Reconciling Malicious and Accidental Risk in Cyber Security

Wolter Pieters[1,2]*, Zofia Lukszo[3], Dina Hadžiosmanović[1], and Jan van den Berg[1]
[1]TU Delft; Technology, Policy and Management; ICT; Delft, The Netherlands
{w.pieters, d.hadziosmanovic, j.vandenberg}@tudelft.nl
[2]University of Twente; EEMCS; Services, Cybersecurity and Safety; Enschede, The Netherlands
[3]TU Delft; Technology, Policy and Management; Energy & Industry; Delft, The Netherlands
z.lukszo@tudelft.nl

Abstract

Consider the question whether a cyber security investment is cost-effective. The result will depend on the expected frequency of attacks. Contrary to what is referred to as threat event frequencies or hazard rates in safety risk management, frequencies of targeted attacks are not independent from system design, due to the strategic behaviour of attackers. Although there are risk assessment methods that deal with strategic attackers, these do not provide expected frequencies as outputs, making it impossible to integrate those in existing (safety) risk management practices. To overcome this problem, we propose to extend the FAIR (Factor Analysis of Information Risk) framework to support malicious, targeted attacks. Our approach is based on (1) a clear separation of system vulnerability and environmental threat event frequencies, and (2) deriving threat event frequencies from attacker resources and attacker strategies rather than estimating them directly, drawing upon work in adversarial risk analysis. This approach constitutes an innovative way to quantify expected attack frequencies as a component of (information) security metrics for investment decisions.

Keywords: adversarial risk analysis, factor analysis of information risk, security metrics, threat event frequency

1 Introduction

When it comes to critical infrastructures such as water, electricity, and transport, policy makers are increasingly concerned about malicious disruptions of these systems. In particular, cyber attacks and even cyber warfare are considered important threats. However, the relation between traditional safety risk assessments for such infrastructures and the required security risk assessments often remains unclear.

Safety risk is usually estimated using a frequency-based approach. For example, storms of a certain category are expected to occur, say, once every decade. Based on such figures in combination with the expected impact of events, one can calculate risk as annual loss expectancy, and decide which countermeasures would be cost-effective. Within risk assessment for malicious threats, like cyber attacks, two approaches are prevalent. Either the safety approach is transferred, assuming externally determined event frequencies (baseline rate of attack, e.g. [17]), or a game-theoretic analysis is performed of strategic attacker and defender moves (e.g. [4]). Although the first approach may work for undirected threats such as viruses, it does not account for the strategic nature of targeted attacks, in which attackers "tailor their attack strategy with the aim of damaging the physical system under control" [6]. The second approach does this, but the results of game-theoretic analyses do not provide *expected frequencies* as output. The

Journal of Internet Services and Information Security (JISIS), volume: 4, number: 2, pp. 4-26

*Corresponding author: TU Delft, Faculty of Technology, Policy and Management, Section ICT, Building 31, Jaffalaan 5, 2628 BX Delft, P.O. Box 5015, 2600 GA Delft, The Netherlands, Tel: +31-(0)15-27-88989, Web: `http://homepage.tudelft.nl/e7x9k` results are therefore not compatible with a frequentist or annual loss expectancy approach to risk, and cannot be easily integrated within the context of traditional (safety) risk assessments.

In this paper, we aim at reconciling safety and security risk, by providing frequency estimations for malicious, targeted threats, in order to support security investment decisions. In particular, we ask the question how can we represent the effects of attacker strategies in such an analysis, being the major difference from the safety context.

To define our concepts, we use the risk definitions provided by The Open Group [24], based on the Factor Analysis of Information Risk (FAIR) taxonomy. They assume the existence of "threat agents" in the environment of a system that may cause damage to system assets. The threat agents cause threat events, with a certain threat event frequency. These threat events may or may not lead to loss events, with associated loss event frequency, depending on the vulnerability of the affected component. For example, an attacker executes an attack scenario (threat event), which may or may not lead to failure of one or more power plants (loss event). This model can be applied to both accidental and malicious threats. However, whereas accidental threats occur randomly, the occurrence of malicious threats is based on attacker decisions.

We extend the FAIR framework to enable integration of safety and security risk assessments. To this end, we combine adversarial reasoning with a frequency-based output of the analysis. We have developed our model in the context of the SESAME project[1], which provides decision support for security investments in electricity networks. In the same project, models of electricity networks, their components, associated vulnerability levels, and consequences of component failure are being developed as well, in order to calculate the expected damage. The present paper outlines the essential insights of the reconciliation approach, and highlights open problems.

As we are primarily interested in estimating threat event frequencies in order to support assessments of cost-effectiveness of countermeasures (return on security investment), our primary claim is that the proposed measures are useful in this context, assuming the current status quo. Given the many developments that may change the battlefield at any time, we cannot provide accurate predictions of future events, although the metrics may turn out to be useful as an ingredient for this purpose as well.

In Section 2, we discuss preliminaries and related work. In Section 3, we provide the definitions of our concepts, based on the FAIR framework [24]. In Section 4, we define the relation between threat event frequency, threat capability, vulnerability, and loss event frequency for safety contexts. In Section 5, we discuss how to use this model to estimate relevant variables in a security context. In Section 6, we illustrate the framework by simulated examples. In Section 7, we evaluate the framework against desirable properties for the metrics. We end with open questions in Section 8 and conclusions in Section 9.

2 Preliminaries and related work

2.1 Security investment

Both the frequency and the impact of events are essential for providing decision support on investments in security measures for critical infrastructures [12]. In particular, to support decisions on security investments, we need to know how much reduction of risk can be achieved by implementing countermeasures. Only if we can achieve this can we do a meaningful cost-benefit analysis. Thus, we would need measures of risk both without and with a specific countermeasure implemented. In such a context, qualitative assessments in terms of Low, Medium, High and similar scales are ineffective, and we would need quantitative measures of both event frequencies and event impacts. Frequency measures are available in safety sciences, e.g. in terms of hazard rates: storms of a certain magnitude are expected to occur once every

[1] www.sesame-project.eu

century, on average. Similarly, accidental threats such as component failure can also be assessed in terms of failure rates.

However, the security domain, dealing with malicious attacks or misuse, creates particular problems that do not occur in safety. Contrary to what is referred to as threat event frequencies or hazard rates in safety risk management, attack frequencies are dependent on attacker behaviour. The frequency of storms is independent of the resilience of power grids, but (cyber) terrorists will estimate their costs and chance of success based on what they know about the system, and then decide how to invest their resources. The frequency of threats is thus dependent on their expected impact. On the one hand, the effect of countermeasures on risk may therefore be *more* than proportional with respect to improved security, as attackers may decide to stop their activities if the security of the systems is increased. On the other hand, increased security may also merely lead attackers to target different components / different attack scenarios in the system, causing the effect to be *less* than proportional. Both matters complicate meaningful assessment of risk with and without a particular countermeasure. Inspiration for solutions needs to be sought in the domain of safety analysis [15, 16], but also in social sciences, in particular economics. In the end, we are interested in estimating the attack rates (here referred to in terms of threat event frequency) in the two different conditions (with and without countermeasure), in order to support investment decisions.

In the context of such assessments, data to support empirical models is often lacking. Many researchers have experienced that companies and governments are reluctant to share information on security incidents, complicating empirical estimation of threat levels. Moreover, even if these data would be available, filling in the empirical model for the situation in which a countermeasure has been implemented would be next to impossible. We want to estimate the benefits of investment *a priori*, as at the time of decision the countermeasure has not been implemented yet, so there is no data available about the attack rate in that situation. This holds in particular when considering targeted attacks (in contrast to, say, falling victim to a randomly travelling virus). Even if data would be available for the situation without investment, past results might not provide a reasonable estimate for the future, unless an otherwise similar organisation has precisely that countermeasure in place, *and* is willing to share its data. Finally, empirical data on past frequencies may not be appropriate predictors for future events, as attacker strategies adapt to changes in their environment. This means that we cannot estimate threat event frequencies based on empirical data, as proposed by [22]. In such situations, recourse to theoretical models is the only option, which is why such models are urgently needed. Our estimates are based on a model of a (rational) attacker, and a judgement of his resources.

In terms of quantifying costs, we assume the impact of attacks and the cost of countermeasures as given. However, it is important to note that costs of countermeasures are not only related to investment and operation, but also to the effect on users and business processes [2, 3]. There may even be situations in which certain attacks increase social welfare, and not taking measures could be rational for some actors [11].

2.2 Adversarial risk analysis

There is some literature on what is called Adversarial Risk Analysis [8, 9, 21, 23], providing game-theoretic approaches to security investment, where investment on both attacker and defender side is abstract and expressed in terms of money. The central question is then how an equilibrium is achieved by repeated investment by the attacker and the defender in response to each others' actions. However, the outputs of such analyses are not bound to time, in the sense of *when* or *how often*. We wish to estimate the expected frequency of attacks in the *current* situation of a concrete system, as well as in a situation where a *specific* countermeasure is implemented. Our example context concerns electricity infrastructures, where grid operators wish to know whether a particular security investment is cost-effective.

There are therefore two main differences between our framework and game-theoretic approaches. Firstly, we work in a context in which the defender first chooses the defenses, and then the attacker executes his strategy without further interference by the defender. Rather than requiring advanced game theory, such problems can be solved by what has been called "leader-follower 'minimax' analysis": a simple game in which the defender moves first, anticipating optimisation by the attacker [7]. Secondly, our framework will not output an equilibrium of optimal (monetary) investments, but rather calculate (1) the expected *frequency* of attack given a selected set of defenses (and associated optimal investment for the attacker, as well as risk induced by the attacks), and thereby (2) the *optimal set of defenses*.

2.3 Attacker models

One of the key items in security analysis is the attacker model. In the domain of security risk assessment, the attacker model is typically expressed in economic terms. In representing the adversaries, we can distinguish the following relevant parameters (derived from [23]): number of adversaries; adversaries' incentives; attacker risk (of detection / punishment); and resources required. The latter three are associated with attacker utility functions [14]. Utility functions map the results of attacks (either in terms of loot, detection/punishment, or mounting activities) to a single value scale for the attacker. As argued by [8], the risk of detection, and therefore the attacker's cost, can be assumed increasing and concave with respect to the effort (resources) spent on the attack. To simplify our initial model, we assume the risk of detection proportional to the effort. It can therefore be seen as part of the effort.

To develop a security model based on an attacker model, the relation between system properties and attacker strategies needs to be defined. The attacker will base his strategy on the (perceived) system properties, and the system security properties (such as risk) will in turn depend on the attacker behaviour. Several approaches already exist for quantifying security attributes, but these generally exclude the "attacker effort as a function of time", and only focus on the "security breaches as a function of effort" (terminology from [15]). When we want to estimate loss event frequencies, and allow the attacker to select the attack scenario, we need to include the attacker effort as a function of time. This requires an explicit model of attacker investment. Modelling attacker effort for each attack scenario as a random variable [16] is not adequate for our context, as the attacker will base this effort on the expected damage of the attack (from his perspective; cf. [7, 13]).

3 Definitions

In order to enable quantification of security properties, they first need to be defined precisely, which is a challenge in itself. Several definitions are possible, depending on standards and references chosen, and definitely also on the goal of the analysis. As we have outlined above, attackers will decide on their attacks based on system properties, and a single likelihood or frequency value for the frequency of successful attacks thus obscures fundamental dependencies. It is therefore essential to clearly separate between internal system properties, and external threat events.

We are also not interested in the probability of failure up to a certain point in time. Rather, we wish to know the *expected number* and *frequency* of failures. We need numbers of attacks and failures to estimate risk in a period in which multiple attacks may occur, not probabilities of at least one event happening in such a period. For example, if four attacks are expected to happen within the set time frame, and two of them are expected to succeed, this provides us with the means to calculate the expected losses. If we only know the likelihood of at least one attack having been successful (i.e. the cumulative failure probability), this does not provide the required information directly.

Based on these considerations, we choose the FAIR framework [24] as a basis.[2] In this taxonomy, risk-related variables are defined starting from the notions of assets and threat agents acting against these assets, potentially causing damage. A threat event occurs when a threat agent acts against an asset, and a loss event occurs when this causes damage. For example, a storm may occur at the location of a power line (threat event), and this may or may not damage the power line (loss event).

Like many other approaches, The Open Group distinguishes between likelihood and impact of events. However, they explicitly use frequencies to represent likelihood, leading to what they call Loss Event Frequency (LEF) and Probable Loss Magnitude (PLM). The former represents the expected number of loss events of a particular type per unit of time (often referred to as failure rate), and the latter represents the expected damage per loss event of that type. Risk can be seen as expected damage due to a certain type of loss event within a given time frame, and it can then be calculated as LEF · PLM.

Within LEF and PLM, The Open Group makes further distinctions. We will not discuss those of PLM here, but focus on LEF. First of all, the Loss Event Frequency can be decomposed into Threat Event Frequency (TEF) and Vulnerability (V). TEF denotes the expected frequency of occurrence of a particular threat (seen as a threat agent acting against an asset; a storm at the location of a power line), and V specifies the likelihood of the threat inflicting damage upon the asset. The value for LEF can then be calculated as TEF · V.

The Open Group defines the Vulnerability V based on Threat Capability (TC) and Control Strength (CS). In this definition, TC denotes some ability measure of the threat agent, and CS a resistance (or difficulty of passing) estimate of the control. We have discussed this relation in detail in [19].

Thus, if a threat event is expected to occur 4 times in 10 years (TEF = 0.4 y^{-1}), and one in two threat events is expected to cause loss (V = 0.5), then 2 loss events are expected to occur in 10 years (LEF = TEF · V = 0.4 y^{-1} · 0.5 = 0.2 y^{-1}). If the expected damage per threat event is € 1000 (PLM), then the risk run due to this threat amounts to € 200 per year (R = LEF · PLM = 0.2 y^{-1}· € 1000 = € 200 y^{-1}), or € 2000 in 10 years. We summarise and formalise these definitions below.

Definition 1. *The* threat event number *(denoted $H(t)$ for* hazard*) is the expected number of threat events within a specific time interval $[0..t]$. The* threat event frequency *(denoted $h(t)$) is the expected number of threat events per unit of time, i.e. the derivative of the threat event number.*

$$h(t) = \frac{dH}{dt} \tag{1}$$

The Mean Time Between Threat Events (denoted MTBT) is the inverse of the threat event frequency (MTBT = 1/h).

Definition 2. *The* vulnerability *(denoted $V(t)$) is the probability that a threat event causes a loss event.*

Definition 3. *The* loss event number *(denoted $\Lambda(t)$ for* loss*) is the expected number of loss events within a specific time interval $[0..t]$. The* loss event frequency *(denoted $\lambda(t)$) is the expected number of loss events per unit of time, i.e. the derivative of the loss event number.*

$$\lambda(t) = \frac{d\Lambda}{dt} \tag{2}$$

The Mean Time Between Failures (denoted MTBF) is the inverse of the loss event frequency (MTBF = $1/h$). The loss event frequency can be calculated from the threat event frequency and the vulnerability:

$$\lambda(t) = h(t) \cdot V(t) \tag{3}$$

[2] This taxonomy is also discussed in our previous work, see [19].

Similarly, the loss event number can be calculated from the threat event frequency and the vulnerability:

$$\Lambda(t) = \int_0^t h(t) \cdot V(t)\, dt \tag{4}$$

When threat event numbers or loss event numbers are calculated for different intervals than $[0..t]$, we use the following notation:

$$H([t_0..t_1]) = \int_{t_0}^{t_1} h(t)\, dt; \tag{5}$$

$$\Lambda([t_0..t_1]) = \int_{t_0}^{t_1} \lambda(t)\, dt \tag{6}$$

When threat and loss event frequencies are constant in time, this simplifies to:

$$H(\Delta t) = h \cdot \Delta t \tag{7}$$

$$\Lambda(\Delta t) = \lambda \cdot \Delta t \tag{8}$$

Definition 4. *The* damage *or* probable loss magnitude *caused by a loss event (denoted D) is the expected monetary cost of the event.*

In this paper, we assume the damage levels as given, although estimating damage is a non-trivial task itself. We mostly assume damage of the same loss event to be constant in time, although this restriction could be lifted when the effects of an attack depend on time, as with different loads at different times/seasons in electricity grids.

Definition 5. *The* risk *associated with a class of threat events (denoted $R(t)$) is the expected damage per unit of time.*

The risk can be calculated from the other variables:

$$R(t) = \lambda(t) \cdot D = h(t) \cdot V(t) \cdot D \tag{9}$$

Note that risk is expressed as a density function of time, i.e. it is the integral of the risk that expresses the damage to expect within a time frame. When considering a particular time interval $[t_0..t_1]$, one can calculate the *average risk* as:

$$\bar{R}([t_0..t_1]) = \frac{\int_{t_0}^{t_1} R(t)\, dt}{t_1 - t_0} = \frac{\Lambda([t_0..t_1]) \cdot D}{t_1 - t_0} \tag{10}$$

Of course, when threat event frequency, vulnerability and damage are all constant in time, the risk will also be constant in time.

Definition 6. *The* countermeasure cost *(denoted $C(t)$) is the expected cost of a countermeasure per unit of time. The* countermeasure-adjusted risk *(denoted $R'(t)$) is the expected risk when the countermeasure is added to the system. A security measure is said to be* cost-effective *if the cost of the measure per unit of time is lower than the reduction in risk (expected damage per unit of time) that it achieves.*

A security measure is cost-effective for time interval $[t_0..t_1]$ if the total cost associated with the countermeasure is less than the total risk prevented by the countermeasure over the time interval:

$$\int_{t_0}^{t_1} C(t)\, dt < \int_{t_0}^{t_1} (R(t) - R'(t))\, dt \tag{11}$$

4 Risk estimates for natural and accidental threats

The above provides a general framework for reasoning about risk in relation to threat event frequencies. However, threat events come in different types. Events like component breakdowns just occur or don't occur, depending on e.g. ageing. Other threat events have a certain threat capability, such as for example the strength of a storm in Beaufort. Generally, the frequency of natural events with high threat capability is lower than the frequency of similar events with low threat capability (fierce storms occur less frequently than mild storms). This can be expressed in distributions such as Poisson, Rayleigh, and Weibull.

4.1 Discrete events

Events with a discrete threat capability are specified on a discrete scale. For practical purposes, events with discrete threat capability can be treated as separate threats. For example, one could define a separate threat type for each hurricane class, and thereby treat them as discrete events. The threat event frequency for a particular threat may be constant, or may vary over time. For example, we may expect the number of storms to be constant over the years (on average), or we may expect an increase due to climate change. In case the threat event frequency does not change over time, it can simply be expressed as a constant number of events per unit of time. The loss event number is then calculated by factoring in the vulnerability (V), representing the probability that a threat event causes a loss event:

$$\Lambda(\Delta t) = V \cdot H(\Delta t) = V \cdot h \cdot \Delta t \tag{12}$$

In case the threat event frequency varies over time, it is expressed as a function of time by means of a density function ($h(t)$), to represent the distribution of threat events over time. This is not a probability distribution, as the concern is not the probability of failure of a component, but the expected number of occurrences within an interval. In using the density function, the integral of h represents the expected number of occurrences in the time interval (see Figure 1).

$$H([t_0..t_1]) = \int_{t_0}^{t_1} h(t)\,dt \tag{13}$$

$$\Lambda([t_0..t_1]) = V \cdot H([t_0..t_1]) = V \cdot \int_{t_0}^{t_1} h(t)\,dt \tag{14}$$

This assumes V constant over time, i.e. the probability of a threat event of a specific type causing a loss event does not change with time. If V does change with time, the formula changes into:

$$\Lambda([t_0..t_1]) = \int_{t_0}^{t_1} V(t) \cdot h(t)\,dt \tag{15}$$

Risk and effectiveness of countermeasures can then be calculated using equations 9 – 11.

In the following, V is always assumed to be constant in time (but, as we will see, typically dependent on threat capability). Extensions to this model would be possible, for example when control strength decreases over time due to ageing, thereby increasing vulnerability, or when, for example, security measures are different during the day and at night.

4.2 Discrete events with magnitude

Events with a continuous threat capability are specified on a continuous scale, for example the Richter scale for earthquakes. As there are infinitely many capability values (at least in theory), these cannot be treated as separate threats. When the frequency varies over time, the threat event frequency becomes

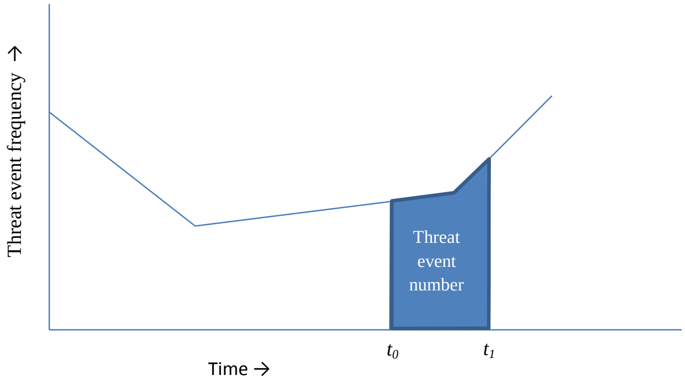

Figure 1: Calculation of the threat event number from a variable threat event frequency, for events with discrete threat capability. The coloured surface under the curve corresponds to the integral that determines the threat event number. One can observe that within the period concerned (t_0 until t_1), the threat event frequency is increasing, and therefore more threat events can be expected towards the end of the period.

a 2-parameter event density function to calculate the expected number of threat events within a certain magnitude range within a certain time interval (e.g. the expected number of earthquakes of magnitude 4-5 in 2014, see also [25]). With m for magnitude, this yields:

$$H([t_0..t_1]) = \int_{t_0}^{t_1} \int_{m_0}^{m_1} h(m,t)\,dm\,dt \tag{16}$$

As the vulnerability also depends on the magnitude, it needs to be included in the integral to calculate the loss event number:

$$\Lambda([t_0..t_1]) = \int_{t_0}^{t_1} \int_{m_0}^{m_1} V(m) \cdot h(m,t)\,dm\,dt \tag{17}$$

4.3 Continuous phenomena

For continuous phenomena, such as wind, one cannot really speak of discrete threat events. Of course a storm is considered an event, but the duration of high wind speeds is very important for the damage. Instead of threat event frequencies, we only have threat magnitudes or levels that vary over time. Therefore, the standard terminology in terms of threat event frequency, vulnerability, and loss event frequency, is not adequate, as there are no separate threat events. For continuous phenomena, different semantics are needed.

Here, we choose to keep the same notation ($\lambda(t)$, $V(m)$, $h(m,t)$), but the terms acquire different meanings. In particular, $h(m,t)$ now represents the *probability density* of the threat level, i.e. the probability that the threat level is within certain magnitude boundaries at time t. Stated differently, the integral of $h(m,t)$ with respect to m is the fraction of the time that the threat magnitude is expected to be within a specified magnitude interval. Thus, at each time t, $\int_{-\infty}^{\infty} h(m,t)\,dm = 1$.

Conversely, $V(m)$ denotes a frequency rather than a probability in this case, namely the expected failure frequency given the threat magnitude level. A high vulnerability means that, given a certain threat level, the failure frequency is high. For example, the failure rate of a power line is high under conditions of high wind speed. Apart from these changes in interpretation, the form of the calculation of the loss event number is the same as for discrete threats with continuous magnitude (Equation (17)).

In summary, h now specifies the distribution of the threat magnitudes, and V represents the failure frequency for each threat magnitude. The reversal of the roles of probabilities and frequencies as compared to the discrete case is shown in Table 1.

Table 1: The inverted relations between probability and frequency in the discrete and continuous case.

	Threat $h(m,t)$	Vulnerability $V(m)$
Discrete	frequency density	failure probability
Continuous	probability density	failure frequency

The two different interpretations of threat and vulnerability in the discrete and continuous case are the first important result of the present study. We will now turn our attention to malicious threats.

5 Risk estimates for malicious threats

The above provides a risk framework for natural threats, in which a clear distinction is made between external threat events and the vulnerability of the system to such events. Both vulnerability and impact can be reduced by means of countermeasures, and the prevented risk of such countermeasures can be evaluated. We distinguish between discrete event models (isolated threat events) and continuous phenomenon models (threat level fluctuating over time, like wind). Both can be used to calculate the expected frequency of loss events, but the associated reasoning is somewhat different.

The frequency-based model is not immediately applicable to malicious threats. For malicious threats, the threat events are no longer randomly occurring, but are based on the strategy of adversaries. The adversaries will be more interested in attacks that are likely to success and/or cause high impact. Therefore, the dependency of the threat events on the (perceived) vulnerability and impact needs to be taken into account, like in game-theoretic approaches. To combine frequentist and adversarial reasoning, the strategic attacker needs to be bound to time, in the sense that there is a time-dependent limit to what he can do. In particular, we need notions of the resources available to the attacker, which attack vectors are chosen by the attacker, the effort or resources an attacker spends on an attack, and how many attacks he can launch with the available resources.

5.1 Method

First, we outline our strategy for transforming the metrics for malicious attacks. To accommodate for the specifics of security, we need to represent that attackers will aim their efforts at the most vulnerable components, or those with the highest rewards. It is assumed here that for electricity networks, attackers are of the terrorist type, and interested in maximising damage. This provides a contrast to, for example, a banking context, where attackers are most likely interested in maximising their own gain.

Our attacker model consists of a fixed number of agents with full knowledge of the vulnerability of components and the impact of failure on the system on the side of the attacker (white-box analysis). The attackers have certain resources to spend, in the simplest case constant over time, and they incur costs by launching attacks, and possibly by risks associated with detection and punishment. We assume the terrorist type of attacks do not provide them with additional income / resources.

As attackers will aim for the most vulnerable targets, a more than proportional effect can be expected upon introduction of countermeasures: there is a reduction in vulnerability / impact, plus reduction in the number of attacks due to changed attacker decision making [8]. However, we assume here that the terrorist attacker is only interested in our system, and will switch to a different *attack scenario* upon countermeasure implementation, not a different *system*. (The latter would require modelling the environment – such as vulnerability of other systems – as well, which we leave for future work.) Consequently, there is no clear-cut point where the attacks suddenly cease (and migrate to a different system), but rather a smaller decrease in risk as the overall increase in security forces attackers to invest their resources in less vulnerable parts of the system. This will enable us to assess the effect of the countermeasure on the expected damage done to the system, even when it causes the attacker to switch strategies. So we assume the system to be analysed to be *static* (i.e. no actions on the part of the defender except an initial selection of countermeasures), and try to assess how often attacks on specific components (or even specific attacks on specific components) can be expected.

We also assume that vulnerability and impact models for the network and its components are already known to the defender, and we focus on the frequency with which specific components are expected to be attacked. For vulnerability, we assume the existence of vulnerability functions expressing the probability of success of an attack in terms of threat capability or threat magnitude, which is again a function of resources invested by the attacker. These functions will be dependent on the type of attack: a scenario involving social engineering will have a different relation between invested resources and likelihood of success than a denial-of-service attack. Defining these functions is not part of the present paper, and is covered in [19] and future work.[3] Impact of component failures on the electricity network can be calculated using load flow analysis [1], which can then be translated into monetary loss by assessing the impact of the power outages [20].

To enable a frequency-type result, we further assume that (a) the likelihood of success of an attack depends on the resources invested by the attacker (as defined by the vulnerability function),[4] and (b) the attacker has limited resources, and acquires resources over time. *The attacker, therefore, has to decide how to invest his resources over time, which will determine the threat event frequency for the different possible attacks.* We observe a similar structure here in terms of discrete and continuous models: either the attacker saves resources, and spends them at a single point in time (discrete event model, for example in case of a distributed denial-of-service (DDoS) attack), or the attacker continuously puts resources into attacking the target system (continuous phenomenon model, for example when trying to crack a password). Thus, we can re-use the interpretations from the previous section, but we need to adapt them to include attacker behaviour.

For the discrete event model, the vulnerability represents the likelihood of component failure as a consequence of the attack event (discrete case, vulnerability as probability). For the continuous phenomenon model, the vulnerability represents the expected frequency of failure as a function of invested attacker resources (continuous case, vulnerability as frequency). Intuitively, the latter expresses that components break down more frequently when under attack. Thus, similarly to the non-malicious context, vulnerability has a different interpretation in the discrete and continuous models (probability and frequency, respectively).

The utility function of our terrorist-type attacker is rather simple: maximal damage per unit of time, within available resources. The attacker will thus launch attacks (threat events) that, given the resources available to the attacker, have the highest loss event frequency times impact (= expected failures per unit of time, times consequences). This requires decisions on (a) which attack scenario to execute, and, for the discrete case only, (b) when to launch the attack. We assume that attackers are neutral with respect to

[3] In this paper, we assume vulnerability estimates as given, and focus only on the frequencies. This means that in the examples, we will make use of vulnerability functions without further explanation on their origins. However, it should be noted that the lack of data applies to vulnerability as well, and suitable theoretical models are needed there too, as we have argued in [19]. However, in the ideal case, vulnerability functions could be estimated from data obtained through penetration tests.

[4] Contrary to separate and fixed costs and likelihood in [14].

time, i.e. not interested in quick gain over higher long-term reward, when the expected damage per unit of time is equal.

Assuming that failure impacts are constant, we can combine vulnerability and impact in functions that represent the expected damage per unit of time for each possible decision (attack/wait, and which scenario). These can then be combined in a maximum expected damage function for *all* scenarios (i.e. the maximum expected damage for the available resources, plus the scenario that gives this maximum expected damage upon execution). We will detail these ideas in the following, first for the discrete and then for the continuous case.

5.2 Discrete event model

In the discrete event model, attackers save resources and attack with the accumulated resources at a single point in time. After the attack, the damage is assumed to be repaired, the attacker resources are reset to zero, and the same process will be repeated. At a later stage, additional variables such as repair time could be added. Attackers can, at each point in time, choose to launch an attack with the resources they have built up until that point ($I(t)$). Attackers will also have a skill level s, and the threat magnitude m is a function of the skill and the available resources:

$$m(t) = f(s, I(t)) \tag{18}$$

Attackers will *wait and save resources*, if they can cause higher expected average damage (induced risk) by launching an attack with more resources later; otherwise, they will *execute the scenario with the highest expected damage* given their current resources. For example, an attacker wishing to execute a distributed denial-of-service attack may acquire resources in terms of the size of the botnet available for the attack, and decide on the optimal target server and optimal attack time. To simulate this strategy, we define an expected damage $D'_c(m)$ for each threat capability level m and for each component/scenario c:

$$D'_c(m) = V_c(m) \cdot D_c \tag{19}$$

where D_c is the probable loss magnitude upon success of scenario c. The maximum expected damage for a given threat capability level m is specified by

$$\hat{D}(m) = \max_{c \in C} D'_c(m) \tag{20}$$

The optimal scenario to execute is then $\operatorname*{argmax}_{c \in C} D'_c(m)$.

We can therefore calculate the maximum expected damage at each point in time, and also the maximum average damage (risk) over the elapsed time.

$$\hat{D}(t) = \hat{D}(m(t)) \tag{21}$$

$$\hat{R}(t) = \hat{D}(m(t))/t \tag{22}$$

The attacker will thus attack at the time $\hat{t}$ when $\hat{R}(t)$ reaches its maximum. The scenario that will be executed is

$$\hat{c} = \operatorname*{argmax}_{c \in C} D'_c(m(\hat{t})) \tag{23}$$

After an attack, the attacker's resources will be reset to zero, and the damage will be repaired. As we are interested in the frequency *given a particular system architecture*, we calculate the frequency *without adding additional measures to the system after a successful attack*. When all relevant variables, i.e. V,

D, and I are constant in time, the next cycle will yield exactly the same result, or the best possible attack will take place at exactly the same time (now counting from the time of the first attack). In this case, the expected threat event frequency can be determined as

$$h = 1/\hat{t} \tag{24}$$

For all other scenarios, the threat event frequency is zero. The loss event frequency for scenario $\hat{c}$ is

$$\lambda_{\hat{c}} = V_{\hat{c}}/\hat{t} \tag{25}$$

From the loss event frequency, we can calculate the risk using the standard definitions outlined in Section 3. When needed in the analysis, precise points in time can be used instead of frequencies (and the frequency will then be zero until $\hat{t}$). We could also have calculated the average damage per resource unit instead of per unit of time. However, the model presented is more flexible when other than constant income functions would be considered. In case V, D and/or I are time-dependent, each attack cycle needs to be calculated separately. Heuristics would be needed in this case to prevent the calculations from becoming prohibitively complex, both in terms of computation time and in terms of understandability.

5.3 Continuous investment model

In the continuous investment model, attackers invest resources to attack components continuously. This case points towards a feature that we have skipped until now: gradual damage. All our previous models are memory-less, in the sense that earlier exposure to high threat levels does not lead to higher vulnerability. Only the current attack, or the current threat level, counts. For the malicious case this is inadequate, as partially successful (or only partially executed) attacks may definitely increase vulnerability to future attacker activity.

To solve this issue, we would need to express vulnerability in terms of the cumulative threat since some defined point in time. Moreover, in calculating the cumulative threat, we may wish to assign less weight to exposure that is further away in the past. The highest weight should be assigned to the current threat level.

For our first model, we simplify this issue by taking only the cumulative threat into account, without assigning weights. This can intuitively be understood as a case where an attacker would invest in cutting a tree, with previous efforts (partial success) permanently increasing the vulnerability of the tree. The frequency of success events (vulnerability) will thus increase with the total invested attacker resources. In other terms, the so-called mean time to (security) failure [16] will decrease with the invested resources.[5] After a breakdown, the invested resources for the scenario/component will be reset to zero.

The attacker will acquire resources specified by an income density function $i(t) = \frac{dI}{dt}$. The threat capability $m(t)$ is again a function of invested resources and skill. At each point in time, the attacker can choose in which scenario to invest. Scenario c has a success rate $V_c(m)$, which depends on the invested attacker resources (m_c). The expected damage for this scenario per unit of time (risk), as a function of invested resources, can be calculated as:

$$R_c(m) = V_c(m) \cdot D_c \tag{26}$$

Note that V is a frequency in the continuous model, not a probability, and that is why this equation gives the risk (damage per unit of time).

[5] We do not include diagnosis and repair times here, so mean time to failure equals mean time between failures.

If the attacker would only look at the short term, he would judge the marginal risk to determine his investment. At time t, the marginal risk $AR_c(t)$ for scenario c is the additional expected damage per unit of time achieved by investing more resources.

$$AR_c(t) = \frac{dR_c}{dm}(m_c(t)) = \frac{dV_c(m)}{dm}(m_c(t)) \cdot D_c \tag{27}$$

At each point in time, the attacker would invest his resources in the scenario that has the highest marginal risk, i.e. the steepest resource-risk curve.

$$\frac{dm_c}{dt}(t) = \begin{cases} \frac{dm}{dt} & \text{if } c = \underset{c \in C}{\operatorname{argmax}} AR_c(t), \\ 0 & \text{otherwise.} \end{cases} \tag{28}$$

If the attacker would adopt a long-term strategy, he would invest in the scenario which, given the investment, would have the largest damage divided by the expected time to failure. For a long-term oriented adversary, the aim is to maximise (in the long run) the average expected damage per unit of time (risk) for large t:

$$\bar{R}(t) = \frac{\int_0^t \sum\limits_{c \in C} V_c(m_c(t)) \cdot D_c \, dt}{t} \tag{29}$$

At each time, the attacker would invest such that this function is maximised for large t. Depending on the strategy, the attacker would thus invest in the scenario for which either marginal risk or average risk is maximal. The difference between short- and long-term strategies is especially relevant if investing in a particular scenario would yield very little in the beginning, whereas the expected damage curve would become steeper (and steeper than the others) after investing some initial resources (see Example 3 in the next section).

A scenario c is expected to succeed at the time t_c when the expected number of succes events equals 1. This is when

$$\int_0^t V_c(m_c(t)) \, dt = 1 \tag{30}$$

At that time, the damage is repaired, and the invested resources in that scenario (m_c) are reset to zero. This discontinuity may cause the attacker to switch scenarios, either to scenario c or from c to a different scenario. The loss event frequency for scenario c is

$$\lambda_c = 1/t_c \tag{31}$$

As the attacker may execute multiple scenarios in the continuous model, the risk needs to be calculated as the sum over all scenarios. The overall damage to the network within a specified time frame is calculated as the sum over all scenarios of their expected number of success events times the damage upon failure. The (average) risk is then the total damage divided by the total time.

6 Examples and simulations

In this section, we will present several examples to illustrate the approach and its theoretical properties. For explanation purposes, the examples have been simplified to yield easily understandable results. Realistic vulnerability functions may have different properties, but the essentials remain the same.[6] Also,

[6] See also footnote 3.

we do not indicate the units here, as the data is simulated anyway. To give some indication, units of time could be thought of as months, risk as euros per month, and vulnerability is a likelihood in the range [0..1].

As a theoretical rather than an empirical measure of attack rates, the examples, combined with the theoretical considerations in the next section, are meant as a provisional validation of the theory. The full decision support system structure, including examples of the complete analysis, will be prepared for publication at a later stage.

6.1 Discrete event model

Example 1. *Consider a system with 2 components. Both will cause damage € 1000 upon failure. Component 1 has vulnerability function $V_1(m) = \frac{m^3}{m^3+m^2+1}$, representing the probability of failure upon an attack with threat capability m. Component 2 has vulnerability function $V_2(m) = \frac{m^4}{m^4+m^2+1}$. We assume that the skill level is irrelevant here, and we therefore assume that $m(t) = I(t)$. The attacker has income density function $\frac{di}{dt} = 1$, i.e. the attacker will earn 1 resource unit per unit of time, or $m = t$. Up to $t = 1$, the attacker will thus be able to invest 1 unit.*

In Figure 2, the resulting induced risk functions $R_c(t)$ are shown. Note that in the beginning, the attacker would attack component 1 (if he would attack), whereas component 2 is more attractive later on. The optimal time to attack is around $t = 1.52$, targeting component 2, with the risk (average damage) being around € 406. The associated vulnerability is around 0.617. We thus have a mean time between attacks of 1.52, and a success probability of 0.617, yielding a mean time between failures of 2.46, or a loss event frequency of 0.406. This, multiplied by the damage, yields again the risk, also from the defender's point of view (as the analysis is white-box).

Note that in the discrete model, it is required that $V(0) = 0$. Otherwise, the expected risk (damage per unit of time) for very small t would be very high (up to infinite with t approaching 0), and the attacker would simply launch loads of "mini-attacks" with almost zero effort. This makes logistic vulnerability models unsuitable in combination with discrete attack rate models, as V is never zero in logistic functions.

6.2 Continuous investment model

Example 2. *Consider again a system with 2 components. Both will cause damage € 1000 upon failure. Component 1 has vulnerability function $V_1(m) = \frac{1}{2}m$, i.e. the expected failure frequency is equal to half the invested resources. Component 2 has vulnerability function $V_2(m) = \sqrt{m}$. The attacker has again constant income ($\frac{dm}{dt} = 1$).*

The attacker will decide on an optimal investment strategy. Therefore, he will choose functions $m_1(t)$ and $m_2(t)$, such that $m_1(t) + m_2(t) = m(t)$. To decide on the strategy, the attacker will calculate the expected time to failure for each of the components, given that he invests in them. For a short-term oriented adversary, the aim is to maximise the marginal risk. For this to work, the adversary needs to take into account that after the mean time to failure, the component needs to be replaced and the invested resources are reset to zero. The attacker can calculate the expected number of failures within a time frame by calculating the integral of the vulnerability function. This also allows calculation of the mean time to failure. If the attacker would only invest in component 1, he can calculate the mean time to failure by solving

$$\int_0^t \frac{1}{2} t \, dt = 1 \tag{32}$$

for t, yielding $t = 2$. For component 2, he solves

$$\int_0^t \sqrt{t} \, dt = 1 \tag{33}$$

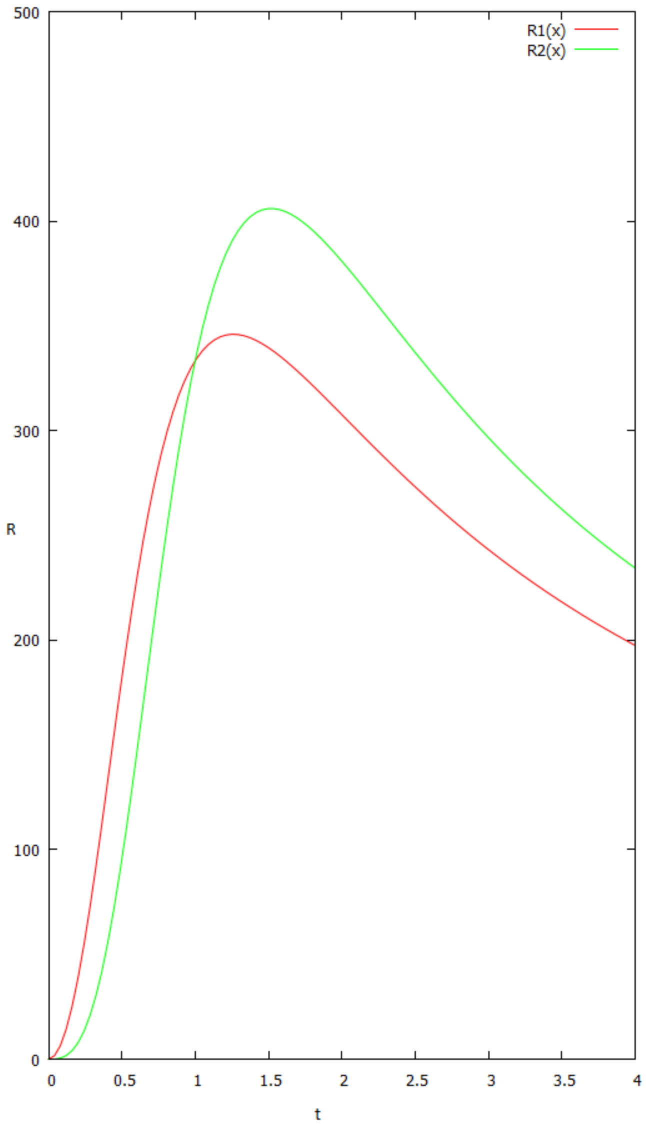

Figure 2: A simulation of attacker-induced risk as a function of attack time for a system with two components, using the discrete event model. The lines represent the attacker-induced risk for the components when the attacker would attack that component at that time, using his built-up resources. The optimal time and component to attack are derived from the highest point in the curves.

yielding $t = \sqrt[3]{\frac{9}{4}} \approx 1.3$. *Thus, if the attacker would only be able to invest in one component, he would invest in component 2, and the mean time to failure would be about 1.3. However, what if he can divide his resources? In that case, he can optimise his strategy by investing in the component with the highest increase in expected damage per unit of time (Figure 3). In that case, by dividing his resources between the two components, he can cause an average of 4 failures in 4.3 time units, instead of 3 in 3.9 time units.*

As the curve of component 2 has steeper slope, the attacker will start investing there. When the slope drops below 0.5, the attacker will switch to component 1. The mean time to failure occurs when the surface beneath the curve reaches 1. At that point, the component is expected to break down and be replaced, with the invested resources reset to zero. From the figure, one can tell that the mean time to

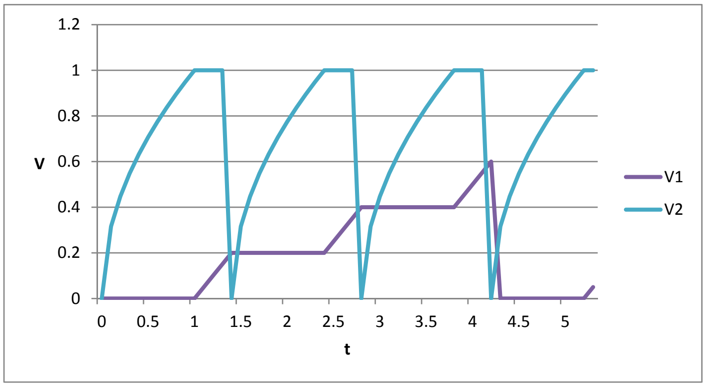

Figure 3: A simulation of attacker investment with 2 components, using the continuous investment model. The lines represent the vulnerability levels of the components. The vulnerability of component 1 increases linearly with invested resources, whereas the vulnerability of component 2 increases with the square root of the resources.

failure for component 1 is about 4.3, and for component 2 about 1.4. In this way, the attacker can cause a higher failure frequency (and thereby damage) than when investing only in component 2.

With mean times to failure of 4.3 and 1.4, the associated risk would be $\frac{1}{4.3} \cdot €1000 + \frac{1}{1.4} \cdot €1000 = €947$ *per unit of time.*

Example 3. *When we assume that* $V_2 = m^2$ *instead, the picture looks different (Figure 4). As the curve of component 1 has steeper slope at* $m = 0$*, the short-term attacker will start investing there, and never switch.*

However, if the attacker would adopt a long-term strategy, he would invest in component 2 instead, giving a higher loss event frequency (mean time to failure $2\sqrt{2} \approx 2.83$ *for component 1 vs.* $\sqrt[3]{6} \approx 1.82$ *for component 2.)*

Example 4. *Assume that in the context of Example 2, a countermeasure for component 2 is proposed. This countermeasure will reduce the vulnerability such that* $V_2 = \frac{1}{4}m$*, instead of the original* $V_2(m) = \sqrt{m}$*. This means that the attacker will now invest all resources in component 1, which has higher marginal damage upon investment (*$\frac{1}{2}$ *versus* $\frac{1}{4}$*). As already shown in Example 2, this will yield a mean time to failure of 2, corresponding to a risk of €500 per unit of time, compared to €947 for the original situation. If the cost of the countermeasure is less than €447 per unit of time, the investment would be cost-effective.*

6.3 The metrics in practice

As indicated above, the defined metrics are meant to support security investments, not as accurate predictors of empirical events. However, it would be interesting to study some empirical aspects, notably the relation between rational and actual attacker behaviour. Both attacker model and vulnerability model could benefit from data obtained from logs, forensic analysis, and experimental methods, and future research is needed to define precisely how the theoretical models could be improved based on such data. Extreme value theory may provide inspiration here [10].

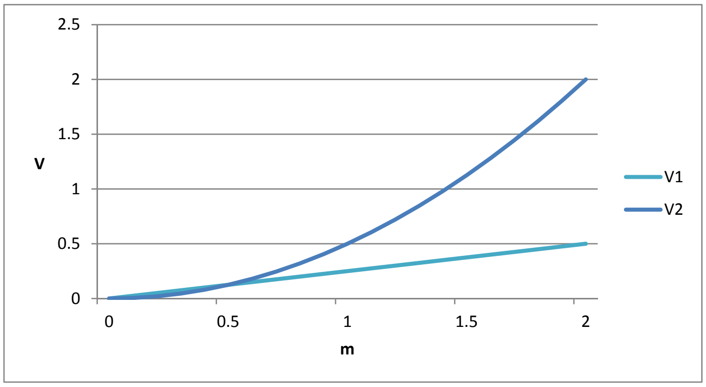

Figure 4: Another example of attacker investment with 2 components. The lines represent the vulnerability levels of the components. Note that the horizontal axis represents m here, not t, as the attacker is not able to invest in both components at the same time. The vulnerability of component 1 increases linearly with invested resources ($V_1(m) = \frac{m}{4}$), whereas the vulnerability of component 2 now increases with the *square* of the resources ($V_2(m) = \frac{m^2}{2}$).

Penetration testing is not particularly suitable as a setting, because the goal is to test various attack scenarios, and there is no incentive for the attacker/tester to select the "best" scenario. Serious games, such as red-team-blue-team assignments, would provide a better context. In this case, one can monitor the resources (time) available to the attackers and the defenders, the selected attacks, the "investments" by the defenders, etc.

In the SESAME project, the metrics can be applied in a decision-support system for security investments in electricity infrastructures. In the TRE$_S$PASS project[7], which focuses on socio-technical security models, they can be used in extensive, iterative case studies on cloud infrastructures, telecommunication networks, and customer privacy protection. In the TRE$_S$PASS project, we are also planning interviews with hackers of different kinds, in order to gather empirical evidence on their motivations and strategies. These could be used as input for improving the metrics, as well as for defining different types of attacker models.

7 Properties of the metrics

The metrics defined in this paper provide estimations of threat event frequencies and associated risk for malicious attacks, where the attacker tries to maximise damage to the system. These metrics can then be used in estimating the cost-effectiveness of countermeasures. We would like the metrics to have certain theoretical properties. In this section, we discuss to what extent the metrics satisfy these properties.

Property 1: Effectiveness of attacker resources The first property states that when we assume more resources on the part of the attacker, the attack rate should increase. This property is satisfied in both models, as the attacker will reach the optimal resources for attack earlier if he has more resources available.

[7] www.trespass-project.eu

Property 2: Effectiveness of vulnerability reduction The second property states that, for any reasonable risk metric, reduced vulnerability leads to equal or reduced attack rate for the less vulnerable component/scenario, and reduced overall risk. This means that, (1) the attacker will not invest more in this scenario if its vulnerability is reduced (all other variables remaining equal), and (2) if the attacker switches to a different scenario because of the reduced vulnerability, he cannot thereby increase the risk.

The risk-related part of the property only holds if the time-bias of the attacker and the time-bias of the defender are the same, i.e., the property will hold for the marginal risk if the attacker has a short-term strategy, and will hold for the average risk if the attacker has a long-term strategy. For example, if the attacker has a long-term strategy, the original target scenario may not have the highest marginal risk (but only the highest average risk in the long term), and the attacker may switch to a different component with higher marginal risk.

Property 3: Rationality of resource spreading The third property states that investing in different attacks provides utility to the attacker under certain constraints. A model does not seem very realistic if it only allows for one attack as an outcome. This property does not hold in the discrete event model, as the attacker will make the same decision on the target scenario every time, unless something changes in the system. This could be remedied when we include repair times during which the scenario is unavailable, but this would only change the situation if the attacker would launch another attack before the previously attacked component has been repaired. Because of this limitation, under the above assumptions, the continuous model would be the preferable one with respect to this property, as spread investments can be explained.

Property 4: Ability to estimate model from data The fourth property states that it should be possible to adjust the model based on observations of actual attacks. We have argued in the introduction that past frequencies do not constitute appropriate predictors for future attacks, as attacker strategies adapt to the environment. However, an adequate frequency metric would allow adjustment based on observed past frequencies. In our model, it would be possible to *estimate attacker investment* based on past frequencies, for a generalised "attacker" that covers all adversaries. That is, using data of past attacks and the vulnerability of the components involved, it can be estimated how much effort the attacker put into the attacks. Assuming constant income for the attacker (which is tricky by itself, and may require another model to enhance the predictions), we can estimate future frequencies, taking changes to the system (countermeasures) into account. Thus, rather than considering past frequencies as an appropriate attacker model, we estimate *past resources* from past frequencies, and use the resource level as the attacker model.

8 Open questions

As the proposed approach to quantify loss event frequencies is new, there are many open problems to be discussed in the research community. In the following we discuss some open questions within the proposed research paradigm.

Recovery times The present models assume a negligible recovery or repair time. A question for future research is how repair times would influence the results of the model. As we have argued above, repair times (during which attack does not make sense) may make it more attractive to invest in different scenarios during repair.

Learning effects When attackers execute the same type of attack multiple times, their skill may increase, and consequently, they may need fewer resources for the next attack. The explicit distinction of skills and resources, and the simulation of associated behaviours, would be another possible addition to the model. Similarly, attackers might obtain additional income from successful attacks, and might spend this on new attacks.

Knowledge about attack status While executing an attack, the attacker may acquire knowledge about the status of the attack. This may adjust his estimations on the likelihood of success as a function of invested resources. This may influence his investment strategy.

This issue is particularly interesting in the case of multi-step attacks [18], where the attacker gets feedback on the success of each step. Adapting the prediction strategy to multi-step attacks is a profound research question. In particular, one would need to assess which multi-step attacks lead to which impact if successful, and how likely each of the steps is too succeed, depending on the effort spent. Inspiration for possible directions can be found for example in [14].

Weights of previous investments What is the best model for representing the effect of previous investments in attacks? In this paper, we assumed that investments remain fully effective forever. With a more advanced model, we can represent degradation of previous investments over time. Assuming the use of a time weighting function, with the highest value for present investment, one could for example calculate the threat magnitude level for a scenario as

$$m_c(t) = \int_0^t w(u,t)\frac{dm_c}{dt}(u)\,du \tag{34}$$

with $\frac{dm_c}{dt}$ attacker investment density. Weights can then be assigned based on the difference between investment time u and current time t, for example

$$w(u,t) = e^{\alpha(u-t)} \tag{35}$$

Multiple attackers Another extension to the model could be approaches to study multiple attackers and joint strategies. In the case of multiple attackers, attackers $A_1..A_n$ at each point in time can confront the system with their built-up resources. They can do so in isolation or in cooperation. In the first case, the aggregate vulnerability would be based on the success probabilities of individual resources, and would be calculated as:

$$V = 1 - \prod_{A_i}(1 - V(m_{A_i})) \tag{36}$$

This represents the probability that the attack scenario succeeds due to any of the attackers, assuming the probabilities are independent. The calculation first assesses the probability that the component survives all attacks.

When the attackers cooperate, the aggregate vulnerability would be the success probability of the sum of the individual resources, as they now pool their resources for a single attack:

$$V = V(\sum_{A_i} m_{A_i}) \tag{37}$$

As these calculations are different, attacker cooperation could influence investment strategies.

Interaction with failures caused by other threats Components do not only fail due to attacks, but also due to natural or accidental threats. When this happens, the attacker-induced risk is lower than estimated in the present models, as a baseline risk already exists. It could be investigated how this would affect attacker strategies, and thereby threat event frequency estimates.

Different attacker utility In this paper, we have discussed the case of a terrorist attacker. When considering for example financially motivated attackers, one would have to take the difference between damage to the system and attacker gain into account. Thus, the outcome of an attack would be perceived differently by the attacker and the defender, also creating different risk perspectives, and thereby requiring more complicated models. Besides, the attacker may use financial gain from one attack as resources for executing another.

In addition, the risk of getting caught could be added as an additional parameter, instead of including this as part of the invested resources [5, 8]. This would allow distinguishing between risk-seeking and risk-averse attackers. Furthermore, we may want to experiment with probabilistic rather than deterministic attacker models, where attackers will not always select the optimal attack.

Timed countermeasures Finally, consider what would happen if the defenders would try to find the optimal time for countermeasure deployment. If costs of a countermeasure are distributed unevenly over time, can we say something about the optimal time to invest? Approaches like net present value and real options analysis could be useful here, to account for the fact that investments at different times should be weighed differently.

9 Conclusions

This paper proposes a new approach for cyber risk analysis that combines frequentist and adversarial approaches to risk in a single framework. Rather than estimating likelihood of threats as a single value, the paradigm separates threat event frequency from vulnerability, in order to be able to assess the cost-effectiveness of countermeasures. In this paper, we focused on the question how to estimate the threat event frequencies and loss event frequencies of a system, in particular for malicious threats. We discussed two possible models for estimation of threat event frequencies from expected attacker resources, and identified directions for future research.

Combined with our earlier work on vulnerability [19], this research leads to a complete framework for determining how often attack scenarios in a system are expected to succeed. We apply this method within two projects, where the focus is on security assessment of electricity infrastructures and socio-technical systems, respectively. In the electricity project (SESAME), the assessment of component failures can be combined with cascading failure analysis and impact assessment to estimate the risk caused by the different threats, as well as the risk prevented by countermeasures.

Even though the foundations of the model are theoretical, data is still required for practical applications. In particular, one would need to have estimates of the damage caused by different failures in a system. For electricity networks, while one may simulate the network evolution after a failure, damage caused to society by a blackout is more elusive. In any system that is subject to cyber attacks, expert knowledge on damage caused by failures is essential to estimate risk. Also, the estimation of realistic vulnerability functions (i.e. the function from threat capability to success probability) is not trivial, and may require experiments such as penetration testing by professional hackers. Finally, empirical input to the model in terms of previous attacks can improve its accuracy, but this requires the sharing of information about such attacks. Platforms to facilitate this, as well as trust in proper handling of the sensitive information involved, would be of great value to the overall context of countermeasure evaluation.

In future work, we will further extend the threat event frequency analysis, to accommodate the open questions outlined above. In particular, we will focus on extending the framework to multi-step attacks. This is another distinguishing feature of security as opposed to safety, next to the attacker adaptation to system design, which is the key contribution of the present results.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreements SEC-261696 (SESAME) and ICT-318003 (TRE$_S$-PASS). This publication reflects only the authors' views and the Union is not liable for any use that may be made of the information contained herein. The authors wish to thank (in alphabetical order) Andrew Bollinger and Sanne van der Ven for helpful brainstorm sessions.

References

[1] S. Arianos, E. Bompard, A. Carbone, and F. Xue. Power grid vulnerability: A complex network approach. *Chaos*, 19(1):013119, 2009.

[2] A. Beautement, M. A. Sasse, and M. Wonham. The compliance budget: managing security behaviour in organisations. In *Proc. of the 2008 New Security Paradigms Workshop (NSPW'08), Plumpjack Squaw Valley Inn, Lake Tahoe, California, USA*, pages 47–58. ACM, September 2008.

[3] R. Böhme and J. Grossklags. The security cost of cheap user interaction. In *Proc. of the 2011 New security paradigms workshop (NSPW'11), Marin County, California, USA*, pages 67–82. ACM, September 2011.

[4] E. Bompard, C. Gao, R. Napoli, A. Russo, M. Masera, and A. Stefanini. Risk assessment of malicious attacks against power systems. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 39(5):1074–1085, 2009.

[5] A. Buldas, P. Laud, J. Priisalu, M. Saarepera, and J. Willemson. Rational choice of security measures via multi-parameter attack trees. In *Proc. of the 1st international conference on Critical Information Infrastructures Security (CRITIS'06), Samos Island, Greece, LNCS*, volume 4347, pages 235–248. Springer Berlin Heidelberg, August-September 2006.

[6] A. A. Cárdenas, S. Amin, Z.-S. Lin, Y.-L. Huang, C.-Y. Huang, and S. Sastry. Attacks against process control systems: Risk assessment, detection, and response. In *Proc. of the 6th ACM Symposium on Information, Computer and Communications Security (ASIACCS'11), Hong Kong, China*, pages 355–366. ACM, March 2011.

[7] L. A. Cox Jr. Game theory and risk analysis. *Risk Analysis*, 29(8):1062–1068, 2009.

[8] M. Cremonini and D. Nizovtsev. Understanding and influencing attackers' decisions: Implications for security investment strategies. In *Proc. of the 5th Workshop on the Economics of Information Security (WEIS'06), Robinson College, University of Cambridge, England, UK*, June 2006.

[9] M. Cremonini and D. Nizovtsev. Risks and benefits of signaling information system characteristics to strategic attackers. *Journal of Management Information Systems*, 26(3):241–274, 2009.

[10] P. Embrechts. Extreme value theory: Potential and limitations as an integrated risk management tool. *Derivatives Use, Trading & Regulation*, 6(1):449–456, 2000.

[11] V. Garg, N. Husted, and J. Camp. The smuggling theory approach to organized digital crime. In *Proc.of the 2011 eCrime Researchers Summit (eCrime'11), San Diego, CA, USA*, pages 1–7. IEEE, November 2011.

[12] B. Grobauer, T. Walloschek, and E. Stocker. Understanding cloud computing vulnerabilities. *IEEE Security & Privacy*, 9(2):50–57, March-April 2011.

[13] E. Jonsson. Towards an integrated conceptual model of security and dependability. In *Proc. of the 1st International Conference on Availability, Reliability and Security (ARES'06), Vienna University of Technology, Austria*, pages 646–653. IEEE, April 2006.

[14] E. LeMay, M. D. Ford, K. Keefe, W. H. Sanders, and C. Muehrcke. Model-based security metrics using adversary view security evaluation (ADVISE). In *Proc. of the 8th International Conference on Quantitative Evaluation of Systems (QEST'11), Aachen, Germany*, pages 191–200. IEEE, September 2011.

[15] B. Littlewood, S. Brocklehurst, N. Fenton, P. Mellor, S. Page, D. Wright, J. Dobson, J. McDermid, and D. Gollmann. Towards operational measures of computer security. *Journal of Computer Security*, 2(2–3):211–229, 1993.

[16] B. B. Madana, K. Goševa-Popstojanovab, K. Vaidyanathanc, and K. S. Trivedia. A method for modeling and quantifying the security attributes of intrusion tolerant systems. *Performance Evaluation*, 56:167–186, 2004.

[17] W. L. McGill, B. M. Ayyub, and M. Kaminskiy. Risk analysis for critical asset protection. *Risk Analysis*, 27(5):1265–1281, October 2007.

[18] V. Nunes Leal Franqueira, R. H. C. Lopes, and P. A. T. van Eck. Multi-step attack modelling and simulation (MsAMS) framework based on mobile ambients. In *Proc. of the 24th Annual ACM Symposium on Applied Computing (SAC'09), Honolulu, Hawaii, USA*. ACM, March 2009.

[19] W. Pieters, S. H. G. Van der Ven, and C. W. Probst. A move in the security measurement stalemate: Elo-style ratings to quantify vulnerability. In *Proc. of the 2012 New Security Paradigms Workshop (NSPW'12), Bertinoro, Italy*, pages 1–14. ACM, September 2012.

[20] J. Reichl, M. Schmidthaler, and F. Schneider. The value of supply security: The costs of power outages to Austrian households, firms and the public sector. *Energy Economics*, 36:256–261, 2013.

[21] D. Rios Insua, J. Rios, and D. Banks. Adversarial risk analysis. *Journal of the American Statistical Association*, 104(486):841–854, 2009.

[22] J. J. C. H. Ryan and D. J. Ryan. Expected benefits of information security investments. *Computers & Security*, 25(8):579–588, 2006.

[23] S. E. Schechter. Toward econometric models of the security risk from remote attacks. *IEEE Security & Privacy*, 3(1):40–44, 2005.

[24] The Open Group. Risk taxonomy. Technical Report C081, The Open Group, 2009.

[25] R. R. Youngs and K. J. Coppersmith. Implications of fault slip rates and earthquake recurrence models to probabilistic seismic hazard estimates. *Bulletin of the Seismological Society of America*, 75(4):939–964, 1985.

Author Biography

Wolter Pieters has Master degrees in both computer science and philosophy of technology from the University of Twente, and a PhD degree in information security from Radboud University Nijmegen, The Netherlands. Currently he is technical leader of the TRE$_S$PASS project at the University of Twente, and assistant professor cyber risk at Delft University of Technology. In the TRE$_S$PASS project, he addresses cyber security risk management in socio-technical systems through the concept of attack navigators, including research on security policies and security metrics. He also published on electronic voting, verification of security properties, and philosophy and ethics of cyber security.

Zofia Lukszo studied applied mathematics at Technical University of Lodz and philosophy at University of Lodz, Poland. In 1996 she received at the Eindhoven University of Technology, the Netherlands, the Ph.D. degree for the thesis “A Practical Approach to Recipe Improvement and Optimization in the Batch Processing Industry”. Since 1995 Zofia has worked on the Faculty of Technology, Policy and Management at Delft University of Technology, the Netherlands. She is there an associate professor in the Energy and Industry group. She is also a leader of the programme Intelligent Infrastructures within the international research programme on Next Generation Infrastructures. The Intelligent Infrastructure sub-programme concentrates on a wide range of problems in the way infrastructures are functioning today, and aims to develop new, intelligent concepts for modelling, optimization and control of their operation resulting in more effective, efficient, safe and reliable utilization.

Dina Hadžiosmanović studied computer science at the University of Sarajevo, Bosnia and Herzegovina. In 2014 she received a PhD degree in system security from University of Twente, The Netherlands. Her thesis focused on improving cyber security in industrial control systems. Currently, she works as a researcher at the Delft University of Technology on different aspects of cyber security in critical infrastructures like smart grids and flood barriers. In addition, she is closely involved in the TRE$_S$PASS project, where she works on incorporating infrastructure information into adaptive risk management.

Jan van den Berg is currently full professor cyber security at Delft University of Technology. He also acts as chairman of the Center of Safety & Security of the 3 universities Leiden University, Delft University of Technology and Erasmus University Rotterdam, and as scientific director of the Cyber Security Academy The Hague. His PhD (1996) was in the area of computational intelligence. He further did research in the areas of (intelligent) data analytics (with applications in finance, agriculture, e-learning, a.o.), knowledge discovery, and information security.