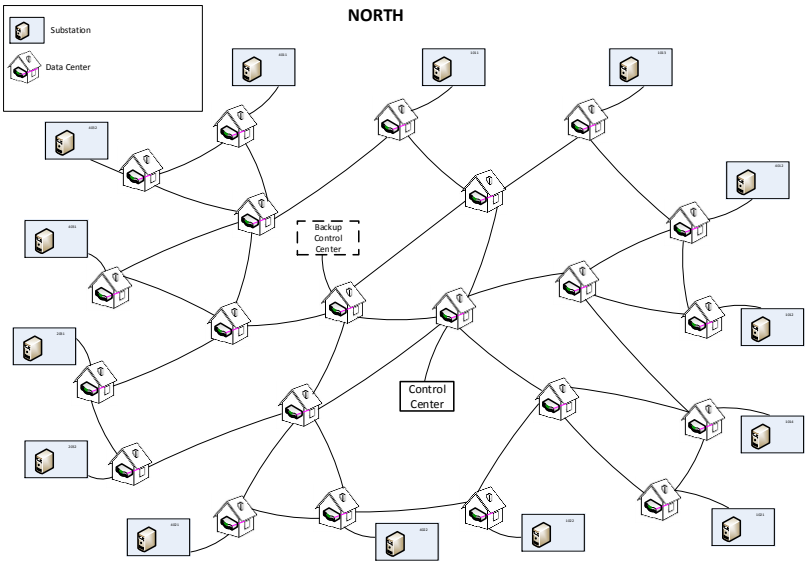


*Figure 3. Communication topology (EMS + SCADA).*

Each bay is responsible for (dis)connecting one element from the transmission network. *Protection devices* (breakers) serve to disconnect power elements from the transmission network, e.g. because of the over-loading of a line or a generator. *Control devices*, on the other hand, are used to connect or disconnect power elements from the network and are typically used by either the operators in the respective control centres or by *special purpose software* (SPS) designed to undertake some of the operators' functions automatically.

Some functions are implemented using *functionally redundant components*, others are not.

We model the behaviour of the entire system using a hybrid model: a combination of probabilistic and deterministic models to capture different aspect of system behaviour. Each element in the system model – whether it be a power element or an element for instrumentation and control – is modelled as a *stochastic state machine*. The effects of component failures on power-flow across the network is captured by a deterministic power-flow model (a DC approximation). A more detailed description of the method used to create the system model is given in [13]. Each state machine has at least two states – "OK" and "Failed". Some state machines (e.g. all power elements) may have an additional state, e.g. "Disconnected". Some other components containing software, such as those that facilitate control and instrumentation, may have the additional state "Compromised", the semantic of which we detail later.

Depending on the element type, its model, in addition to a state machine, may include specific additional *properties* needed to capture the functionality of the

component. For instance, the model of a generator will have a property defining the maximum power that the generator can produce; the model of a load will have an additional property defining the power consumed; a power line will have two properties – one that defines the line capacity and another that defines the power (current) through the line. The full list of properties for the different components that occur in the system model are beyond the scope of this chapter. However, we provide an illustration of these concepts in Appendix A. The listed properties for a given element may vary depending on the level of detail used in the model – e.g. depending on whether DC or AC power flow calculations are used to establish the new state of the power network following a disruption. The complete model used in the study can be found in [14].

The reader familiar with state-based probabilistic models may have realised that these properties extend the state space of the state machines *implicitly*. This complication is handled in our model by having a clear separation between the state captured by the states of the respective state machine and the state extension captured by the values of the various properties. The discrete part of the state evolves according to the logic built into the topologies of the state machines of the individual components. The dependencies between the state-machines are captured by different models, chief among them power-flow calculations that use the state of components across the entire power network, and predicates that determine the operational status of protection/control/measurement devices based on the state of all related components, etc. This is pragmatic approach allows us to apply in the studies well-known methods of solving Markov processes despite the use of properties, attached to some of the components.

### 3.2 Modelling protection devices

In our study we compare the behaviour of systems using non-replicated protection devices, with the behaviour of systems with (some) replicated protection devices – replication being the use of functionally equivalent but "diverse" channels in the devices.

The state machines of both a 1-channel and 2-channel protection devices are shown in Figure 4 and Figure 6, respectively. In these diagrams we refer to a "Compromised" state which will be defined in detail in section 3.3.

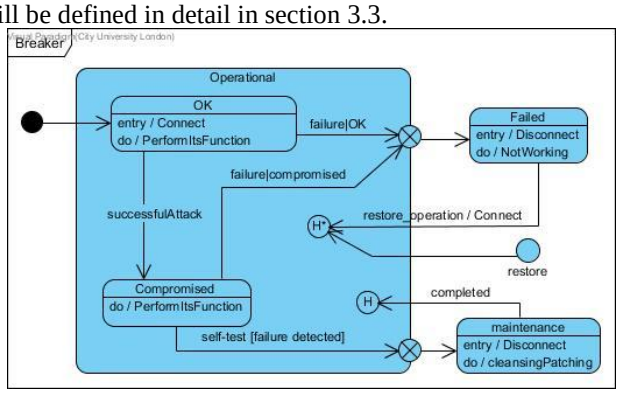


*Figure 4. A UML state machine diagram of a single channel protection device.*

A Markov chain is shown in Figure 5, which corresponds to the state-machine shown in Figure 4 under the assumption of exponentially distributed sojourn-times.

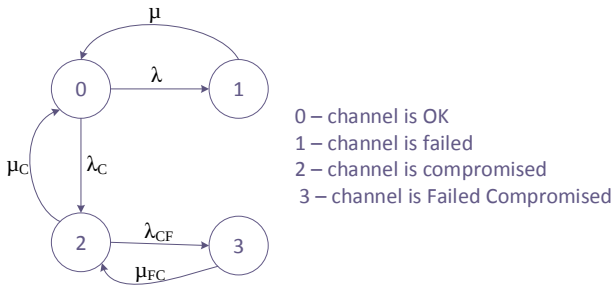


*Figure 5. A Markov chain diagram illustrating the behavior of a single channel protection device.*

The states of the Markov chain correspond to the states in the UML diagram, except for two differences: i) there are 2 failed states – "Failed" and "Failed Compromised" states. This is necessary since the state machine in Figure 4 models a non-Markov process: which state is the device restored to from a failure is dependent on the state the device was prior to failure ("OK" or "Compromised"); ii) the maintenance state is not explicitly shown[5]. The maintenance occurs while the device is in one of the failed states.

The transition rates are as follows:

$\lambda$ – rate of failure in non-compromised state.

$\mu$ - rate of repair after a failure.

$\lambda_C$ - rate of attack. This parameter is *related* to a system-wide rate of attacks (discussed later). During an attack, depending on an attacker's preferences, a particular protection device is selected (e.g. no preferences, preferences for larger assets, etc.). As a result of such a selection, the rate of attack of a specific protection device is modelled as a fraction of the rate of attacks on the system.

$\mu_C$ - rate of inspection (i.e. of returning to OK state from a compromised state). This is a system-wide parameter. In our studies we ignore the time taken to inspect different devices and restore their operation after a compromise. We model the inspections by defining the distribution of times between successive inspections (e.g. exponential distribution with mean – a day, a week, etc.). The inspection is assumed to restore *simultaneously* the normal operation of all compromised devices.

$\lambda_{CF}$ - rate of failure in compromised state.

$\mu_{FC}$ - rate of repair in Failed-Compromised state.

The behavior of a 2-channel protection device is shown in Figure 6.

The 2-channel version implements a 1-out-of-2 architecture. |That is, the 2-channel device fails only if both channels have failed. As long as at least one of the channels is operational (in either "OK" or "Compromised" state), the 2-channel breaker itself is

[5] The maintenance activity in the UML diagram is one which affects *all* of the devices across the network, and one via which transitions from the "Compromised" to the "OK" state in the Markov chain are realized. This could be modelled as a "shared activity", but would require an adequate modelling support (e.g. available in Mobius SAN v2.5).

also assumed to work correctly. The diagram uses the advanced features of UML 2.X to model the behaviour of each of the channels, including their possible failure, repair and maintenance. A channel can be restored on its own (triggered by the event "restore_operation") or by a repair of the 2-channel system (triggered by "restored" event). In the former case the respective channel is returned to the operational state it was prior to this channel failure captured by "deep history" (H*) pseudo-state[6]. The latter case leads to forking a signal "restore" to both channels, which in turn returns each of the channels to the respective deep history pseudo-state. The "Breaker" state machine further models the device maintenance and eliminates the effects of a malicious compromise of the device via either the possibility of "cleansing" [6] – e.g. restoring the device to a known clean software configuration – or by patching it. In the model we assume that maintenance is always successful, hence it returns the state machine to OK state (modelled as "shallow history", H, in the diagram, Figure 4).

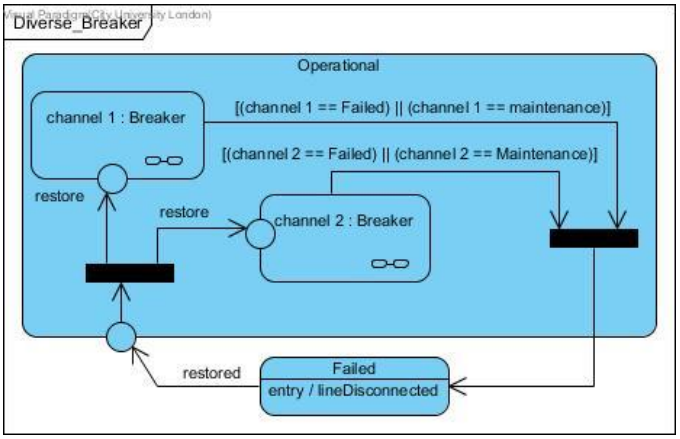


*Figure 6. UML state machine diagram depicting the behaviour of a 2-channel protection device.*

When a breaker fails, the corresponding component (line, generator, etc.) becomes *disconnected*. In either case, the failed protection device would respond to commands from the control centre (to connect or disconnect the respective line).

An external event, "successfulAttack", triggers the transition "OK" → "Compromised". A return to the OK state requires maintenance.

In a "Compromised" state the protection device (or a channel, in the case of a two channel device) continues to operate, but it may fail in circumstances, in which the device in "OK" state would not, i.e. the *failure-rate* in a "Compromised" state is higher

[6] The terms "deep history" and "shallow history" is part of the UML state-machine jargon. These refer to pseudo-states which are used with composite states (i.e. states, which consists of two or more sub-states). The pseudo state "deep history" is used to signify that when a state machine enters a composite state, it will in fact enter the sub-state of this composite state in which the state machine was prior to leaving the composite state for the last time. The "shallow history" pseudo-state, instead, will always enter the same "initial" sub-state of the composite state.

than it would be in an "OK" state. The rationale for this modelling choice is the desire to capture the effect of *advanced persistent threats* (APT), e.g. Stuxnet [15], under which the affected devices may continue to operate for some time before a failure occurs[7]. This model of how cyber-attacks affect device (channel) behaviour in a compromised state is a special case of the model developed in [16], where the interested reader may find further detail.

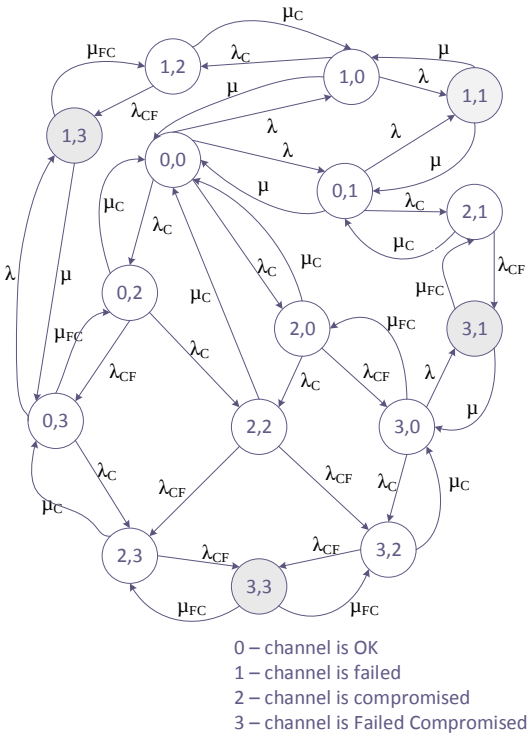


*Figure 7. A Markov chain diagram illustrating the behavior of a 2-channel protection device. The labels attached to the states indicate the state of each of the channels, e.g. "2,0" means that channel 1 is in a "Compromised" state ("2") and the second channel is in an "OK" state ("0").*

Figure 7 depicts a Markov chain which models the behaviour of the 2-channel protection device assuming exponentially distributed sojourn-times for all transitions. The state space of the chain is a Cartesian product of the space of the channels (defined in Figure 5). The transitions correspond to the rates defined for the single channel device. The transition from state "2,2" to "0,0" captures the fact that the inspection of all devices is assumed an *atomic operation*, hence the states of both channels, when found "Compromised", are changed simultaneously.

The shaded states, "1,1", "1,3", "3,1" and "3,3" are the states when the 2-channel

[7] In fact, APT introduce subtle changes in the behaviour of the compromised devices (software), which are difficult to distinguish from normal operation, hence the compromised state may remain undetected for a long time.

device itself fails – both channels are in either "Failed" or "Compromised-Failed" state. According to our assumption that the device is a 1-out-of-2 protection device, when the chain is in one of these shaded states, the 2-channel protected device (a line, a generator or a load) is *disconnected* from the power network, which in turn triggers a redistribution of the power in the power network, i.e. the power-flow calculation will be computed, which in turn may lead to more devices being disconnected if lines get overloaded.

Clearly, while in an operational state – "OK" or "Compromised" – the protection device works according to its specification: it either keeps the respective protected device connected to the power network or, when a power overload threshold is exceeded, the protection device will enact a powerline trip. This behaviour is not visible in the diagrams. Note that while the UML state-machine diagram captures normal operational behaviour – the device keeps the protected asset either connected to, or disconnected from, the power network – capturing these details with the Markov chain would be problematic. This is because transitions between operational sub-states (connected/disconnected) are triggered by changes of the entire power system – changes which are external to the protection device and typically follow deterministically after a change of the power network topology. External stimuli can be modelled with a UML state-machine while this is problematic for Markov chains.

We compare the effect on a system model's behaviour using two different models of a compromised breaker:

i) as soon as the line breaker is compromised its tripping threshold is set to a value which is 10% above the load to/from/through the protected asset (line, load or generator) at the time of the compromise. This tripping threshold can be significantly lower than the "correct" threshold, linked to the capacity of the respective asset, e.g. a line. We used this model in [4], following reports in the literature that similar attacks have indeed been observed [17]. The device failure may have no immediate consequence: it is only manifested once the state of the power network changes (e.g. as a result of accidental failure of a power component), resulting in the redistribution of electrical power flowing across the power network. If the flow through the protected asset then exceeds the incorrect tripping value set by a successful attack, the associated breaker will disconnect the asset;

ii) significantly increasing the rate of failure of a channel in a compromised state. An example would be an increase from a rate of failure *once in 10 years* in the non-compromised state to a rate of failure *once a day* in a compromised state (i.e. over *3 orders of magnitude*). Once the breaker fails (with a high failure-rate), under this model, it disconnects the respective protected element (a line, a generator or a load).

Clearly, the two models possess quite different levels of abstraction. The first one *deterministically* defines the breaker failure behaviour and requires detailed knowledge about the actions taken by an adversary. Such knowledge is only available for *known attacks*, for which extensive forensic analysis has been undertaken and their consequences established with certainty. Such knowledge, however, is not available for attacks which have *not* been seen or studied, e.g. those that use 0-day vulnerabilities.

The first model, therefore, cannot be used for analysis in the face of *unknown attacks*. The second, more abstract model of a compromised breaker, instead, merely hypothesizes that a breaker compromise leads to a higher failure intensity, without defining specifics beyond the failure mode (that should the breaker fail it disconnects the protected component). Importantly, a lack of specific knowledge about *unknown attacks* is not, by itself, an impediment for using the abstract model to study the effects of these attacks.

Via suitable parameterization of the abstract model, can one reproduce system behaviour that is *close* to the behaviour arising when using the specific model of the compromised breaker instead? If it turns out that this is indeed possible, then there is an argument for using the abstract model in risk-assessment that includes unknown cyber-attacks. Varying the parameters of the abstract model might allow one to explore a spectrum of possible losses, without a detailed knowledge about how (unknown/future) cyber-attacks may compromise the respective devices.

### 3.3 Modelling cyber-attacks

Now we briefly describe the adversary model, which captures the behaviour of the attacker. This model is derived from [4] and is extended to capture the knowledge that an adversary may have about the deployed architecture of the breaker, e.g. whether defence-in-depth in the form of replicated breakers is deployed.

For the system under study we assumed that each substation has a dedicated firewall (indicated by the "brick wall" in Figure 2), which isolates the sub-station from the rest of the world. We also assume that an intrusion detection/prevention system (IDS/IPS) monitors traffic in the sub-station's LAN. When an IDS/IPS detects illegitimate traffic, it blocks an adversary from accessing those assets located at the substation.

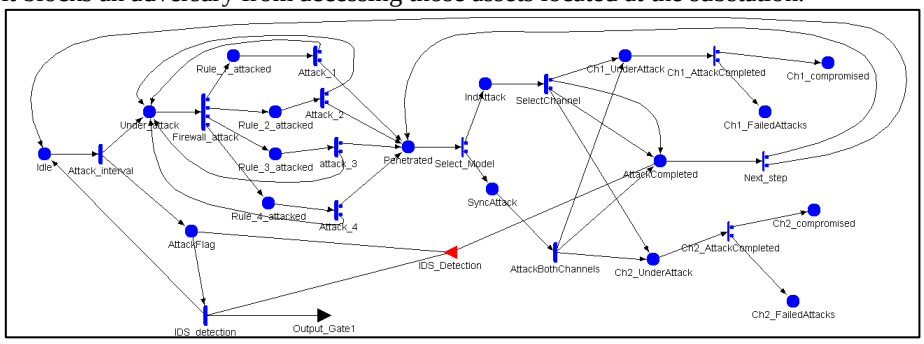


*Figure 8. Model of adversary applied to NORDIC 32.*

Our study is limited to the effect of a ***single type of attack*** on the modelled system: a cyber-attack via the firewall of a sub-station. The model is shown in Figure 8 using the *Stochastic Activity Networks* (SAN) formalism.

This model assumes that the adversary is periodically idle (represented by the SAN *place labelled* "Idle"). With some regularity, defined by the *activity* Attack_interval, the adversary launches a cyber-attack on the system by trying to penetrate the Firewall

(modelled in Figure 8 by the activity Firewall_attack) of *one* of the 32 sub-stations defined in the NORDIC-32 model.

The selection of a substation to attack[8] is driven by either a *uniform distribution*, defined over the 32 sub-stations ("Indiscriminate attacker profile") or by a *non-uniform distribution* defined in a way to capture the *preferences* of the adversary, which are discussed elsewhere [18]. In this chapter, we limit the study to an indiscriminate adversary. Under the current model we also assume that the firewalls of all sub-stations are equally easy/difficult to penetrate. This model shows the steps that follow the adversary's initial selection of a sub-station to attack:

- The adversary may target each of the firewall *configuration rules*. The decision of which rule to attack is modelled by the *activity* Firewall_attack. In Figure 8 we assume that there are 4 rules to choose between, which is just an example. The model assumes that the rules are equally likely to be chosen by an attacker – the probabilities associated with the outputs of the Firewall_attack activity are all set to 0.25.
- Once a rule is selected (modelled by the places Rule_1 – Rule_4), the adversary spends time trying to break the selected rule. This is modelled by *activities* Attack_1 – Attack_4, respectively. Her efforts may be successful or unsuccessful. In the case of a failed attempt, the adversary returns to the state Under_attack and may try another rule.
- However, in the case of a successful penetration through the firewall, the state "Penetrated" is reached, in which case the adversary now has two further options for proceeding[9]:
  - compromise the protection device of a *single line*,
  - compromise the protection devices of *all lines* in the particular substation.

  If the adversary succeeds, she leaves the substation. This choice is modelled by the *instantaneous activity* Next_step, which returns the adversary to the state "Idle".
- When the protection device (breaker) is replicated, the adversary is presented with two further choices when in a "Penetrated" state (modelled by "Select_Model" instantaneous activity):
  - *Independent attacks*: one of the channels of the breaker is selected at random and attacked. If the attack is successful, this channel enters "ChX_compromised" state; the state of the second channel remains unaffected by the attack. This models the behaviour of an adversary *unaware* of the particular form of defence-in-depth (diverse replication of the breaker) she is facing. Note, with this adversary model, the two channels of the breaker may still eventually become simultaneously compromised, e.g. as a result of 2 separate attacks on the same substation, each attacking different channels of the same breaker;
  - *Synchronised attacks*: here, the adversary tries to compromise both channels of the breaker in the same attack, using suitably devised attacks for each of the

[8] Figure 8 does not show how the adversary chooses a sub-station.

[9] The actions that an adversary can take are not modeled in detail in Figure 8. The specific logic of successful attacks – either changing the tripping threshold immediately or increasing the failure rate in the future, however, is implemented in the solver (simulator) of NORDIC-32.

channels. So each channel may be compromised as a result of a single attack. This adversary model captures an adversary with detailed knowledge of the deployed defence-in-depth. In this case the likelihood of compromising both channels of a protection device is clearly significantly higher than in the previous model with independent channel attacks.

- IDS/IPS is modelled by the *activity* IDS_detection, which is enabled if the model state is "Penetrated". This activity *competes* with the activities for the adversary selecting and attacking the breaker channels. The adversary may be detected before she completes the attacks – as soon as the *activity* IDS_detection fires, the attack is aborted and the adversary is returned to "Idle".

Finally, a channel of a protection device attacked multiple times may end up being *compromised multiple times* [16]. In this study, however, we ignore the implications of this possibility, assuming that the cumulative effect of multiple compromises of the same channel is no worse than the effect of a single successful attack; this is, admittedly, a simplifying assumption.

## 4 Results

First we compare the behaviour of the system model with two different adversary models: i) the adversary model described in [4]. Under this model, if the adversary succeeds in getting unauthorised access to the protection devices of a substation, she changes the tripping threshold from being set at a value slightly over the capacity of the protected device (line, generator and load) to a value that is merely 10% above the current flow through the protected device at the time of the successful attack; ii) the adversary model [16] discussed above. Under this model the rate of failure of a compromised protection device is set to a value of $10^3$ greater than the rate of failure of the uncompromised device (i.e. before the successful attack on it).

In both cases a measure of interest is the expected value of the supplied power, as a fraction of the maximum power that can be supplied in the model, 10,940 MW. This measure is calculated via Monte-Carlo simulation. The NORDIC–32 system is simulated as running for 10 years of operation, repeated 300 times. The 300 simulation runs are a sample, allowing us to compute a sample estimate of the expected supplied power, as well as confidence intervals for this statistic at stated confidence levels.

In all simulated cases, in addition to those measures which seek to prevent an adversary from accessing assets of a substation, a periodic activity of "proactive recovery" (or cleansing) of the protection devices is in place. This activity restores the successful operation of protection devices by eliminating any effects of compromises that have taken place. When an adversary alters tripping thresholds, cleansing restores tripping thresholds to their nominal values. For the abstract model of a compromise, cleansing restores the failure-rates to the values assumed for non-compromised states of the protection device. The cleansing procedure is assumed "perfect", i.e. its outcome

is always a success[10]. Later in the chapter we discuss the implications of relaxing this assumption.

### 4.1 High fidelity vs. abstract adversary models

In this section we summarise the results from the studies with the two adversary models. These are detailed in Table 1. The labels attached to the columns are as follows:

- μ represents the expected supplied power as a fraction of the nominal supplied power. The expectation is calculated over a number of simulation runs, *N*, typically 300. The average supplied power, P, is a random variable. For each simulation run, *i*, P takes some value $p_i$. We define $\mu$ as the expected value of P, and it is computed as: $\mu \equiv E[\mathrm{P}] = \frac{\sum_{i=1}^{N} p_i}{N}$. Values of μ close to 1 (100%) represent cases with small average losses, while large deviations of μ from 100% indicate more significant losses, e.g. those due to cyber-attacks.
- σ is the standard deviation of P.
- LB and UB are the lower and upper bound, respectively, of the 95% confidence interval for μ computed under the assumption that P is normally distributed.
- p-values are computed for the Anderson-Darling statistic, in a test for statistical normality. A value of the test statistic is computed for each sample of simulation runs (typically a sample-size of 300), and the associated p-value for the sample is the probability of observing a value for the test statistic that is no less extreme than the value computed from the sample, assuming the sample was indeed drawn from a normal distribution. This p-value should be compared with the required significance level, typically 0.05, to pass a judgement about normality – as values smaller than the significance level suggest that the hypothesis about normality should be rejected.

The top part of the table summarises the observations when the adversary model follows a specific cyber-threat closely – changing protection thresholds of the protection devices. Successful attacks of this kind have no immediate visible consequence and may manifest themselves only if/when the topology of the power network changes and the flow of power alters in such a way as to exceed the thresholds of some compromised devices and, thereby, trigger line trips. The problem may escalate over time, and unless the tripping thresholds are restored to their proper values the losses will be very significant, as the top 3 rows in the table indicate. Such large losses are clearly intolerable, and the problem with thresholds is likely to be identified and fixed. For this reason, although the studies point to a potentially serious type of attack, the fix is relatively simple.

[10] Clearly, this is a simplifying assumption, which may not hold true in practice: the cleansing procedure itself may be fallible or it may be unavailable due to an insufficient number of personnel or insufficient amount of resource required for its enactment.

*Table 1. Lost power due to attacks tampering with the tripping threshold of a protection device*

| | Case | μ | σ | LB | UB | p-value |
|---|---|---|---|---|---|---|
| Attacks change protection threshold | Daily attacks, no inspections | 0.0319 | 0.0171 | 0.03 | 0.0338 | <0.0005 |
| | Weekly attacks, no inspections | 0.2180 | 0.1143 | 0.205 | 0.2309 | <0.0005 |
| | Monthly attacks, no inspections | 0.7185 | 0.2101 | 0.695 | 0.7423 | <0.0005 |
| | Yearly attacks, no inspections | 0.9681 | 0.0552 | 0.962 | 0.9744 | <0.0005 |
| | Weekly attacks, inspect daily | 0.9832 | 0.0015 | 0.983 | 0.9833 | 0.74823 |
| | Weekly attacks, inspect weekly | 0.9800 | 0.0028 | 0.980 | 0.9803 | 0.00070 |
| | Weekly attacks, monthly inspections | 0.9692 | 0.0089 | 0.968 | 0.9702 | <0.0005 |
| | Weekly attacks, yearly inspections | 0.7653 | 0.1243 | 0.751 | 0.7794 | <0.0005 |
| Attacks reduce reliability | Weekly attacks, no inspections | 0.7913 | 0.0106 | 0.790 | 0.7925 | 0.031739 |
| | Weekly attacks, monthly inspections | 0.9772 | 0.0021 | 0.977 | 0.9774 | 0.90484 |
| | Weekly attacks, yearly inspections | 0.9226 | 0.0185 | 0.920 | 0.9249 | <0.0005 |
| No attacks | Baseline | 0.9845 | 0.0012 | 0.984 | 0.9846 | 0.53410 |

Looking at those rows of the table which summarise the effect of inspections, one sees that the frequency of inspections affects losses, which is not surprising. Monthly inspections leave the losses within the 2% (in comparison with the base line) – 0.9692 vs. 0.9841 for the average supplied power.

Let us now compare the model results with the two models of attacks – using a detailed model of stealth attacks vs the more abstract model of the effect attacks would have on compromised protection devices. The two highlighted rows of the table show losses calculated with the two models. It is striking how close the average losses are: 0.9692 for the stealth model vs. 0.9772 for the abstract model of the compromised protector. Although the difference between these averages is statistically significant[11] the absolute difference is negligible – less than 1%! This observation suggests that despite conceptual differences between these models of how attacks compromise protection devices, the average losses from stealth attacks for the particular case (of weekly attacks and monthly inspections) for this particular system, NORDIC-32, can be *estimated quite accurately* using a model which operates at a much higher level of abstraction. And more importantly, the abstract model does not rely on detailed knowledge of the mechanisms of how the stealth attacks might alter the behaviour of the protection devices, which makes the abstract models potentially very attractive for assessing the risk from future, unknown attacks.

Perhaps it is noteworthy that these results were "*easily"* obtained from an initial *informal exploration* of the abstract model's parameter space. We ran a short campaign with an order of magnitude increase of the failure rate as a result of a compromise. The effects on the system model were negligible. We then tried an increase of 3 orders of

[11] We do not present the results from testing statistically the hypothesis that the means of the two samples are the same, but did conduct this test and the null hypothesis was strongly rejected.

magnitude and this choice of parameterisation for the abstract model produced the agreement between the models we report here.

In practice, the parameterisation of the abstract model is likely to be done more systematically. Here, we list a number of options worth considering:

- one may carry out *systematic sensitivity analysis* exploring how failure-rate[12] increases affect system behaviour. In case specific models of past cyber-attacks are available, one could try to identify a range of parameters for the abstract model, for which the system model behaves comparably to how it behaves when the more detailed models of compromise are used, e.g. repeat, for a whole slew of known attacks, a study similar to the one we reported above. Selecting the abstract model's parameter values from within this range might give some indication about the system's preparedness against both known (which is usually where security assessment stops!) and unknown cyber-threats which happen to have consequences that are captured accurately enough by the abstract model parameterised from the range identified in the sensitivity study.
- Clearly with the abstract model of consequences, the failure-rate may increase to infinity, which would result in an instantaneous failure of the compromised device. Instead of using the failure rate increase (i.e. a parameter, relative to the rate of failure of the non-compromised device), one may parameterise the abstract model using an *absolute failure-rate*. With this, a sensitivity analysis may still be employed to determine a useful range of values: from instantaneous failure to a rate which corresponds to mean-time-to-failures of a few units of meaningful time, e.g. seconds, minutes or hours, depending on the specific context.

### 4.2 Quantification of Defence-in-depth using the abstract model

In this section we look at the effect of applying defence-in-depth (DiD) in the form of 2- channel protection devices deployed at certain points across the network, instead of 1-channel protection devices. The options that we considered are: applying DiD to devices protecting lines only, generators only, loads only or all power elements (lines, and generators, and loads)[13].

In all of the cases with 2-channel protection devices we study the behaviour of the system model subjected to different attacks on protection devices:

- *Independent attacks*: each time an adversary succeeds in gaining access to a 2-

[12] Clearly, by referring to the rate of failure, we implicitly envisage an exponential distribution of the time to failure, which is often used as it reduces the problem of parameterisation to a single value. Should there be a reason ruling out the use of exponential distribution, the abstract model parameterisation will become more complex. It will involve a selection of a suitable family of probability distributions and applying sensitivity analysis to their respective parameters.

[13] Clearly, limiting the total number of 2-channel protection devices, and trying to identify the optimal places to deploy these resources in the system is yet another example of a worthwhile study.

channel protection device, she compromises *only one of the channels*, selected at random. Under this mode of attack, compromising both channels is still possible, but would require at least two separate successful attacks on each of the channels of the same device, respectively.

- *Synchronised attacks*: every successful attack on a protection device results in both channels being compromised (i.e. simultaneously by the same attack).

As stated above, we ignore the effects of the second, third, etc., successful attacks on the same channel, a simplifying assumption made for convenience (to reduce the number of modelling parameters). Under rather broad conditions, reducing the periods between proactive recoveries will reduce the probability of multiple compromises of the same protection channel, which provides some justification for the adopted simplification. Clearly, a sufficiently frequent "proactive recovery" reduces the probability of a protective channel being compromised more than once to a negligibly small number, justifying ignoring the effects of multiple attacks on the same protection device.

The results from our studies are reported in Table 2 and grouped according to the *mode of attack* – independent or synchronised, whether inspections (i.e. cleansing) are applied or not and, if applied, the rate of the inspections. At the bottom of the table, the simulation results are presented for a system with single channel protection devices. This last case is included to demonstrate some of the benefits from DiD.

*Table 2. Defense-in-depth: Independent vs. synchronized attacks on protection devices.*

| System Model | | | μ | σ | LB | UB | p-value |
|---|---|---|---|---|---|---|---|
| **Independent attacks** | Baseline (attacks disabled) | | 0.9845 | 0.0012 | 0.984 | 0.985 | 0.534 |
| | No Inspections | Weekly attacks (all) | 0.9443 | 0.0079 | 0.943 | 0.945 | 0.481 |
| | | Weekly attacks (generators) | 0.9529 | 0.0025 | 0.953 | 0.953 | <0.005 |
| | | weekly attacks (lines) | 0.9577 | 0.0046 | 0.957 | 0.958 | 0.3977 |
| | | weekly attacks (loads) | 0.9629 | 0.0013 | 0.963 | 0.963 | 0.4911 |
| | Monthly Inspections | weekly attacks (all) | 0.9843 | 0.0012 | 0.984 | 0.984 | 0.145 |
| | | weekly attacks (generators) | 0.9837 | 0.0013 | 0.984 | 0.984 | <0.005 |
| | | weekly attacks (lines) | 0.9843 | 0.0011 | 0.984 | 0.984 | 0.4103 |
| | | weekly attacks (loads) | 0.9838 | 0.0012 | 0.984 | 0.984 | 0.008 |
| | Yearly Inspections | weekly attacks (all) | 0.9801 | 0.0036 | 0.98 | 0.980 | <0.005 |
| | | weekly attacks (generators) | 0.9702 | 0.0043 | 0.97 | 0.971 | 0.117 |
| | | weekly attacks (lines) | 0.9816 | 0.0023 | 0.981 | 0.982 | <0.005 |
| | | weekly attacks (loads) | 0.9752 | 0.0027 | 0.975 | 0.975 | 0.003 |
| **Synchronised attacks** | Baseline (attacks disabled) | | 0.9845 | 0.0012 | 0.984 | 0.985 | 0.416 |
| | No Inspections | Weekly attacks (all) | 0.8930 | 0.0057 | 0.892 | 0.894 | 0.418 |
| | | Weekly attacks (generators) | 0.9505 | 0.0016 | 0.950 | 0.951 | 0.187 |
| | | weekly attacks (lines) | 0.9264 | 0.0037 | 0.926 | 0.927 | 0.884 |
| | | weekly attacks (loads) | 0.9609 | 0.0011 | 0.961 | 0.961 | 0.462 |
| | Mon thly | weekly attacks (all) | 0.9810 | 0.0014 | 0.981 | 0.981 | 0.518 |
| | | weekly attacks (generators) | 0.9774 | 0.0016 | 0.977 | 0.978 | 0.278 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | weekly attacks (lines) | 0.9825 | 0.0013 | 0.982 | 0.983 | 0.718 |
| | | weekly attacks (loads) | 0.9798 | 0.0012 | 0.980 | 0.98 | 0.056 |
| | Yearly Inspections | weekly attacks (all) | 0.9560 | 0.0089 | 0.955 | 0.957 | <0.005 |
| | | weekly attacks (generators) | 0.9588 | 0.0036 | 0.958 | 0.959 | 0.980 |
| | | weekly attacks (lines) | 0.9667 | 0.0063 | 0.966 | 0.967 | <0.005 |
| | | weekly attacks (loads) | 0.9672 | 0.0018 | 0.967 | 0.967 | 0.929 |
| 1-channel protection device | (All) weekly attacks, no inspections | | 0.7912 | 0.0106 | 0.79 | 0.792 | 0.032 |
| | (All) weekly attacks, monthly inspections | | 0.9772 | 0.0021 | 0.977 | 0.977 | 0.905 |
| | (All) weekly attacks, yearly inspections | | 0.9226 | 0.0185 | 0.920 | 0.925 | <0.005 |
| Baseline (1-channel protection device) | | | 0.9845 | 0.0012 | 0.984 | 0.985 | 0.302 |

Comparing the three rows labelled "Baseline" clearly indicates that the cases are statistically indistinguishable from the point of view of the selected measure of interest (supplied power): the collected measures are practically identical. Statistical tests of whether the samples from the simulation runs, collected for all 3 cases, come from the *same distribution*, provided us with no evidence to suggest that the hypothesis should be rejected. This observation is somewhat surprising, as it suggests that using replicated protection devices brings *no benefits* for the modelled system, provided the system operates in a *trusted environment* without attacks. The reason might be that the rate of failure of the protection devices is very low (MTTF ~ 10 years), which makes redundancy unlikely to improve a device's reliability in the face of accidental failure.

The rows at the bottom of Table 2 (labelled 1-channel device) provide measures from the attack and inspection rates used to study DiD: weekly attacks and no/monthly and yearly inspections, respectively. A comparison of 1-channel and of 2-channel protection devices indicate clear benefits from employing replication in an untrusted environment. The benefits are more clearly pronounced for independent attacks.

Now let us compare the model behaviour with, and without, DiD, and under different attack modes: independent and synchronised attacks.

- No inspections. Without inspections, the losses under independent attacks are clearly smaller than the losses under synchronised attacks: the expected value of supplied power under independent attacks is closer to the values recorded for the Baseline studies than the losses from synchronised attacks.
  Stratification – attacks are applied to all protection devices vs. to generators only, lines only and loads only – provides additional insight as to where DiD would bring the most serious benefits. Under the independent attacks model, losses from attacks on generators are greater than the losses from attacks on the lines or on loads. Under synchronised attacks, however, the pattern is different. With no inspections the largest losses from synchronised attacks are recorded for attacks on the lines, while the losses from attacks on generators are lower than from attacks on both lines and loads.
  - Inspections (either monthly or yearly). Adding inspections changes the ordering between the cases quite subtly.
    - For independent attacks, even yearly inspections make the system

comparable to the Baseline case: the additional power lost due to weekly cyber-attacks is only a small fraction of a percent. Clearly, the combination of replication in protection, together with the favourable attack regime (one channel at a time), is sufficient for the effects of cyber-attacks to be *almost entirely compensated*; the additional losses are very small. Increasing the rate of inspections (monthly) reduces the additional losses due to cyber-attacks even further, which is not surprising.

- For synchronised attacks the fact that the two channels of a protection device can be compromised by the same attack, leads to device failure shortly (on average 7.5 hours later) after a successful attack. A device failure, in turn, leads to disconnecting the respective protected component (a generator, a line or a load) from the power network, i.e. the topology of the power network changes, and some power losses become inevitable. Yearly inspections are simply not frequent enough to mitigate the additional losses due to cyber-attacks: with yearly inspections the losses due cyber-attacks are almost 3 times greater than they are due to accidental failures (the baseline case). Our results suggest that monthly inspections can mitigate – to a large extent – the additional losses: the model behaviour with monthly inspections is very close to the Baseline case, especially for the cases when power-line protections are under attack. Intuitively, this last observation is not surprising: disconnecting some lines may be of no immediate consequence. Whether disconnecting a line will lead to losses or not depends on the topology of the power network before and after disconnecting the line. Our study also suggests, that the monthly inspections are less effective in mitigating losses from attacks on protection devices attached to generators and loads. Although intriguing, this observation is not surprising either: disconnecting a load in the power network leads to an immediate loss of power. The effect of disconnecting a generator is less obvious: in some cases the effect may be nil, e.g. if the operational generators have spare capacity sufficient to pick up the required power and the topology of the network is such that it does not get overloaded. If a large generator is disconnected[14], however, a power loss is imminent and substantial.

# 5 Discussion

Our studies demonstrate the quantitative analysis of cyber-risks in a complex industrial system. Contrary to a commonly adopted approach to cyber-risk assessment (e.g. [19] relying on “high”, “moderate”, or “low” qualitative indicators of impact), we demonstrate that the impact of cyber-attacks can be meaningfully established using a

[14] One of the generators in NORDIC – 32 provides more than 40% of the power in the network. The smallest generator – provides less than 10% of the total power.

model of the *particular cyber-physical system*. While we share the view that establishing the likelihood of various cyber-attacks is difficult and, perhaps, unknowable[15], the quantitative method of cyber-risk assessment we put forward here seems useful. Dismissing quantitative methods because of a lack of credible methods to capture likelihoods seems to miss the point. Yes, even if, somehow, the "true" likelihoods of attacks can be captured today, these are likely to become hopelessly inaccurate when the landscape of cyber-threats changes tomorrow. However, instead of giving up on quantitative risk-assessment because of this difficulty, one could opt for performing sensitivity analysis over a range of plausible likelihoods. Using such an approach could establish useful bounds for risk indices of interest (e.g. the lost power in our studies). This is much better than using questionable indices with values {high, moderate, low} calculated on a scale devoid of mathematical rigour, and that typically ignore the specific application context!

Now, arguably, there is a fundamental issue with security assessment activities that are solely based on establishing whether "best practice and engineering principles" have been followed. While there is no doubt that such assessment approaches are sensible, they do fall short in answering the question of whether the system is "secure enough". Clearly, while undertaking an assessment (certification) gives some confidence that the system is prepared against anticipated (i.e. known) attacks, such confidence can be misleading, especially if the system scores very well in the assessment (certification). The problem is the assessment provides no indication of how good the system defences are against *unknown* cyber-threats (e.g. those that exploit 0-day vulnerabilities).

One approach to tackling this problem was developed recently in [16], which we have now attempted to validate here. By using an abstract adversary model consistent with [16], we reproduce the expected power loss experienced by the NORDIC-32 power system subjected to sophisticated stealth attacks. Here, with the power system undergoing monthly maintenance inspections and being subjected to weekly attacks, each successful attack resulted in a modified tripping threshold for some protection device. The corresponding abstract adversary model does not explicitly represent these threshold changes; instead, the consequence of a successful attack is represented as an increase in the failure rate of the affected device. And yet, the expected losses, 0.9692 and 0.9772, under these very different alternative ways of capturing the effects of successful attacks, are in close agreement. Our studies highlight the potential for the behaviour of a CPS, subjected to a previously unknown sophisticated cyber-attack, to be suitably mirrored by subjecting the CPS to attacks from a properly parameterised abstract adversary model. Fully demonstrating such a substitute of "the specific" with "the abstract" will take more than our simulation studies, but we believe the work we report here is important as it indicates a useful way forward in addressing unknown cyber-attacks. Finding out how the system might be affected by unknown attacks may prompt system operators to look for additional controls to bring risk down to an acceptable level.

The final set of results – quantifying the effect of DiD – is also intriguing. With these

[15] This is a "known unknown"

we confirm the observations made in [16], but with a much larger CPS – that a model of an adversary attacking replicated assets (in this case protection devices) is significantly affected by the adversary's knowledge of the architecture of the assets. The improvements DiD bring against *independent attacks* (i.e. when a single channel of a replicated asset is attacked) are more significant than the improvements against synchronised attacks. The magnitude of the difference depends on how replication is complemented by inspections – measures to cleanse software from the effects of cyber-attacks.

The main message of the study, however, is in demonstrating the feasibility of deploying DiD *rationally*. If an operator has identified a number of plausible and affordable alternatives, say A, B, C, etc. to deploying DiD, then she doesn't need to count on "gut feelings" to choose amongst these. No; instead, she can run model-based studies with each of the available alternatives and compare the resulting improvements. Such studies, however computationally demanding, are a small price to pay in comparison with investing in a sub-optimal alternative that gives little to no improvement. The feasibility of the approach is demonstrated in this work: we identified a number of alternative deployments of DiD – all equipment protected, protection for only generators, or lines, or loads – and report on the benefits each of these candidate DiD-deployments bring.

## 6 Related research

In addition to the references given earlier, we would like to outline a number of related sources.

There have been studies applying different modelling techniques to *known* attacks. A couple of examples are [20, 21]. The first reference applies a probabilistic technique to define a model of Stuxnet, and demonstrates how model parameters can be assigned plausibly. The second example, instead, uses a non-probabilistic formalism. These authors claim that documenting the particular malware is, itself, an important contribution. Neither of the two models, however, is used by the respective authors for an analysis of open research problems. Our focus is quite different here: instead of merely constructing a model of something that has been seen, we use a model as a tool to study practical problems such as the effectiveness of DiD in different, adverse environments.

Somewhat related to the work presented in this chapter is our own previous work on modelling the effect of cyber-attacks on the reliability of an embedded device with fault-tolerant software [22]. The style of modelling there and in this chapter are conceptually similar, but the scope of the analysis is quite different. The tools to implement the work are also quite different. In [22] we developed a detailed model of a specific device – to study a specific attack on the safe-state of the device – using the *stochastic activity networks* (SAN) formalism. In this chapter, however, we use complex hybrid models of power systems which combine both probabilistic (stochastic state machines) and deterministic (e.g. power-flow models) parts. A SAN is depicted in Figure 8 merely as an aid in describing the adversary model.

The synchronized attacks that we studied in detail are *conceptually similar* to common mode/cause failure; a topic which has been studied extensively in the context of system/software safety and highly available computer systems.

There is also conceptual similarity in the proposed approach of replacing specific models of adversaries with more abstract counterparts and the popular approach to dependability analysis based on fault injection – trying to learn about true faults via injection of faults believed by the proponents of the methods to be representative. Despite the conceptual similarities – replacing "reality" with surrogates – there are significant conceptual differences. Many of the fault-injection based studies merely assume that injecting faults is a valid approach. In our work, we try to gain confidence in those model parameter values potentially related to unknown attacks by identifying those parameter values which make the abstract model suitably mimic the "real thing".

Finally, we would like to acknowledge the ADVISE formalism, a part of the popular Mobius tool (https://www.mobius.illinois.edu/). The ADVISE formalism captures, probabilistically, the motivation of an adversary, the assets of a particular system and the rewards that successful attacks will bring to the successful adversary. ADVISE operates at a high level of abstraction, which may pose some difficulties in estimating risk indices requiring detailed causal mechanisms for their computation, such as the expected loss of power which we used in our studies. Modelling synchronized attacks, which require detailed knowledge of defense-in-depth, with ADVISE is likely to pose additional difficulties, too.

## 7 Conclusions and future research

This chapter provides a number of results concerning a quantitative assessment of cyber-risks in cyber-physical systems (CPS) – one which we proposed a few years ago. We use a complex model of the power-transmission system, NORDIC-32, extended with measurement, protection and control, all in line with the recent standard for interoperable sub-stations, IEC 61850.

We report on two important advances:

- experimental evidence that, via suitable parameterisation of an abstract model, the expected losses due to a specific attack can be established fairly accurately. This result is significant as it points to an intriguing prospect of quantifying risks from unknown cyber-attacks.
- demonstrating that our model-based approach can be used to support rational, evidence-based decisions, about how to maximise the benefits from investing in defence-in-depth (DiD). We studied the effectiveness of DiD – a combination of design diversity in protection devices and proactive recovery of the channels – as a defence against two types of attackers:
    - a naïve attacker, unaware of the nature of DiD, who would select only one of the channels of the protection device to compromise at any one time, and,
    - a knowledgeable attacker. One with detailed knowledge of the DiD they face, and able to launch attacks which defeat the DiD.

This work can be extended in a number of ways. The encouraging result, that there are easily identifiable circumstances under which an abstract adversary model can be used to accurately establish losses from a fully defined attacker, needs further scrutiny. In what ways can this be harnessed to give more insight into unknown attacks? In part, this will require exploring more specific models of attacks, studying how well the abstract model can mimic these, and using sensitivity analysis to establish ranges of the abstract model's parameters that result in plausible, but as-of-yet unseen, cyber-attacks.

It is unclear at this stage whether the abstract adversary model used in this work is universally applicable, i.e. whether accurate estimates of the loss can be achieved for any attack type – we suspect not. The modelling circumstances under which such parity can be accomplished, as well as the generality of the approach, requires further investigation. In order to shed more light on the problem, we plan to look at sophisticated attacks, e.g. compromising WAMS software (mentioned earlier) or other "special purpose software" (SPS). Such cyber-attacks could lead to system operators being presented with plausible, but nevertheless incorrect, data on the state of the CPS, causing these operators to take erroneous decisions in the control room. It may well turn out that the effects of compromised SPS require a different family of abstract models. Constructing these with the same objective – getting accurate estimates of losses due to attacks on SPS – is an important direction for immediate future research.

## Acknowledgement

This work was supported by the UK EPSRC CEDRICS project, part of the UK Research Institute of Trustworthy Industrial Control Systems (RITICS), by the UK GCHQ and by the AQUAS project funded in part by the EU ECSEL – JU Programme (project ID 737475).

## References – 1 page

1. Eckhardt, D.E. and L.D. Lee, *A theoretical basis for the analysis of multiversion software subject to coincident errors.* IEEE Transactions on Software Engineering, 1985. **SE-11**(12): p. 1511-1517.
2. Popov, P. and B. Littlewood. *The Effect of Testing on Reliability of Fault-Tolerant Software*. in *Dependable Systems and Networks (DSN'04)*. 2004. Florence, Italy: IEEE Computer Society Press.
3. DHS, I.-C. *Recommended Practice: Improving Industrial Control System Cybersecurity with Defense-in-Depth Strategies*. 2016. 58. Available from: https://ics-cert.us-cert.gov/sites/default/files/recommended_practices/NCCIC_ICS-CERT_Defense_in_Depth_2016_S508C.pdf.
4. Netkachov O., P.P., Salako K. , *Model-Based Evaluation of the Resilience of Critical Infrastructures Under Cyber Attacks.* , in *Critical Information Infrastructures Security (CRITIS 2014)* E.G. Panayiotou C., Kyriakides E., Polycarpou M. , Editor. 2016, Springer: Limasol, Cyprus. p. 231-243.

5. Sousa, P., et al., *Highly Available Intrusion-Tolerant Services with Proactive-Reactive Recovery.* IEEE Transactions on Parallel and Distributed Systems, 2010. **21**(4): p. 452-465.
6. Arsenault, D., A. Sood, and Y. Huang. *Secure, resilient computing clusters: Self-cleansing intrusion tolerance with hardware enforced security (SCIT/HES).* in *2nd International Conference on Availability, Reliability and Security*. 2007. Los Alamitos, CA: IEEE Computer Society Press.
7. Teixeira, A., et al., *A Cyber Security Study of a SCADA Energy Management System: Stealthy Deception Attacks on the State Estimator*.* IFAC Proceedings Volumes, 2011. **44**(1): p. 11271-11277.
8. Liu, Y., 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*. 2009, ACM: Chicago, Illinois, USA. p. 21-32.
9. Christensen, C.M., *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*. 1997: Harvard Business School Press.
10. Netkachova, K., et al., *Using Structured Assurance Case Approach to Analyse Security and Reliability of Critical Infrastructures*, in *SAFECOMP 2015 : ASSURE Workshop*. 2015, Springer: Delft, Netherlands. p. 345-354.
11. Stubbe, C.M., *Long term dynamics, phase II*. 1995, CIGRE TF 38.02.08.
12. Peppas, D., *Development and Analysis of Nordic32 Power System Model in PowerFactory* in *School of Electrical Engineering , Electric Power Systems* 2008, Royal Institute of Technology: Stockholm, Sweeden. p. 77.
13. Bloomfield, R.E., et al., *Preliminary interdependency analysis: An approach to support critical-infrastructure risk-assessment.* Reliability Engineering & System Safety, 2017. **167**: p. 198-217.
14. Netkachov, O. *HPS: High Performance Simulation Engine of cyber-physical systems*. 2018; Available from: http://openaccess.city.ac.uk/19330/.
15. Falliere, N., L. O Murchu, and E. Chien *W32.Stuxnet Dossier*. 2011. 69. Available from: http://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/w32_stuxnet_dossier.pdf.
16. Popov, P., *Models of reliability of fault-tolerant software under cyber-attacks* in *The 28th IEEE International Symposium on Software Reliability Engineering (ISSRE'2017)*. 2017, IEEE: Toulouse, France. p. to appear.
17. Zetter, K., *Countdown to Zero Day: Stuxnet and the Launch of the World's First Digital Weapon*. 2016: Broadway Books.
18. Netkachov, A., Popov, P., Salako, K. *Quantification of the Impact of Cyber Attack in Critical Infrastructures*. in *1st International Workshop on Reliability and Security Aspects for Critical Infrastructure Protection (ReSA4CI 2014)*. 2014 Florence, Italy (co-located with SAFECOMP 2014): Springer International Publishing.
19. ISA, *ISA-62443-3-2, Security for industrial automation and control systems: Security Risk Assessment, System Partitioning and Security Levels*. 2017, International Association of Automation (ISA). p. 38.
20. Kriaa, S., M. Bouissou, and L. Pietre-Cambacedes, *Modeling the Stuxnet attack with BDMP: Towards more formal risk assessments*, in *7th International Conference on*

*Risks and Security of Internet and Systems (CRiSIS)* F. Martinelli, et al., Editors. 2012, IEEE: Cork, Ireland. p. 8.

21. Maynard, P., K. McLaughlin, and S. Sezer. *Modelling Duqu 2.0 Malware using Attack Trees with Sequential Conjunction*. in *2nd International Conference on Information Systems Security and Privacy*. 2016. Rome, Italy: SciTePress.
22. Popov, P.T., *Stochastic Modeling of Safety and Security of the e-Motor, an ASIL-D Device.*, in *34th International Conference on Computer Safety, Reliability, and Security (SAFECOMP 2015)*, Koornneef F. and v.G. C., Editors. 2015, Springer: Delft University of Technology, Netherlands. p. 385-399.

## Appendix A: Model of Power Line

```
{
  "name": "Link",
  "type": "state-machine",
  "comment": "Represents physical lines between substations. ",
  "properties": {
    "from": {
      "type": "Lookup",
      "required": true,
      "properties": {
        "list": "machines",
        "filter": "name === 'Substation'",
        "value": "name"
      }
    },
    "to": {
      "type": "Lookup",
      "required": true,
      "properties": {
        "list": "machines",
        "filter": "name === 'Substation'",
        "value": "name"
      }
    },
    "kV": {
      "type": "String",
      "required": true
    },
    "x": {
      "type": "Number",
      "required": true
    },
    "max": {
      "type": "Number",
      "required": true
    },
    "overloaded": {
      "type": "Boolean",
      "required": true
    },
    "connected": {
      "type": "Boolean",
      "required": true
    },
```

```
      "failure": {
        "type": "Activation",
        "required": true
      },
      "recovery": {
        "type": "Activation",
        "required": true
      },
      "length": {
        "type": "Number",
        "required": true
      }
    },
    "structure": {
      "states": [
        "ok",
        "fail"
      ],
      "initial": "ok",
      "transitions": {
        "ok": {
          "fail": [
            {
              "type": "property",
              "property": "failure"
            }
          ]
        },
        "fail": {
          "ok": [
            {
              "type": "property",
              "property": "recovery"
            }
          ]
        }
      }
    }
}
```

This code fragment provides the definition of a Power Line and includes the respective state machine and a set of properties defined for the line.

## Appendix B: a detailed description of attacks on a breaker

```
{
  "name": "Breaker Component",
  "type": "state-machine",
  "structure": {
    "states": [
      "ok",
      "fail",
      "compromised-ok",
      "compromised-fail"
    ],
    "initial": "ok",
```

```
12.    "transitions": {
13.      "ok": {
14.        "fail": [
15.          {
16.            "type": "probabilistic",
17.            "distribution": "exponential",
18.            "parameter": 0.1,
19.            "comment": "once in 10 years"
20.          }
21.        ]
22.      },
23.      "fail": {
24.        "ok": [
25.          {
26.            "type": "deterministic",
27.            "parameter": 0.00086,
28.            "comment": "7.5h"
29.          }
30.        ]
31.      },
32.      "compromised-ok": {
33.        "compromised-fail": [
34.          {
35.            "type": "probabilistic",
36.            "distribution": "exponential",
37.            "parameter": 365,
38.            "comment": "daily"
39.          }
40.        ]
41.      },
42.      "compromised-fail": {
43.        "compromised-ok": [
44.          {
45.            "type": "deterministic",
46.            "parameter": 0.00086,
47.            "comment": "7.5h"
48.          }
49.        ]
50.      }
51.    }
52.}
```

The code fragment (in json notation) defines a state machine, which captures the adversary behaviour. The state machine definition starts in line 4, from which its structure is defined: i) the *states* (“`"ok"`, `"fail"`, `"compromised-ok"` and `"compromised-fail"`), `"ok"` is defined as the initial state, and ii) the *transitions* between the states, which define the source and destination state for each of the transitions, together with a *distribution* of the transition duration: distribution type and the parameters, required by the respective distribution type. Most of the transitions in

this example are assumed exponentially distributed: this distribution requires a single parameter. The recovery from a failure (with or without a compromise) is deterministic: a fixed duration of 7.5 hours, a somewhat arbitrary figure. Apart from these two options – exponentially distributed and deterministic – a number of alternatives distributions for the transition durations are available to a modeller to choose from.