Employing key indicators to provide a dynamic risk picture with a notion of confidence

Atle Refsdal and Ketil Stølen

Abstract A security risk analysis will only serve its purpose if we can trust that the risk levels obtained from the analysis are correct. However, obtaining correct risk levels requires that we find correct likelihood and consequence values for the unwanted incidents identified during the analysis. This is often very hard. Moreover, the values may soon be outdated as the system under consideration or its environment changes. It is therefore desirable to be able to base estimates of risk levels on measurable indicators that are dynamically updated. In this paper we present an approach for exploiting measurable indicators in order to obtain a risk picture that is continuously or periodically updated. We also suggest dynamic notions of confidence aiming to capture to what extent we may trust the current risk picture.

1 Introduction

In order for a security risk analysis of a computer system to serve its purpose, we need to trust that the risk levels obtained for the identified risks are (at least roughly) correct. This requires finding good answers to the following questions: 1) How likely is the unwanted incident in question to occur? 2) What is the consequence if this incident occurs? 3) How do the consequence value and the likelihood value combine into a single risk value? Unfortunately, in most cases the answers obtained from a risk analysis will provide a snapshot reflecting a single point in time. Hence, the risk values may soon be outdated as the system or its environment change.

Moreover, finding correct likelihood values is often very hard. This is typically the case if we are analyzing a new system where historical data do not exist, or if the incident in question cannot easily be observed directly, such as an eavesdropper

Atle Refsdal
SINTEF ICT, Norway, e-mail: Atle.Refsdal@sintef.no

Ketil Stølen
SINTEF ICT, Norway, e-mail: Ketil.Stolen@sintef.no

reading a sensitive e-mail. Finding correct consequence values may also be difficult, particularly in cases where the asset we seek to protect is not easily measured in terms of money, such as confidentiality of sensitive data.

All this means that we need to seek ways of obtaining good estimates of likelihood and consequence values. One way of doing this is to base the assessments on measurable indicators that are seen as relevant for the unwanted incident in question, even though its likelihood or consequence value cannot be directly inferred from any of these indicators. For example, if we want to estimate the likelihood that an intruder accesses sensitive data by logging on to a system with the username and password of a legitimate user, it may be useful to know how many passwords have not been changed during the last three months and how many of the users do not comply with the company's password strength policy. If likelihood, consequence and risk levels are defined as functions of indicators, we also ensure that risk levels can be updated as soon as the indicators are updated, rather than representing a snapshot at a given point in time.

Having assigned likelihood values to the relevant threat scenarios and unwanted incidents, we are also interested in estimating the level of confidence we may have that the risk levels obtained are in fact correct, and to uncover weaknesses of the analysis. One way of achieving this is to check whether the risk picture is consistent with respect to likelihood values. This can be done by assigning likelihood not only to the unwanted incident that harms an asset, but also to the potential threat scenarios that may lead to this incident. The likelihood of the unwanted incident can then be compared to the likelihood of the threat scenarios.

This paper presents an approach for providing a dynamic risk picture and for assessing to what degree we can be confident that the risk levels obtained are correct. A basic assumption of the approach is that an infrastructure is available for defining and monitoring the measurable indicators required. Providing such an infrastructure is an important goal for the project MASTER (see http://www.master-fp7.eu/), which addresses the challenge of managing assurance, security and trust for service-oriented systems. Although the work presented has been carried out within the context of the MASTER project, the approach we present is general in the sense that we just assume the availability of a palette of monitored indicators; the infrastructure required to obtain them is not considered.

The rest of this paper is structured as follows: In Section 2 we present the conceptual model on which the approach is based. A high-level description of the approach, as well as the underlying assumptions, is given Section 3, which ends with presenting three steps that need to be performed in order to carry out the dynamic risk monitoring. Based on an example case, these three steps are further explained in Sections 4, 5, 6. We then explain how the internal consistency of the risk picture can be checked in Section 7, and discuss measures of confidence in the analysis in Section 8. Some related work is presented in Section 9, before we conclude in Section 10.

2 Conceptual model

Figure 1 shows the conceptual model for risk and closely related concepts on which our approach is based. The model is shown as a UML class diagram with explanatory text on the associations. A risk involves an unwanted incident, such as sensitive patient data being disclosed to outsiders. The unwanted incident may occur with a certain likelihood. When it occurs, an asset will be damaged (and its value reduced) – this is the consequence of the risk. An asset is something of value that we seek to protect. Assets can be anything from physical objects such as computers to abstract entities such as confidentiality of information or the reputation of a stakeholder. If the asset we are concerned with is the reputation of the hospital, and the identified incident is sensitive patient data being disclosed to outsiders, then the consequence related to this incident could be a certain reduction of (or damage to) the hospital's reputation. In the diagram, we have assigned consequence directly to the risk, rather than to the unwanted incident. This has been done in order to emphasize that the consequence of an incident, unlike its likelihood, is not a property of the incident per se, as the consequence also depends on the particular asset in question and its measure.

In order to obtain a clear risk picture and be able to choose and prioritize between treatments, we need to assign a risk value to each risk that has been identified. This is done by applying the risk function, which takes the consequence and the likelihood of the unwanted incident of the risk as input. Hence, consequence and likelihood need to be measured according to some suitable scale. Typically, likelihood is measured in terms of frequency or probability. Consequence may be measured by for example monetary value or the number of data items affected by the incident, dependent on the nature of the asset in question. The risk function is defined by the risk analysis team and depends on the scales chosen for measuring consequence and

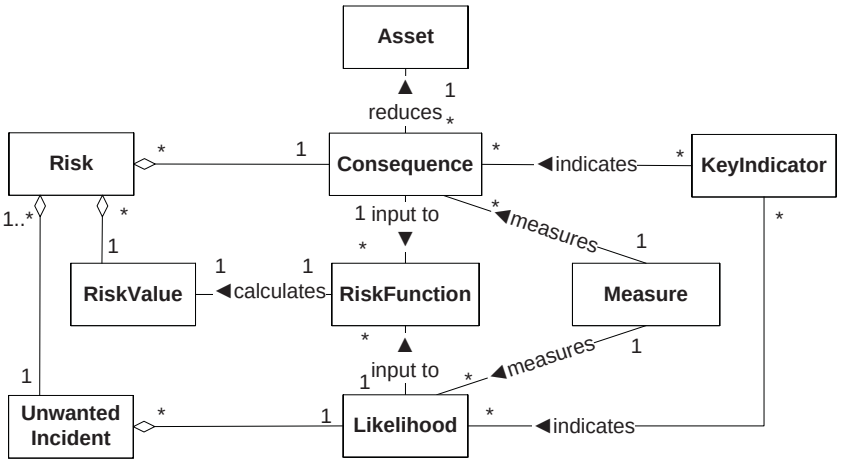

Fig. 1 Conceptual model for risk and closely related concepts.

likelihood. If we are measuring likelihood in terms of frequency and consequence in terms of monetary value, we may use multiplication to obtain a risk value. For example, from a consequence of 10000 euros and a likelihood of 3 times per year we get a risk value of 30000 euros per year. If qualitative scales are used for measuring consequence and likelihood, then the risk function defines how the possible consequence and likelihood values combine into a risk value. For example, a consequence value “catastrophic” and a likelihood value “seldom” may combine into a risk value “high”.

As discussed in Section 1, obtaining correct consequence and likelihood values is a major challenge. Therefore it may be highly useful to be able to identify relevant and measurable key indicators on which estimates can be based. Notice that the multiplicity symbol * on each end of the associations between the KeyIndicator class and the Likelihood class in Figure 1 means that there is a many-to-many relationship between key indicators and likelihoods; one likelihood may be obtained from several key indicators, and one key indicator may be used to obtain the likelihood of several unwanted incidents. The same holds for the relation between key indicators and consequences.

3 The approach

Figure 2 outlines our vision for a dynamic risk monitor defined on the top of some monitoring infrastructure. In this paper we just assume the availability of a monitoring infrastructure offering a palette of continuously monitored key indicators that we may select from. The key indicators are quantitative measures that are considered relevant for finding the likelihood or consequence of an unwanted incident. In many cases, indicator values will be calculated automatically by the system. For example, the system can recognize a certain kind of event and count the number of occurrences of such events. In other cases, the indicator values will be obtained manually. For example, the number of errors detected in a periodic review of a sample of the records stored in a database may be input into the system to provide an indicator value. Our envisaged dynamic Risk Monitor consists of three modules as indicated in Figure 2, and the rest of this paper is devoted to establishing the data required for it to function, given the assumed availability of a palette of key indicators.

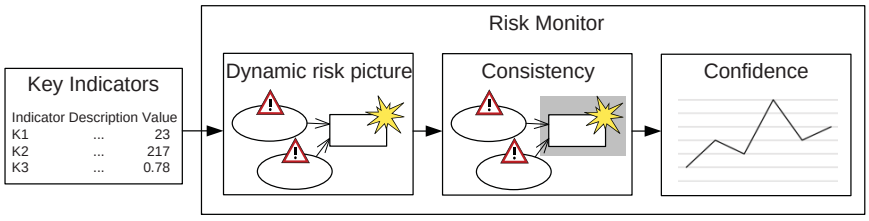

Fig. 2 Risk Monitor modules.

The "Dynamic Risk Picture" module allows the user to monitor the likelihood, consequence, and risk values, thereby providing a more high-level view than the "Key Indicators" infrastructure. Values may be presented in graphical diagrams that show how threat scenarios lead up to unwanted incidents; likelihood values may be assigned to threat scenarios as well as unwanted incidents. The values are obtained from functions for calculating likelihood and consequence values from sets of key indicators, as well as for calculating risk values from likelihood and consequence values. These functions are defined during the risk analysis of the system, as the relevant risks will depend on the system in question.

The "Risk Consistency" module checks whether the risk picture is consistent at a given point in time. This can be done by comparing likelihoods for threat scenarios assumed to lead up to an unwanted incident with the likelihood of the actual incident. For example, if a certain threat scenario is assigned probability "twice a year", and is assumed to lead to a certain unwanted incident with probability 0.5, then the likelihood assigned to the unwanted incident should be at least "once a year". But it can also be higher, if there are other threat scenarios leading up to the same incident. In Section 7 we present calculation rules taken from [BDS08] that can be utilized in order to check whether the likelihood values assigned to a diagram are consistent.

Finally, the "Confidence" module offers a quantitative measure of confidence in the current risk picture, thereby providing an aggregated view from which the correctness of the analysis can be assessed. The aim is to estimate to what extent we may trust that the risk levels are correct based on the degree of inconsistency detected in the risk picture.

The programming of the dynamic risk monitor may of course be work consuming, but not very challenging from a research point of view. The real issue of research is how to come up with the data to display, and this is what we concentrate on in the rest of the paper. We propose an approach of three steps that need to be carried out before monitoring can start, which we will describe in further detail in the next sections:

1. *Perform an initial risk analysis of the system.* This step serves a number of purposes. It provides information about what are the relevant risks and a rough analysis of the risk levels, so that a decision can be made of which risks need to be monitored. Furthermore, it provides information about how threats exploit vulnerabilities to initiate threat scenarios leading to unwanted incidents, which is essential for the later steps.
2. *Identify relevant key indicators for the risks to be monitored.* This is done based on the understanding obtained through the initial analysis in the previous step. Indicators may be related not only to an unwanted incident that is directly associated with a risk, but also to vulnerabilities and threat scenarios leading up to this incident.
3. *Find functions for likelihood, consequence, and risk values.* Likelihood and consequence values are calculated from sets of key indicators, while risk values are calculated from the likelihood and consequence value for the unwanted incident in question.

4 Performing the initial risk analysis of the system

We consider a hospital concerned about protecting the integrity of patient records. All details have been made up. Hence, the unwanted incidents, threat scenarios, risk levels and other aspects of the analysis presented here do not reflect any real case.

The patient record database can only be accessed from terminals in the hospital's office area. Access to the terminals is protected by user names and passwords. A password strength policy has been issued informing employees of requirements with respect to passwords length, use of numerical characters and so on. In addition, users are expected to change their password every third month. Users do not in general have access to all the patient records. For example, a doctor has only access to the records of her/his own patients. Before leaving a terminal, users are supposed to log off. Users are logged off automatically if a terminal has been inactive for a certain time (the delay logoff interval).

The doors into the office area are normally kept closed and locked at all times, and are fitted with keycard locks. In order to open a door, a keycard has to be inserted into the lock. The door will then open up and remain open for a certain interval in order to allow entry. If the keycard lock of a door is defective the door will unlock. Keycard locks are fitted with a failure detection system that generates a signal if the lock is defect.

The asset that we seek to protect for the purpose of this example is integrity of patient data. There is, of course, many ways in which we can imagine that this asset may be harmed. In order to limit the scope, the analysis will be restricted to external threats, and we consider only cases where such data are accessed by intruders into the office area that are not part of the hospital staff. To conduct the initial risk modeling and analysis of the system we employ CORAS. However, other approaches may also be used.

CORAS [dBHL$^+$07] provides a method, a language, and a tool for asset-oriented risk analysis. The CORAS method consists of seven steps. For the purpose of this paper, we focus on only a few of these. In order to assign likelihood and consequence values to unwanted incidents, we need to establish some suitable scales for this that are useful for making assessments. Table 1 shows the scales that will be used to measure consequence for the "Integrity of patient records" asset in this paper, as well as the scale that will be used for measuring likelihood. Note that by choosing

Table 1 Consequence scale (left) and likelihood scale (right)

Consequence	Description	Likelihood	Description
Catastrophic	$>$ 400 records affected	Very often	$>$ 100 times per year
Major	101-400 records affected	Often	21 – 100 times per year
Moderate	21-100 records affected	Sometimes	6 – 20 times per year
Minor	3-20 records affected	Seldom	3 – 5 times per year
Insignificant	0-2 records affected	Very seldom	$\leq$ 2 times per year

to measure consequence for integrity of patient data only in terms of the number of records affected, we do not distinguish between different levels of importance for different records. If necessary, we could have identified separate assets for records based on their importance.

After deciding upon the suitable scales for consequence and likelihood, the analysts establish the risk evaluation criteria that states which level of risk the client accepts for each asset. The result is typically recorded as a risk matrix that shows which combinations of consequence and likelihood values that are acceptable and which are not. Table 2 shows such a matrix. A gray box means that the risk level is so high that the risk needs to be further evaluated for treatment.

Table 2 Risk evaluation matrix

		Consequence				
		Insignificant	*Minor*	*Moderate*	*Major*	*Catastrophic*
Frequency	*Very seldom*					
	Seldom					
	Sometimes					
	Often					
	Very often					

After establishing the risk evaluation criteria, the next steps concern identifying potential unwanted incidents and the scenarios leading up to these incidents. The result is documented in *threat diagrams*. Threat diagrams show how threats exploit vulnerabilities to initiate threat scenarios and unwanted incidents, and what assets are harmed if the unwanted incident occurs. A threat scenario is a scenario that may lead to an unwanted incident or to another threat scenario. Figure 3 shows the symbols used to denote threats (of three different kinds), vulnerabilities, threat scenarios, unwanted incidents and assets. Except from vulnerabilities, these elements are referred to as vertices. Figure 4 shows a threat diagram for the hospital example. Note that this is not intended to represent a complete analysis. In addition to the symbols shown in Figure 3, a threat diagram contains relations represented by arrows between the vertices, possibly via one or more vulnerabilities. These vulnerabilities are then considered to be a part of the relation. Hence, a threat diagram consists of a set of vertices and a set of relations between the vertices. For a formal definition of the language, see [BDS08]. There are three kinds of relations: “initiate”, “leads-to” and “impact”. An “initiate” relation goes from a threat to a threat scenario or an unwanted incident, and shows that the threat initiates the threat scenario or unwanted incident. Possibly, the threat achieves this by exploiting one or

Fig. 3 Basic building blocks of CORAS threat diagrams

more vulnerabilities, which are then shown on the arrow from the threat to the threat scenario or unwanted incident.

The "initiate" relation in the left-hand part of Figure 4 shows that the threat "Intruder" exploits the vulnerability "Weak access control to hospital offices" to initiate the threat scenario "$S1$: Intruder accesses patient records terminal". [1] From threat scenario $S1$ there is a "leads-to" relation showing that this scenario may lead to the threat scenario $S2$ via the vulnerability "Weak logoff dicipline". This means that the intruder accesses a terminal where the previous user has not logged off before leaving the terminal. The "leads-to" relation from $S2$ show that this scenario may lead to the unwanted incident $U1$, which impacts the asset "Integrity of patient records". From threat scenario $S1$ there is also another "leads-to" relation to $S3$ showing that the intruder may achieve the same unwanted incidents by logging on to a terminal with the user name and password of an employee.

Having identified the unwanted incidents, threats, vulnerabilities, and threat scenarios, as well as the "initiate", "leads-to" and "impact" relations between them, the next step is to assign likelihood and consequence values. This has also been done in Figure 4. Likelihood values are inserted in brackets on the threat scenarios and unwanted incidents, while consequence values are inserted on the "impact" relations from unwanted incidents to assets. Probability intervals have also been assigned to the "leads-to" relations from threat scenarios to unwanted incidents.

5 Identifying relevant key indicators

In order to calculate and monitor risk values based on key indicators, we first need to identify the indicators that are of relevance for the risks in question. As seen in Section 4, CORAS threat diagrams illustrate graphically how unwanted incidents result from threats exploiting vulnerabilities to initiate threat scenarios. These diagrams can be exploited in a structured brain storming in order to identify relevant key indicators. The analysis leader can direct the attention of the analysis team to

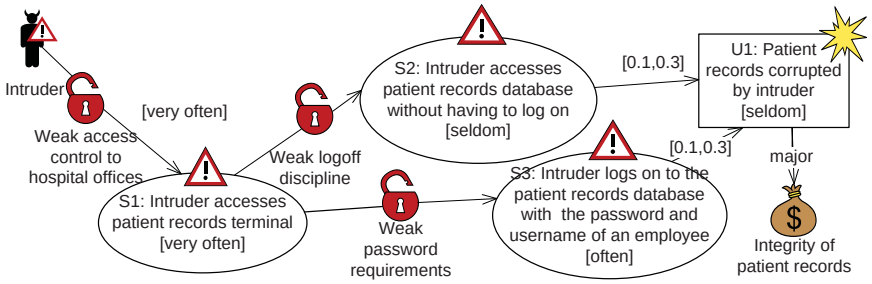

Fig. 4 Threat diagram with likelihood and consequence estimates

[1] We use $S1, S2, S3$ as shorthand names for threat scenarios, and $U1$ for the unwanted incident.

the different elements of the diagram one at a time, each time asking for suggestions for suitable indicators and noting these down at the relevant place in the diagram. Thereby, the team is encouraged to think about not only indicators directly associated with the unwanted incident, but also indicators that are more closely related to vulnerabilities and scenarios leading up to the incident. Figure 5 shows a possible result of applying this process on the diagram in Figure 4. Indicators have been chosen in order to illustrate different aspects of the approach, and do not represent a real (or exhaustive) analysis. Key indicators are shown as boxes with a dial in the upper right-hand corner, and are attached to vulnerabilities, threat scenarios, unwanted incidents and "impacts" relations. The indicators $K1$ "Time interval for opening of doors" and $K2$ "Total time that keycard locks have been defect during the last 3 months" have been attached to the vulnerability "Weak access control to hospital offices". The reasoning is that it will be easier for an intruder to access the office area if doors remain open for a relatively long time after someone has entered or left, or if a keycard lock is defect so that the door is unlocked. For the vulnerability "Weak logoff discipline", the indicators $K3$ and $K4$ have been identified as relevant, measuring how often it occurs that users are automatically logged off due to inactivity, and the length of the logoff delay interval, respectively. For assessing the likelihood of the unwanted incident "Patient data corrupted by intruder", we assume that periodic reviews are held of random samples of the patient records. The doctors of the patients in question are asked to go through the records to check that the data are correct. For example, recorded treatments should match the patient's disease. The indicator $K5$ measures the percentage of records with errors reported by doctors in the sample review. The indicators $K6$ "Number of passwords that do not fulfill password strength policy" and $K7$ "Number of passwords that have not

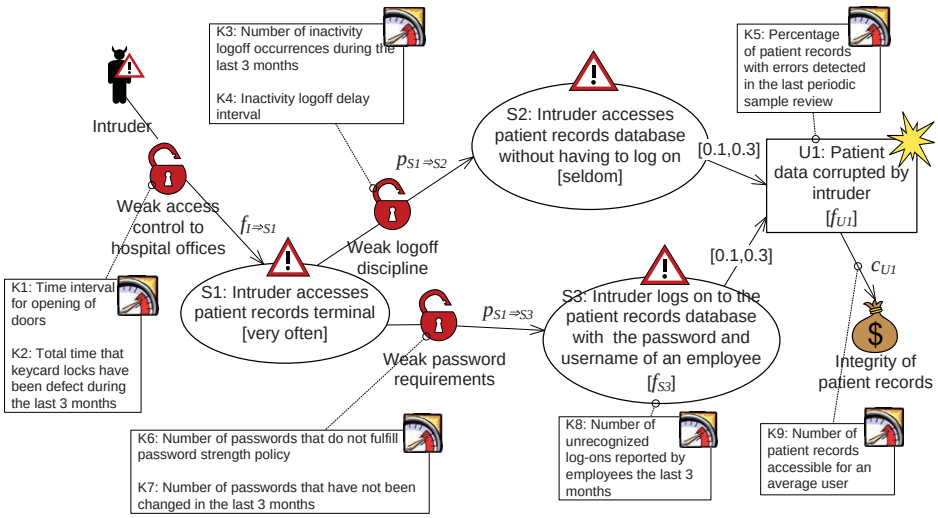

Fig. 5 Threat diagram with indicators attached

been changed in the last 3 months" have been identified for the vulnerability "Weak password requirements". For threat scenario $S3$, the indicator $K8$ has been introduced as an aid in assessing the likelihood. The idea is that a message with the date and time of the previous log-on pops up each time a user logs on, and the user is asked whether she or he can confirm that this is correct. The indicator $K8$ measures the number of occurrences where users give a negative answer during the last three months. Finally, the indicator $K9$ measures the number of patient records that are accessible for an average user. This indicator is attached to the "impacts" relation from the unwanted incident "Patient data corrupted by intruder" to the asset "Integrity of patient records", showing that this indicator can be used as an aid in assessing the consequence of this incident.

Note that likelihood and consequence values have been replaced by function names for the vertices and relations for which indicators have been identified. These functions are explained in the next section.

6 Finding functions for likelihood, consequence, and risk values

After identifying relevant key indicators for the risks to be monitored, the next step is to define functions for calculating likelihood and consequence values from the indicators. This is done for all the vertices and relations for which relevant indicators have been identified. For example, for the "initiate" relation from the threat "Intruder" to $S1$ we need to define a function that calculates the frequency with which an intruder initiates this threat scenario from the indicators $K1$ and $K2$.

We use function names with subscripts to show which function we are dealing with, according to the following convention: The first letter denotes the type of the output of the function; f for frequency, p for probability, and c for consequence. The subscript denotes which frequency/probability/consequence we are talking about, and we use $\rightarrow$ in the subscript when referring to a relation between two vertices. For example, $f_{I\rightarrow S1}(K1,K2)$ denotes the function for calculating the frequency with which the intruder initiates $S1$ (I is shorthand for "Intruder") from $K1$ and $K2$, $p_{S1\rightarrow S3}(K6,K7)$ denotes the function for calculating the conditional probability that $S1$ leads to $S3$ from $K6$ and $K7$, and $f_{S3}(K8)$ denotes function for calculating the frequency of $S3$ from $K8$.

When defining the functions for the vertices and relations where indicators have been attached, we may get some guidance from the values of the indicators at the time when the initial analysis was performed, as the functions should give likelihood (and consequence) values in the intervals obtained in the initial analysis when applied on these values. Table 3 shows the indicator values that we assume apply at the time of the initial analysis in our example. Exactly how to define the functions depends on each particular case, and must be based on the expertise and judgment of the analysis team, as well as existing statistical data if available. We now describe how it might be done for the example. Note that the definitions below have been made up in order to illustrate the approach.

Table 3 Key indicators with values at the time of the initial analysis.

Name	Description	Domain	Unit	Value
$K1$	Time interval for opening of doors.	$\{2,\ldots,30\}$	seconds	16
$K2$	Total time that keycard locks have been defect during the last 3 months.	$\{0,\ldots,2160\}$	hours	24
$K3$	Number of inactivity logoff occurrences during the last 3 months.	$\{0,\ldots\}$	-	21
$K4$	Inactivity logoff delay interval.	$\{5,\ldots,30\}$	minutes	8
$K5$	Percentage of patient records with errors detected in the last periodic sample review.	$[0,1]$	-	0.01
$K6$	Number of passwords that do not fulfill password strength policy.	$\{0,\ldots,200\}$	-	40
$K7$	Number of passwords that have not been changed in the last 3 months.	$\{0,\ldots,200\}$	-	35
$K8$	Number of unrecognized log-ons reported by employees the last 3 months.	$\{0,\ldots,200\}$	-	6
$K9$	Number of patient records accessible for an average user.	$\{0,\ldots,2500\}$	-	224

We start with the function $f_{I\rightarrow S1}(K1,K2)$. From Figure 4 we see that an initial estimate has been made for the current value of this function. The analysts therefore note that the function should, when applied to the above values for $K1$ and $K2$, give a frequency value "very often", which corresponds to more than 100 times per year.

Based on existing data about physical access control, their own experience in the field, and knowledge about the hospital in question, the analysts expect that if a door is unlocked (due to a defect keycard lock), there will be on average one intruder every day, or 365 intruders per year. As three months equals 2160 hours, and $K2$ gives the number of hours a keycard lock is defect during a three month period, the contribution to $f_{I\rightarrow S1}(K1,K2)$ due to defect keycard locks is $\frac{K2}{2160}\times 365$. Furthermore, the analysts expect that as long as the keycard locks function properly, the number of intruders per year will be proportional to the opening interval of doors. In the worst case, where doors remain open for 30 seconds after being opened, the analysts consider that it will be almost as easy to gain access as when a keycard lock is broken, as there is a lot of traffic in and out of the hospital; they therefore expects 300 intruders per year in this case. In the best case, where doors only remain open for 2 seconds after being opened, it will be much harder to gain access by following after a hospital employee. In this case the analysts expect 20 intruders per year. This gives the contribution $(1-\frac{K2}{2160})\times 10\times K1$ for the time periods when no keycard locks are defect. All in all, the above considerations give the following function:

$$f_{I\rightarrow S1}(K1,K2)=\frac{K2}{2160}\times 365+(1-\frac{K2}{2160})\times 10\times K1 \tag{1}$$

Applying this function on the arguments $K1=16$ and $K2=24$ gives a value of 162 per year, which is indeed in accordance with the initial estimate.

For defining the function $p_{S1\rightarrow S2}(K3,K4)$, the analysts assume that the value will be at least 0.01, and at most 0.95. Within these limits, the probability is expected

to be proportional with the product of $K3$ and $K4$ and be given by $\frac{K3 \times K4}{x}$ for some suitable x, as the product of $K3$ and $K4$ says something about the length of time during the observation period there will be terminals where the user have forgotten to log off. Clearly, the value of x has to be set so that a suitable probability is obtained. The analysts decide to estimate this number from the case where $K3 = 300$ and $K4 = 10$. In this case, the value of $p_{S1 \to S2}(K3, K4)$ is expected to be 0.5. Hence, the value of x is given by $\frac{300 \times 10}{x} = 0.5$, which gives $x = 6000$. The analysts confirm that this is a suitable value for x by inserting alternative values for $K3$ and $K4$, in each case verifying that the resulting probability is within their expectations. These considerations give the following definition of $p_{S1 \to S2}(K3, K4)$:

$$p_{S1 \to S2}(K3, K4) = \begin{cases} 0.01 & \text{if } \frac{K3 \times K4}{6000} \leq 0.01 \\ 0.95 & \text{if } \frac{K3 \times K4}{6000} \geq 0.95 \\ \frac{K3 \times K4}{6000} & \text{otherwise} \end{cases} \tag{2}$$

For defining the function $p_{S1 \to S3}(K6, K7)$ for calculating the conditional probability that $S1$ leads to $S3$ from $K6$ and $K7$, the analysts assume that the value will be at least 0.01, which is the case where all users change their password every third month and follow the password strength policy (i.e. when $K6 = K7 = 0$), and at most 0.7, which is the case when none of the users do this (i.e. when $K6 = K7 = 200$). The analysts consider an old password to be just as bad as a weak password (and a password that is both old and weak to be twice as bad), and expects that the probability will depend linearly on the sum of $K6$ and $K7$ between these limits. This gives the following definition:

$$p_{S1 \to S3}(K6, K7) = \frac{0.69 \times (K6 + K7)}{400} + 0,01 \tag{3}$$

Note that, according to the above definitions and the possible values of the indicators, the probability that $S1$ leads to $S2$ and the probability that $S1$ leads to $S3$ may sum up to more than 1. The reason for this is that we assume that an intruder may initiate both $S2$ and $S3$, so that they are not mutually exclusive. For example, when gaining access to a terminal where the previous user has forgotten to log off, the intruder may decide to try to log on as a different user in order to gain access to patient records that are not accessible for the previous user. In general, CORAS diagrams do not require that the sum of probabilities for the outgoing relations from a vertex should add up to 1 or less, as one scenario may lead to a number of different scenarios or unwanted incidents simultaneously.

For calculating the frequency of $S3$ from $K8$, the analysis team assumes that the number of unrecognized log-ons reported by employees (and measured by $K8$) reflects the actual situation reasonably well. Therefore, they decide to simply use the value of $K8$ multiplied with 4 to obtain the number of occurrences per year, rather than per 3 months.

$$f_{S3}(K8) = K8 \times 4 \tag{4}$$

Inserting the current value of $K8 = 6$, this gives a frequency of 24 times per year, which is in accordance with the initial estimate "Often", i.e. from 21 to 100 times per 10 years.

For calculating the frequency of $U1$ from $K5$, the analysts decide to set a minimum value of 2 per ten years, independently of $K5$. This is done in order to avoid a situation where the risk value associated with $U1$ becomes 0 in the cases where the frequency is so low that it might not be captured by the periodic sample review. Above this minimum value, the frequency is assumed to be proportional with $K5$. In order to obtain a frequency from the probability $K5$, the analysts reason as follows: There are 2500 patient records in all, so the number of records with errors are $2500 \times K5$. The records have an average age of 5 years, and errors are assumed to have been introduced during the last 5 years, which means that the frequency of error introduction is $\frac{2500 \times K5}{5}$ per year. These considerations give the following function:

$$f_{U1}(K5) = \begin{cases} 2 & \text{if } K5 \times 500 < 2 \\ K5 \times 500 & \text{otherwise} \end{cases} \tag{5}$$

Inserting the current value of $K5 = 0.01$, this gives a frequency of 5 times per year, which is in accordance with the initial estimate "Seldom", i.e. from 3 to 5 times per year.

The last place in the diagram where an indicator has been identified is the "Impacts" relation from $U1$ to the asset "Integrity of patient data". It therefore remains to define a function that calculates the consequence of $U1$ from the relevant indicator $K9$. For an incident where data is corrupted, the consequence may clearly depend on the nature of the corruption. The intruder may, for example, add false information or delete some or all fields of one or more records. However, it was decided in the initial analysis that consequence should be measured simply in the number of records affected. The analysts decide to assume that the intruder, when corrupting patient data, manages to corrupt all the records available for the user in question. As $K9$ measures the number of patient records available for an average user, they therefore decide to define the function simply as follows:

$$c_{U1}(K9) = K9 \tag{6}$$

7 Evaluating consistency

Obtaining a correct threat diagram with suitable indicators and functions for calculating likelihood values from indicators will clearly be quite challenging. As illustrated by the previous section, the subjective judgments made by experts and analysts will typically play a major role. It is therefore important to have ways of discovering weaknesses and aspects that need to be reconsidered. Being able to automatically check whether the values obtained are consistent is therefore important. By consistent we mean that the likelihood value assigned to a vertex should match

the value that can be obtained from the likelihood values of the preceding vertices and incoming "leads-to" relations. For example, the frequency of $S3$ should match the frequency we obtain from the frequency of $S1$ and the probability assigned to the "leads-to" relation from $S1$ and $S3$. The CORAS calculus introduced in [BDS08] provides rules that can be employed to calculate the likelihood of a vertex indirectly in this way. This allows us to check the consistency of likelihood values for the vertices where we have one value calculated directly from the indicators attached to the vertex and another value calculated indirectly from preceding vertices and "leads-to" relations. In our example this is the case for $S3$ and $U1$.

Before showing how the consistency for these two vertices can be analyzed, we now explain some of the most important rules of the CORAS calculus. For a more comprehensive presentation, see [BDS08]. In the following we use t to denote a threat, while v, v_1, v_2 denote vertices that may be threat scenarios or unwanted incidents. We use $t \xrightarrow{f} v$ to denote that t initiates v with frequency f, while $v_1 \xrightarrow{p} v_2$ denotes that v_1 leads to v_2 with conditional probability p. $v(f)$ denotes that vertex v occurs with frequency f.

The "initiate" rule captures the semantics of the "initiate" relation. The frequency of the occurrences of vertex v due to threat t is equal to the frequency with which t initiates v. This is captured by the following rule, where $t \sqcap v$ can be understood as "the subset of the scenarios/incidents v initiated by threat t".

Rule 1 (Initiate) *For threat t and vertice v related by "initiate", we have:*

$$\frac{t \xrightarrow{f} v}{(t \sqcap v)(f)}$$

The "leads-to" rule captures the conditional probability semantics embedded in the "leads-to" relation. The likelihood of the occurrences of v_2 that are due to v_1 is equal to the frequency of v_1 multiplied with the conditional probability that v_1 will lead to v_2 given that v_1 occurs. This is captured by the following rule, where $v_1 \sqcap v_2$ can be understood as "the subset of the scenarios/incidents v_2 that result from v_1".

Rule 2 (Leads-to) *For the vertices v_1 and v_2 related by "leads-to", we have:*

$$\frac{v_1(f) \quad v_1 \xrightarrow{p} v_2}{(v_1 \sqcap v_2)(f \times p)}$$

If two vertices are mutually exclusive the likelihood of their union is equal to the sum of their likelihoods. This is captured by the following rule, where $v_1 \sqcup v_2$ denotes all instances of the scenarios/incidents v_1 and v_2.

Rule 3 (Mutually exclusive vertices) *If the vertices v_1 and v_2 are mutually exclusive, we have:*

$$\frac{v_1(f_1) \quad v_2(f_2)}{(v_1 \sqcup v_2)(f_1 + f_2)}$$

Finally, if two vertices are statistically independent, the likelihood of their union is equal to the sum of their individual likelihoods minus the likelihood of their intersection.

Rule 4 (Independent vertices) *If the vertices v_1 and v_2 are statistically independent, we have:*

$$\frac{v_1(f_1) \quad v_2(f_2)}{(v_1 \sqcup v_2)(f_1 + f_2 - f_1 \times f_2)}$$

To illustrate the use of the consistency rules, we now assume that some time has passed after the initial analysis, and that the indicator values have changed to the following values: $K1 = 17, K2 = 8, K3 = 23, K4 = 8, K5 = 0.03, K6 = 61, K7 = 72, K8 = 5, K9 = 230$. Note that the initial likelihood and consequence estimates assigned in Figure 4 are outdated at this point, as they applied at the time of the initial analysis.

In order to simplify the presentation, in the following we set $p_{S2 \to U1} = p_{S3 \to U1} = 0.15$. As no indicators were identified for these probabilities, the value will not change. Furthermore, we assume that the diagram in Figure 4 is meant to be complete, in the sense that there are no other threats or threat scenarios that may initiate or lead to any of the described threat scenarios or unwanted incidents. This means that all instances of $S1$ are initiated by the threat Intruder (denoted by I), and hence that $S1 = I \sqcap S1$. Furthermore, it means that all occurrences of $S2$ and $S3$ are due to $S1$ (i.e. that $S2 = S1 \sqcap S2$ and $S3 = S1 \sqcap S3$), and that all occurrences of $U1$ are due to $S2$ or $S3$, i.e. that $U1 = (S2 \sqcap U1) \sqcup (S3 \sqcap U1)$. This completeness assumption allows us to view the likelihood estimates obtained through use of the above rules as actual values, rather than lower limits.

We first look at the consistency of the frequency of $S3$. This value can be obtained either indirectly from the preceding vertices and relations by application of the above rules, or directly from (4). With the new value for $K8$, the latter approach gives the frequency 20 per year for $S3$. Taking the indirect approach, we start by calculating the frequency with which the intruder initiates $S1$. The new values for $K1$ and $K2$ gives $f_{I \to S1}(K1, K2) = 171$. We then apply Rule 1 to obtain the frequency 171 per year for $S1$. Now we want to apply Rule 2 to calculate the frequency of $S3$ from the frequency of $S1$. First we calculate $p_{S1 \to S3}(K6, K7) = 0.24$ from the new indicator values. Rule 2 then gives us the frequency $171 \times 0.24 = 41$ per year for $S3$. Hence, we have a difference of 21 per year for the two estimates of frequency for $S3$. For simplicity, in the following calculations we will use $S3 = 41$ rather than $S3 = 20$, but we could also have tried both values in order to see which gives the highest degree of consistency.

Next, we want to check the consistency of frequency estimates for $U1$. Again, taking the direct approach is easy; applying (5) on the new value of $K9$, we get the frequency 15 per year for $U1$. For the indirect approach, we note that both $S2$ and $S3$ may lead to $U1$ according to Figure 4. However, as $S2$ and $S3$ are not considered to be mutually exclusive or statistically independent, we cannot obtain an exact value from Rule 3 or Rule 4. We are therefore confined to calculating maximum and minimum values.

Clearly, the minimum frequency of $U1$ cannot be lower than the highest of the frequencies we obtain from coming to $U1$ from one of $S2$ or $S3$, i.e. the frequency of either $S2 \sqcap U1$ or $S3 \sqcap U1$. To find the former we first need to calculate the frequency of S_2. Applying Rule 2, we obtain this value by calculating the frequency of $S1$ with the probability $p_{S1 \rightarrow S2}(K3, K4)$ that $S1$ leads to $S2$, which with the new indicator values gives the frequency $171 \times 0.03 = 5$ times per year for $S2$. Applying Rule 2 on the "leads-to" relation from $S2$ to $U1$ then gives a frequency of ca 1 per year for $S2 \sqcap U1$. Similarly, we use Rule 2 to obtain a frequency of ca 6 per year for $S3 \sqcap U1$. Hence, according to these calculations the minimum frequency of $U1$ is 6 times per year.

For the maximum frequency of $U1$, we use Rule 3 as if $S2$ and $S3$ were mutually exclusive. Adding up the frequencies of $S2 \sqcap U1$ and $S3 \sqcap U1$ we thus obtain $1 + 6 = 7$ times per year. Hence, according the indirect calculations, the frequency of $U1$ should be between 1 and 7 times per year, which is lower than what was obtained through the direct calculation.

The kind of calculations and comparisons demonstrated here can be performed automatically as the indicator values change. This can be utilized to give a warning in cases where the risk picture is inconsistent. In the above example we saw that a certain discrepancy was detected for both vertices that were checked. In practice it is hardly realistic to expect the values to coincide exactly. It is up to the analysts to decide how much two values must differ in order to count as inconsistent, and what should be the exact criteria for triggering a warning.

8 Measuring confidence

The purpose of the Confidence module is to offer a quantitative measure of confidence in the overall risk picture based on the degree of inconsistency that has been detected. There are a number of ways in which notions of confidence may be estimated. The measure could, for example, be based on the number of nodes where inconsistent likelihood estimates are assigned, or on the average difference between conflicting estimates of the same likelihood, or on some more sophisticated statistical analysis. Furthermore, the change of indicator values over time could be considered. If the risk picture has remained consistent over a period of time where indicator values have changed, this gives greater reason to believe in the correctness of the analysis, and in particular the correctness of the functions from indicators to likelihood values, is correct than if consistency has only been observed with one set of indicator values.

Clearly, the degree to which the analysis actually allows consistency to be checked should also be taken into account. Values that are not checked for consistency should count as neither consistent nor inconsistent. In our example above, we are able to check the consistency of likelihood values for $S3$ and $U1$, but not for $S1$ and $S2$. Therefore, $S1$ and $S2$ should not contribute to a high confidence value, even if no inconsistency is detected for these vertices.

In cases where two alternative and inconsistent likelihood estimates are obtained for a vertex, the measure of confidence can be employed as an aid to help decide which value is most likely to be correct. This can be done by checking which value gives the highest confidence value. In the example above we obtain two different values for the likelihood of $S3$ from the indirect and the direct calculations. When calculating the likelihood of $U1$ indirectly from the likelihood of $S2$ and $S3$, we therefore need to decide which value to use for $S3$. Clearly, if we choose a wrong likelihood value for $S3$ we may also get a wrong likelihood value for $U1$, possibly resulting in inconsistent estimates also for $U1$, and thus a lower overall confidence value. We should therefore choose the value that gives the highest confidence value, unless we have other reasons to believe that this value is wrong. After deciding which of two conflicting values is assumed to be most correct, the next step will then be to make the necessary corrections in order to bring the assumed wrong value in line with the correct one, for example by redefining a function for calculating likelihood from key indicators.

A suitable definition of confidence value based on the degree of inconsistency may serve not only as a measure of the assumed correctness of the risk picture, but also as an aid in improving the analysis. How to find the most suitable ways of measuring the degree of inconsistency and the confidence value is a interesting research question in its own right, that we will not pursue further in this paper.

9 Related work

For demonstrating the approach presented in this paper, we have chosen to use CORAS for threat modeling and assignment of quantitative likelihood, consequence and risk levels. However, the approach is generic in the sense that other languages and modeling techniques may also be employed. As we have seen, a suitable language needs to be flexible with respect to annotations of likelihood values and be able to capture inconsistent likelihood estimates, as this allows us to uncover weaknesses of the subjective estimates made by the analysts. We now present some related work, with a particular view on these aspects.

Fault tree analysis (FTA) [IEC90] and related techniques like attack tree analysis [Sch99, MO05] are often used to obtain the likelihood of an unwanted incident in the context of risk analysis. In fault tree analysis, the top vertex represents an event/fault that is decomposed into intermediate and leaf vertices by the use of logical operators. The likelihood of the top vertex is calculated from the likelihood of the leaf vertices, which are assumed to be independent, as well as the operators used to compose vertices. Attack trees are much like fault trees, but focus on the attacks that a system may be exposed to. Moreover, attack trees allow also other values than probability to be assigned to the vertices, for example the cost of an attack, or qualitative statements such as “possible” or “impossible”. As values are assigned only to the leaf vertices by the analysts, there is no possibility of assigning inconsistent values for fault trees or attack trees. In addition, likelihood values are only assigned

to vertices, and not to the relations between vertices. From a methodological perspective this means that we cannot define a probability function for a relation from the indicators identified for that relation independently from the related vertices.

Bayesian networks [BG07] are directed acyclic graphs that may be used to represent knowledge and to make quantitative assessments about an uncertain domain, and can therefore be employed in risk assessment. Nodes represent random variables, while edges between nodes represent probabilistic dependencies between variables. A Bayesian network can be used to compute the joint probability distribution over a set of random variables. Like fault trees and attack trees, probabilities are not assigned to the edges of Bayesian networks. Instead, each node is decorated with parameters that for each node give its conditional probabilities, where the conditions represent the state of its parent nodes. The underlying mathematical model of Bayesian networks is more complicated than that of CORAS diagrams, but also more powerful. Rather than capturing inconsistent likelihood estimates, use of Bayesian networks usually focus on updating likelihood values based on new evidence. Fenton et al [FKN02, FN04] uses Bayesian networks to address the problem of quantifying likelihood based on different types of evidence, and demonstrate how their approach can be applied to assess the frequency of defects in software or components.

Phillips and Swiler [PS98] present a method for risk analysis of computer networks based on attack graphs. An attack graph is not produced directly by the analysis team, but generated automatically from configuration files containing information about the network, attacker profiles containing information about the assumed attacker's capabilities, and attack templates containing information about known generic attacks. Nodes in the attack graph represent attack states (effects of the attack so far), while edges represent a change of state caused by the attacker. Each edge is assigned a weight estimate representing a success probability or some other measure, such as average time to succeed or effort level for the attacker. Multiple weights may be assigned to edges. However, these are not intended to capture inconsistent estimates. Instead they represent potentially conflicting criteria, for example that the attacker wishes to minimize both cost and probability of detection. From the attack trees various kinds of analysis may be performed, such as finding a set of low-cost attack paths or cost-effective defenses. The approach presented in [PS98] does not address the question of inconsistency or confidence in the analysis. It is also less generic than the one we propose, as it is specially tailored to analysis of computer networks, with no emphasis on human behavior.

Closely related to what we call key indicators, the field of IT Security metrics provides an approach to measuring information security. The NIST Performance Measurement Guide for Information Security [CSS+08] aims to assist in the development, selection, and implementation of suitable measures to this end. It also provides a number of candidate measures, for example "Percentage of information system security personnel that have received security training" or "Percentage of individuals screened before being granted access to organizational information and information systems". Such measures are suitable candidates for key indicators. Unlike the work we have presented, the approach taken in [CSS+08] does not neces-

sarily aim to establish explicit frequency, consequence, and risk levels from the identified set of measures.

An interesting approach to the uncertainty involved in risk analysis based on subjective estimates is taken in [JBK04], which explains the use of subjective logic in risk analysis. Subjective logic [Jøs07] is a probabilistic logic that explicitly takes uncertainty about probability values into account. For example, it is possible to calculate to what degree an actor believes a system will work based on the actor's belief about the subsystems. In [JBK04], subjective beliefs and uncertainty about threats and vulnerabilities are used as input parameters to the analysis, allowing the uncertainty associated with the result of the analysis to be explicitly represented.

10 Conclusion

We have presented a vision and an approach for risk monitoring where risk values are calculated from measurable key indicators. The resulting risk picture is dynamic in the sense that risk values are automatically updated as soon as the indicators change. This means that we get a risk picture that remains valid over a period of time rather than representing a snapshot. Moreover, it allows us to consider not only the actual risk levels at a given point in time, but also to analyze trends. For example, if a risk level has been steadily increasing over time, this might suggest that mitigating measures should be considered even if the current risk level is lower than the acceptance threshold. We have demonstrated the approach on the CORAS method, but the same ideas can be used for other risk modeling languages. We claim however that CORAS is particularly suitable due to its flexibility with respect to likelihood annotations.

The approach allows the internal consistency of the risk picture to be assessed in order to reveal weaknesses and issues that need to be reconsidered. A notion of confidence calculated from the degree of inconsistency found in a risk picture has also been proposed in order to assess to what degree the risk picture may be assumed to be correct. This is important because subjective judgment and estimates play a major role in the approach. The aim is not to eliminate the need for such subjective judgment, which would be unrealistic, but to provide support for making the judgment and evaluating its result. Clearly, defining functions from key indicators to likelihood and consequence values based on the subjective judgment of experts will be a major challenge. As noted in [Vos08], eliciting from expert opinion has a number of pitfalls and requires great care, but there are techniques for avoiding the pitfalls. Providing tailored guidelines and methods for defining the necessary functions from key indicators to likelihood and consequence values is an interesting topic for further research.

When making decisions that depend on risks, having a list of potential risks is not enough. We also need to understand how high the risks are. The work presented here is a step towards the goal of obtaining such an understanding.

Acknowledgements The research leading to these results has been funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement FP7-216917.

References

[BDS08] Gyrd Brændeland, Heidi Dahl, and Ketil Stølen. A modular approach to the modelling and analysis of risk scenarios with mutual dependencies. Technical Report A8360, SINTEF ICT, 2008.

[BG07] Irad Ben-Gal. Bayesian networks. In Fabrizio Ruggeri, Ron S. Kenett, and Frederick W. Faltin, editors, *Encyclopedia of Statistics in Quality and Reliability*. John Wiley & Sons, 2007.

[CSS+08] Elizabeth Chew, Marianne Swanson, Kevin Stine, Nadya Bartol, Anthony Brown, and Will Robinson. Performance measurement guide for information security (NIST Special Publication 800-55 revision 1). Technical report, National Institute of Standards and Technology, 2008.

[dBHL+07] F. den Braber, I. Hogganvik, M. S. Lund, K. Stølen, and F. Vraalsen. Model-based security analysis in seven steps — a guided tour to the CORAS method. *BT Technology Journal*, 25(1):101–117, 2007.

[FKN02] Norman Fenton, Paul Krause, and Martin Neil. Software measurement: uncertainty and causal modeling. *Software, IEEE*, 19(4):116–122, Jul/Aug 2002.

[FN04] Norman Fenton and Martin Neil. Combining evidence in risk analysis using bayesian networks. Agena White Paper W0704/01, 2004.

[IEC90] IEC. *IEC 61025, Fault Tree Analysis*, 1990.

[JBK04] Audun Jøsang, Daniel Bradley, and Svein J. Knapskog. Belief-based risk analysis. In *Proceedings of the Australasian Information Security Workshop (AISW)*, volume 32 of *Conferences in Research and Practice in Information Technology (CRPIT)*, pages 63–68. Australian Computer Society, 2004.

[Jøs07] Audun Jøsang. Probabilistic logic under uncertainty. In *Proceedings of Thirteenth Computing: The Australasian Theory Symposium (CATS)*, volume 65 of *Conferences in Research and Practice in Information Technology (CRPIT)*, pages 101–110. Australian Computer Society, 2007.

[MO05] Sjouke Mauw and Martijn Oostdijk. Foundations of attack trees. In Dongho Won and Seungjoo Kim, editors, *ICISC*, volume 3935 of *Lecture Notes in Computer Science*, pages 186–198. Springer, 2005.

[PS98] Cynthia Phillips and Laura Painton Swiler. A graph-based system for network-vulnerability analysis. In *Proceedings of the 1998 workshop on new security paradigms*, pages 71–79. ACM, 1998.

[Sch99] Bruce Schneier. Attack trees: Modeling security threats. *Dr. Dobbs Journal*, 24(12):21–29, 1999.

[Vos08] David Vose. *Risk Analysis. A quantitative guide*. John Wiley & Sons, third edition, 2008.