A Move in the Security Measurement Stalemate: Elo-Style Ratings to Quantify Vulnerability

Wolter Pieters
Delft University of Technology
Technology, Policy and Management
Infrastructure Systems & Services
P.O. Box 5015
NL-2600 GA Delft
The Netherlands
w.pieters@tudelft.nl

Sanne H.G. van der Ven
University of Amsterdam
Social and Behavioural Sciences
Psychological Methods
Weesperplein 4
NL-1018 XA Amsterdam
The Netherlands
s.h.g.vanderven@uva.nl

Christian W. Probst
Technical University of Denmark
Informatics and Mathematical Modelling
Richard Petersens Plads
DK-2800 Kongens Lyngby
Denmark
probst@imm.dtu.dk

ABSTRACT

One of the big problems of risk assessment in information security is the quantification of risk-related properties, such as vulnerability. Vulnerability expresses the likelihood that a threat agent acting against an asset will cause impact, for example, the likelihood that an attacker will be able to crack a password or break into a system. This likelihood depends on the capabilities of the threat agent and the strength of the controls in place. In this paper, we provide a framework for estimating these three variables based on the Elo rating used for chess players. This framework re-interprets security from the field of Item Response Theory. By observing the success of threat agents against assets, one can rate the strength of threats and controls, and predict the vulnerability of systems to particular threats. The application of Item Response Theory to the field of risk is new, but analogous to its application to children solving math problems. It provides an innovative and sound way to quantify vulnerability in models of (information) security.

Categories and Subject Descriptors

K.6.5 [**Management of Computing and Information Systems**]: Security and Protection

General Terms

Management, Measurement, Security

Keywords

control strength, Elo, Item Response Theory, rating systems, risk assessment, security metrics, threat capability, vulnerability

NSPW'12, September 18–21, 2012, Bertinoro, Italy.

1. INTRODUCTION

1.1 Problem Statement

One of the big problems of risk assessment in information security is the quantification of risk-related properties, such as the expected frequency of threat events, vulnerability of a system to such events, and impact on the system. In this research area, the goal is not finding new threats or attack scenarios (which is another big problem), but rather quantifying the risk of known ones. Simply counting incident numbers is not very adequate though, especially if they are self-reported [12]. Therefore, we need more detailed models of what constitutes a risk and how much damage to expect from it [22].

Among the different components of risk, the notion of vulnerability is particularly problematic. There are at least three issues with the use of the term:

1. The term is used in both countable and uncountable forms, to express two different things:
   (a) A particular security weakness in a system design: "This software has two buffer overflow vulnerabilities", or
   (b) The ease with which a system can be damaged by threats: "This system has a higher vulnerability than that one";
2. If we assume the second meaning, it is unclear how to define the vulnerability (level) of a system to a particular threat;
3. Even if we know the vulnerability levels of a system to specific threats, it is unclear how to define an aggregate vulnerability level of a system, to provide justification for the statement that one system is more vulnerable than another.

In this paper, we deal with the second problem, namely the definition of the level of vulnerability of a system with respect to a particular threat. This means that we will use vulnerability in the uncountable sense, i.e., the ease with which damage is caused by a threat. The solution we envision is based on the psychometric approach of Item Response Theory, related to the widely known Elo ratings used to rank chess players.

1.2 Ingredients

To define our concepts, we use the risk definitions provided by The Open Group [15]. They assume the existence of "threat agents" in the environment of a system that may cause damage to system assets. According to The Open Group, vulnerability expresses the likelihood that a threat agent acting against an asset will cause impact. Vulnerability is based on the "skills" of the threat (Threat Capability) and the "skills" of the defender (Control Strength).

Thus, in a non-malicious context, a storm with a certain wind speed will have a certain likelihood of causing a power line breakdown. This likelihood will increase with the wind speed. On the other hand, different power lines will have different resistance against storms, and the likelihood of breakdown will decrease with the resistance.

There are three variables involved here: wind speed, resistance, and breakdown. For each particular case of the threat agent storm acting against the asset power line, we may be able to obtain certain information. In particular, we may know the wind speed, the type of power line, and/or whether the "attack" results in a breakdown or not. For malicious threats, we may know the attacker capability, the type of system, and/or the "success" of an attack. Given this information, we would like to predict the likelihood of breakdown for future events (i.e., the vulnerability). How to do this?

In a very similar setting, we have the following problem. We have a population of children who attempt so solve instances from a population of math problems. Initially we may know neither the capability of the children nor the difficulty of the problems. Can we, based on observing which children are able to solve which problems, determine the children's capabilities, and the difficulty of the problems? If we can, then we can also predict the success of a particular child in solving a particular (previously unencountered) problem.

Klinkenberg, Straatemeier, and Van der Maas [19] developed just such a system for math problems, called Math Garden, which is based on the Elo rating. The Elo rating was originally conceived to rank players in chess [8], but is now also regularly applied outside the chess domain [9, 10]. Each player has a rating that can be used to estimate the expected outcome of a particular match between two players. This rating is updated after each match: players' ratings increase if they win, and decrease if they lose. The rating increases more if they win against a highly ranked player. When applied to math problems, Elo works analogously, although we now have two different types of entities: children and problems. If a child solves a problem, the rating of the child increases and that of the problem decreases. If a child fails, the rating of the child decreases and that of the problem increases. This particular setting is covered by the field of Item Response Theory.

1.3 The New Paradigm

In this paper, we investigate how a similar paradigm might work in a security context, where threat agents "play" against systems. We thus redefine vulnerability metrics as a special application area of Item Response Theory. We study to what extent a rating system can be used for estimating (a) capability of a threat, (b) strength of a control (resistance), and (c) vulnerability, i.e., the likelihood that a particular threat is successful against a particular asset. This provides a method for estimating vulnerability, as a component of quantification of risk. Besides, the Elo approach has the advantage that it also yields ratings of the different threat and control types, which will give information about the vulnerability of other assets to the same threat, or the vulnerability of the same asset to other threats.

1.4 Paper Outline

In Section 2, we discuss related work, including the use of Elo style ratings as well as information security models. In Section 3, we provide the definitions of our concepts, based on the Risk Taxonomy of The Open Group [15]. In Section 4, we define the relation between threat capability, control strength, and vulnerability. In Section 5, we discuss how to use this model to estimate relevant variables in a risk context. In Section 6, we investigate the necessary infrastructure to support such applications. We end with open questions in Section 7 and conclusions in Section 8.

2. RELATED WORK

2.1 Rating Systems

To enable the quantification of vulnerability, threat capability, and control strength, we make use of existing rating systems. Several existing rating systems enable rating of the capabilities of entities, and we will discuss their properties in this section. Throughout the section, we use θ to represent a rating where both entities have the same type, e.g., the Elo system for chess players. When the entities have two (different) types, we use δ for the difficulty of the problem or asset and β for the ability of the solver or the attacker. This may not match the notation in the original papers.

Furthermore, the outcome of a "match" between two entities is 1 for the entity that "wins", 0 for the entity that "loses", and 0.5 for each in case of a "draw". Other possible outcome scales will be discussed later.

A central assumption of the rating systems discussed is the following: if the ratings of two entities are equal, then the expected result is 0.5 for each of them. If there is no draw option (the outcome is always 0 or 1), then we can state equivalently that the likelihood of winning is 0.5, when the ratings of the entities are equal.

2.1.1 One Type, Dynamic Rating: The Elo System

The Elo Rating System [8] is used for rating players in chess. The system consists of two independent sets of equations. The first set of equations is the update rule: after a match between two players, the ratings of these players are updated, based on the discrepancy between the expected outcome and the observed outcome of the match.

The new ratings $\hat{\theta}$ of two players i, j with ratings θ after a match with result S for each player (0 for lose, 0.5 for draw and 1 for win) are calculated as follows:

$$\hat{\theta}_i = \theta_i + K \cdot (S_i - E(S_i)) \tag{1}$$

$$\hat{\theta}_j = \theta_j + K \cdot (S_j - E(S_j)) \tag{2}$$

In these equations, the factor K reflects uncertainty, or how quickly the rating should change, based on the results of a single match. New players enter the system with a default rating, and if K is too low, it takes a long time before they reach their true rating. On the other hand, if

K is too high, the system is unstable: ratings fluctuate too much. The optimal value for K is found empirically, which can be done in different ways. Also, new players need to be assigned provisional ratings. In chess, K decreases with the number of matches played and the strength of a player.

The *expected* outcome for player i in a match against player j can be expressed as:

$$E(S_i) = \frac{1}{1 + 10^{(\theta_j - \theta_i)/\mathcal{S}}} \tag{3}$$

The factor $\mathcal{S}$ determines the scale of the rating and is set to 400 in chess. Also note that equations (1) and (2), the updating equations, are independent of equation (3), the expected results equation. This means that if desired, the expected match result can also be obtained in a different way while maintaining the rating updating system. This is done in Math Garden, as also shown later in this paper [19].

2.1.2 Two Types, Static Rating: Rasch Analysis

In the original Elo Rating System, we have one type of entity, namely chess players. In other settings, often two types of entities play a role, for example when persons solve problems. In such a setting, one would like to establish both the ability rating of the persons and the difficulty rating of the problems. This is the field of Item Response Theory (IRT), often used for the construction of aptitude tests.

The probability of the participant producing a correct answer to a problem depends jointly on the ability of the participant and the difficulty of the problem. This probability is described by a logistic model. The simplest model is the Rasch model, or one parameter logistic model (1PL model) [32]. The mathematical form of the Rasch model is as follows, with δ_i being the item difficulty, β_j the person ability, and S_{ij} the success of person j against item i, where 1 means solved and 0 means failed:

$$P(S_{ij} = 1) = \frac{e^{\beta_j - \delta_i}}{1 + e^{\beta_j - \delta_i}} = \frac{1}{1 + e^{\delta_i - \beta_j}} \tag{4}$$

Note that the probability of obtaining a correct response is conceptually very similar to the expected outcome of a match in the Elo system. The curves produced by the functions are also quite similar, although the Rasch formula uses e rather than 10 as exponentiation base. This choice is rather arbitrary. For a detailed comparison of the Elo and Rasch systems, see [36].

IRT test construction starts with a calibration phrase. A large set of problems is administered to a large number of participants. This enables the simultaneous estimation of both the difficulty ratings of the problems in the set and the ability ratings of the participants. After the calibration phase, a selected subset of items in the desired difficulty range can be administered to a participant of unknown ability. The more items the participant solves, the more precisely their ability can be estimated. In the model, it is assumed that there is a latent trait, with corresponding quantitative scale, which is measured by the tests under consideration. This scale is thus not predefined, but it emerges as a result of fitting the Rasch model to the data.

An advantage of this approach is that two participants can solve different problems from the set, e.g., a young child solves only easy problems and an older child solves only more difficult problems, but nevertheless, their ability ratings are on the same scale. When test administration takes place on a computer, the items can even be selected based on previous answers, so the participant only receives items that are tailored to their ability. The fact that items must first be calibrated is a limitation of the IRT approach. Calibration is a costly, time-consuming procedure and it prevents items from being added or changed after calibration, as their difficulty rating will not be available or accurate in that case.

2.1.3 Two Types, Dynamic Rating: Math Garden

The Math Garden project [19] circumvents this problem by combining the IRT approach with the Elo updating rule. The Math Garden provides an online platform where children can solve math problems. In this system, both children and problems can be added on the fly. All children and problems that are new to the model obtain a preliminary first rating, based on superficial characteristics (age for children, rough estimation of the difficulty for the items). When child j solves problem i, the Rasch model is used to estimate the probability of a correct answer. After the problem is solved, the Elo updating rule is applied to adjust the rating of both the child and the problem: if the answer is correct the child "wins", and if the answer is incorrect the problem "wins". In this way, newly added items are calibrated "automatically" by the update rules.

For calculation of the expected outcome S_{ij} of child j solving problem i, the Rasch rather than the Elo formula is used (with e as exponentiation base). This is combined with the following Elo update rules (same notation as above):

$$\hat{\delta}_i = \delta_i + K_i \cdot (E(S_{ij}) - S_{ij}) \tag{5}$$

$$\hat{\beta}_j = \beta_j + K_j \cdot (S_{ij} - E(S_{ij})) \tag{6}$$

The factors K again reflect the uncertainty in the ratings, to adjust the update speed. New items or persons with few results should converge quickly, whereas a single unexpected outcome for items or persons with many results should not affect the rating too much. In Math Garden, K increases when a player consistently scores below or above the expected outcome.

Furthermore, for computing the final ranking the response time is included as a factor, leading to additional adaptations. For details, we refer to [19].

2.2 Security Risk Models

In this paper, we apply the Math Garden approach to vulnerability quantification in risk assessment. The final aim is to extend traditional risk analysis with quantitative frequency and likelihood values representing risk-related properties. Several models for risk assessment have been proposed, in which such quantification would be of value. We will not mention all, but focus on a few that are model-based and deal with a security context.

CORAS [24] provides risk assessment based on UML diagrams. SAVEly [3], Exasym [30], Portunes [7], MsAMS [27] and ANKH [29] are graph-based security models. In addition, Portunes and Exasym annotate the nodes with processes. These approaches can be used to generate attack trees [25, 34], representing possible (multi-step) attack paths in the system. Nodes in attack trees can be annotated with quantitative properties, but so far the models have addressed attack possibilities only qualitatively. Such trees can also be augmented with countermeasures, as is done in

attack-defence trees [20] and semantic threat graphs [13]. Also, argumentation-based approaches have been proposed to reason about security [33]. The explicit relation between threats and countermeasures in such models makes them ideally suited to include information on threat capability, control strength, and vulnerability, as defined in the present work. The results presented in this paper will contribute to making such models suitable for quantitative security risk analysis, by providing the likelihood of success of steps in the attacks.

In terms of vulnerability quantification, existing work includes the Common Vulnerability Scoring System (CVSS) [26], which quantifies the severity of vulnerabilities in the sense of particular software weaknesses. Also, quantification of the level of such vulnerabilities has been attempted from the economic angle, assigning a market value to specific vulnerabilities [1, 4]. Here, we are interested in the quantification of the vulnerability, in the uncountable sense, of a system as a whole, based on measures of the strength of adversary and control. Using adversary strength as a security metric has been discussed before in the valuable notion of "weakest successful adversary" [28]. This is an abstract measure of adversary strength, meant to quantify system security, rather than the concrete ratings we propose here.

Vulnerability measurements can be done by means of penetration testing, as we will discuss later. With respect to penetration testing for security assessment, also economics have been addressed, in the sense of optimal penetration testing strategies [5]. Furthermore, economic approaches have been applied to optimal patching strategies [17, 37]. Such economic considerations could inspire future work on strategies to employ the Elo-style ratings in practice, by relating data quality to costs, as well as by relating patching strategies to rating improvement.

2.3 Expert Risk Judgement

In security, the use of Rasch-type models has been proposed to make expert risk judgement more objective [11]. In this setup, different experts would rate the risk associated with different threats (e.g., in terms of frequency and impact), and from this analysis, both the "risk bias" of the experts (risk-taking or risk-averse) and the "objective risk" of the threats would be determined. This is indeed a very interesting idea, and it would merit more attention in the security community. Here, we are concerned with capabilities of threats, strengths of controls, and the associated vulnerability levels, so we are completely on the "objective" side of the above distinction. However, it could be valuable to combine both approaches in future work.

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.

Part of the inspiration for this research stems from the Risk Taxonomy of The Open Group [15]. We found this taxonomy particularly valuable, because it makes an explicit distinction between threat events and loss events, and associated frequencies. 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 what they call Loss Event Frequency (LEF) and Probable Loss Magnitude (PLM).[1] The former represents the expected number of loss events of a particular type per unit of time, 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 separated in 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.

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.

Now comes the interesting observation in relation to Rasch and Elo systems:

> *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. In the storm/power line example, TC would be a value related to wind speed, and CS would be a value related to the strength of a power line.

Unfortunately, the relation between TC, CS, and V as defined by The Open Group seems rather problematic. Firstly, although they do not specify it explicitly, they seem to regard TC and CS as percentages. But percentages of what? The standard seems to assume a probability distribution of TC over all the threat agents, but this is something different than representing the TC of a particular threat agent as a probability. One can say that there is a probability distribution of wind speed, with the most powerful winds only occurring rarely, but it does not make sense to associate a probability value with the wind speed of a particular storm, or even a "threat community" of storms.[2] Further-

[1]These variables are often called likelihood (probability) and impact, but (expected) frequency is indeed more accurate than likelihood. A likelihood or probability is always a number between 0 and 1. If one considers a specified time frame, say 10 years, one is not interested in the probability of occurrence of a certain event (e.g., 0.9), but rather in the expected number of occurrences (e.g., 4). Only the latter allows calculation of the expected damage.

[2]Unless one means a percentile or cumulative probability, stating that 90% of the storms would be weaker than this

more, they then seem to calculate V as TC - CS, with the example 90% − 80% = 10%. This could be interpreted as a threat agent with TC 90% having 10% probability of inflicting damage upon an asset with CS 80%. Even if percentages would be the right way to express TC and CS, this seems plainly wrong, as probabilities of events cannot be subtracted to obtain the probability of another event.

Fortunately, there is another way to represent TC, CS, and V, which yields a meaningful relation both in mathematical and in practical sense. The key observation is that the notions of TC and CS seem to fit extremely well in the Elo and Rasch type models, where – as discussed above – TC corresponds to a person's ability to solve problems, and CS corresponds to the difficulty of a problem. A threat agent acting against an asset thus corresponds to a child solving a math problem. As models already exist for the latter case, we can provide an accurate definition of the relation between TC, CS, and V by employing similar models.

4. EXPRESSING VULNERABILITY

Our goal is thus to represent threat capability, control strength, and vulnerability in a mathematical model, analogous to the Rasch and Elo approaches for children solving math problems. The difference between threat capability and control strength then determines the probability of success, i.e., the probability of the threat agent inflicting damage upon the asset.

Note that our approach has general applicability, as it is based on data (observed outcomes) only. A single variable per entity is assumed to explain the observed variation (i.e., threat capability / control strength). When underlying physical, psychic or social mechanisms are known, more precise models, with additional variables, may be developed for specific cases (e.g., wind and power lines). However, the goal here is a general model that can be used for any type of threat.

Typically, for a given threat capability, the probability of success will first decrease slowly with the control strength. Where the control strength is close to the threat capability, the curve will drop steeply, after which it will start to decrease slowly again. The inverse behaviour will be seen for the probability of success with a fixed control strength, with increasing threat capability. Here, there will be a sharp increase where threat capability and control strength are close. In the math problems case, children with low capability are very unlikely to solve difficult problems, but the probability increases sharply where capability and difficulty are almost equal.

This can be expressed in the logistic formula of the Rasch model (one parameter, or 1PL). The relation between threat capability β_j, control strength δ_i, and success S_{ij} can be expressed as follows (same as Equation 4):

$$P(S_{ij} = 1) = \frac{e^{\beta_j - \delta_i}}{1 + e^{\beta_j - \delta_i}} \tag{7}$$

Examples of the 1PL model are shown in Figure 1. The curves show the probability $P(S_{ij} = 1)$ corresponding to the capability β_j, where each curve corresponds to a different value of δ_i.

one. This would make more sense, but it does not help in expressing vulnerability as the difference between threat capability and control strength.

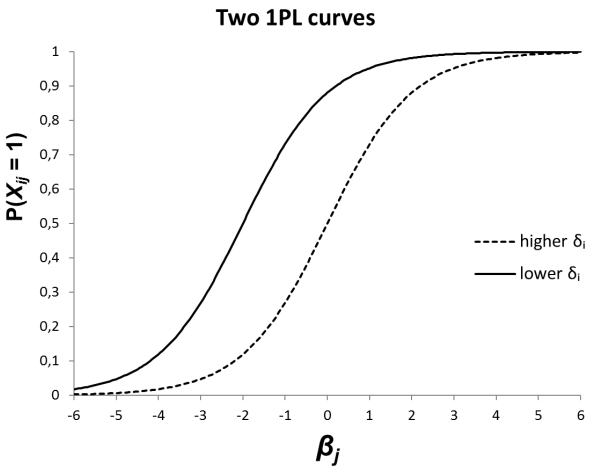

Figure 1: A one-parameter logistic function.

Instead of the one parameter logistic Rasch model (1PL) a two-parameter model (2PL) can also be used [2]. One can then express the discrimination of problems / controls, i.e., how well they can distinguish between persons / threats of different capability. When the success probability is displayed against the threat capability for a specific control, this corresponds to the maximum slope of the curve. With additional parameter α, indicating the slope of the curve, the formula becomes:

$$P(S_{ij} = 1) = \frac{e^{(\beta_j - \delta_i)\alpha_i}}{1 + e^{(\beta_j - \delta_i)\alpha_i}} \tag{8}$$

The 2PL model is more expressive, in the sense that it can adequately represent items / controls with different discriminatory value, but the fitting of the model to the data becomes more complex.

A third parameter can be added to represent bias in the probability, for example in case of multiple choice questions with guessing bias. In that case, the lower asymptote would become higher than zero. The issue of "luck" or guessing might be relevant in security, although it would be harder to define its exact value. Also, one might want to consider a bias on the upper side of the graph, when even the best threat agents would have some fixed probability of failing.

Typically, when using such models in a problem-solving context, one will observe the success of certain persons playing against certain problems, and thereby estimate the parameters of the problems (δ_i, and in case of the 2PL model α_i), and the ability of the persons (β_j). This will then allow the prediction of the success of a person on a problem she has not encountered before. It will also allow the estimation of the ability of an unknown person, based on any set of administered problems (once calibrated). For threats and controls, a similar analysis may be employed based on known incidents.

When we use dynamic ratings, like in the Elo chess rating and in the Math Garden approach, we can define update rules that we apply when events occur (see Equations (5) and (6)). New threats and controls can then be added to the system on the fly, without requiring a new calibration phase. As new security risks emerge quite often, this is an important feature in a risk management context. Again,

the update rate K may be adjusted, such that fast updating occurs for new threats and controls, and slow updating for known threats and controls.

The mathematical translation from Item Response Theory to risk management is thus rather straightforward. Based on measured events, both threats and systems are rated on the latent ability scale. This will then allow us to predict the likelihood of a threat causing damage to a system in the future (the vulnerability). However, in a practical sense, it is not always clear how to define threats, systems and events. It is in particular this problem that we face in the application, and we will discuss it in more detail in the next section.

5. APPLICATION

We are now ready to apply the rating to different kinds of threats. Threats can be classified based on a number of dimensions or properties, for example, whether a threat is observable or not, malicious or benign, or preventable or not. Some of these dimensions may be orthogonal, for example observability and maliciousness. We believe that an exploration of threat dimensions will help in understanding and mitigating threats. For this article we focus on maliciousness, and we distinguish between non-malicious and malicious threats. We finally consider the application of our techniques to (vulnerabilities of) software products.

5.1 Non-Malicious Threats

As an example of a non-malicious threat, we consider the threat of storms to power lines. In this case, the threat capability is related to the wind speed, and the control strength to the construction quality of the lines. We distinguish two main questions for this application case:

1. What is the role of existing threat capability scales, e.g., wind speed?
2. How to determine the relevant population or set of entities: which storms are the "same" storm, and which power lines are the "same" power lines?

The answer to the first problem is relatively straightforward: the constraints of the model will define the scales of threat capability and control strength from the data, as latent traits. Based on the results, one could then reconstruct the relation between the emergent threat capability scale and, say, the wind speed in km/h. It is, however, neither necessary nor sufficient to have a predefined threat capability scale.

For the second question, the issue at stake is how to define the population that the rating will apply to. In the context of children against math problems, the populations are already well-defined, and it is clear when the same child is solving a different problem (apart from technical issues of authentication and shared accounts).

But in the threat context, are we speaking about storms against power lines, attackers against organisations, or nation states against nation states? For each of these cases, what exactly do we rate? Do we rate individual storms / individual people, or classes of storms / organisations that people are members of? This gives rise to the notion of levels of abstraction, which we will come back to in the malicious case. Different levels of abstraction could be used simultaneously: one can rate both organisations and people, but the rating of the organisation and the ratings of its members will be related. In most cases, choosing a single level of abstraction suffices. Thus, one will either rate individual instances, or groups.

The population does not need to be homogeneous, in the sense that the entities have similar properties. As with chess players or school children, the entities may have different backgrounds (or designs), and the model will automatically place the entities along the assumed single latent scale. However, there should be a reasonable assumption that a single scale is meaningful, and that there are no interfering variables. Also, there should be enough data available on the chosen level of abstraction. As a storm occurs only once, it is probably not a good idea to rate individual storms.

In the context of storms and power lines, the easiest solution for defining the population is grouping entities in distinct classes, and then updating the rating of the classes instead of the entities. For grouping the entities, predefined capability scales *can* be useful. One can then, say, use the Beaufort or Hurricane scales for wind speed, considering storms of the same class as identical for analysis purposes.

However, this means that arbitrary boundaries between classes would influence the results. For example, the Beaufort scale would put certain storms in class 9 and others in 10, but a storm high in class 9 is very close to a storm low in class 10. For accuracy, one would rather not lose information by grouping entities like this. When threat agents or controls are not identical but similar, a different strategy is needed. A potential solution is associating the rating with a representative instance of the class (e.g., the mean wind speed for Beaufort 9). If a storm occurs with wind speed somewhere between the representative instances of class 9 and 10, one would then update the ratings of *both* representative instances, with update speed proportional to the closeness to the actual wind speed.

Another solution would also use a similarity measure between entities to determine the update speed, but now *all* entities would be rated instead of representative instances of a class. If a storm destroys a power line, one would then not only update the rating of this storm and this power line, but also those of storms and power lines that are "close" on the similarity scale. The less similar, the lower the update factor. For the similarity measures, both predefined scales (wind speed in km/h) or the emergent Threat Capability scale could be used. Additional research is needed to uncover advantages and disadvantages of these three methods, in combination with tailored simulations or case studies.

Obviously, these considerations apply not only to threat events, but also to controls. Just as it is required to define when two threat events correspond to the same threat, we need to define when two threat events act upon the same control. Again, pre-defined classes can be used (same manufacturing type of a component, same operating system on a computer, etc.). The granularity of such classifications needs to be determined based on the requirements of the case. With large classes, there will be more data per class, but the precision is lower. With small classes, the theoretical precision is higher, but there may not be enough data to support the model.

5.2 Malicious Threats

On the side of malicious attacks, there is typically a human attacker that aims to compromise the security of an

asset or organisation. In an attack on information technology infrastructure, we have an attacker with a certain knowledge and capability, representing the threat to the organisation and consequently being measured by the threat capability. The control strength then is a measure for the organisation's capability to repel the attack, be it in form of policies or physical countermeasures. Again, we assume the relevant attack scenarios to be known here; identifying such scenarios is a different area of interest.

This kind of setting poses several challenges for the Elo rating system. Most notably, neither successful attacks nor successful defences are always reported or easy to detect. This difficulty can be avoided in penetration testing, where organisations systematically investigate the opportunities for attack. The foremost goal is to test shortcomings of security precautions and to report and document them. These tests may include benevolent attempts to gain access to physical as well as digital property, and may include manipulation of people, i.e., social engineering. Based on the outcomes of such tests, one can then try to measure the control strength of different measures, such as technical properties, detection mechanisms, and employee education.

Also, the "attackers" are known in penetration testing, whereas real attackers often remain hidden. Thirdly, penetration tests will generate knowledge on failed attempts, whereas we may not know how many failed attempts real attackers initiated before they finally succeed. Finally, a controlled setting makes it possible to select the attackers and systems that play against each other, maximising the games for optimal information (like selecting math problems appropriate to children's skill levels). These advantages are not limited to the penetration testing setting. A similar constellation is found when considering, e.g., compliance testing, certification, and accreditation of organisations or individuals with respect to standards.

The semantics of an Elo-score for penetration testing is a valuation of both the penetration tester and the tested organisation. This should not (and actually can not) replace standard procedures, such as patching detected vulnerabilities. The proposed Elo-scores based method allows judging the threat capability (of the tester) and the control strength of the organisation for use in risk assessment; it is not a preventive measure.

In applying our ranking approach, we consider three different scenarios. In the first, penetration testers play against organisations. In this scenario we can judge an individual tester's ability to find holes in a defence, and how good organisations are in plugging these holes. In the second scenario, regulations or standards play against organisations. In this scenario we can assess the difficulty of regulations, and organisations' ability to comply with these. Finally, in the third scenario, software products are rated according to their response to bugs. All three scenarios could be used to, for example, implement seal mechanisms for auditing of services [31].

5.2.1 Testers Against Organisations

To begin, we consider the person performing a penetration test as the threat, and the tested organisation as the asset. In this scenario, the threat capability is related to the tester's ability to unveil shortcomings in the organisation's defence, and the control strength is related to the organisation's ability to defend its assets.

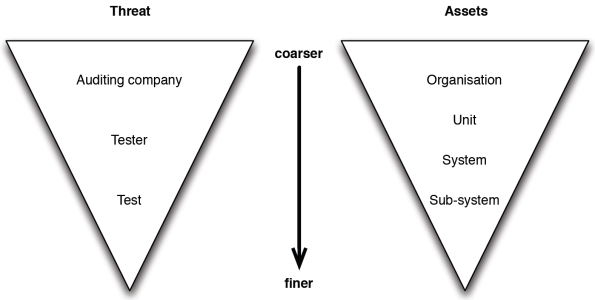

Figure 2: Different levels of granularity for the penetration test example. When moving from finer to coarser levels, the coarser level subsumes results from the finer elements. That is, an auditing company wins whenever, for example, a specific penetration test succeeds.

As in the ranking systems discussed in Section 2.1, testers and organisations start at a common initial rating. Every time a test is performed, its result is used for determining the new rankings of both parties. A successful penetration, i.e. when the target asset of the test is reached, counts as a win for the tester; an unsuccessful penetration counts as a win for the organisation.

By applying the ranking to testers and organisations we measure the tester's skill and the effectiveness of the organisation's protective measures. Testers can most likely get a sufficient number of games by performing penetration tests; organisations will probably only perform a limited number of penetration tests, which might be problematic to obtain meaningful numbers.

As before we face the issue of granularity. For both parties involved we can perform the games on different levels (see Figure 2). On the testing side the coarsest level would be that of the regulation being tested, and the finest that of an individual test performed. On the organisational side, the coarsest level is the organisation itself, while the finest level, depending on the kind of test, is the subsystem or employee being tested. One could imagine to choose even finer levels, such as different types of skills of the tester, or properties of the unit being tested, but it is questionable what could be gained from this.

When moving from finer to coarser levels of granularity, the coarser level subsumes results from the finer elements. This means that the results of any finer level are counted as the result for the coarser level, that is, for example, that an auditing company wins whenever a specific penetration test succeeds. In this way, a single event can cause updates of abilities at various levels.

5.2.2 Regulations Against Organisations

In the second case we consider organisations being certified in relation to some standard or regulation. In this case, the control strength is still related to the organisation, but now it measures its ability to implement a (part of the) regulation correctly, so it refers to the compliance capabilities of the organisation. The threat capability (interpreted as regulation difficulty) is now measured at the level of regulations (or parts thereof), removing the auditing company and its auditors from the equation (Figure 3). A win of the organisations in this case means that a (part of the) regu-

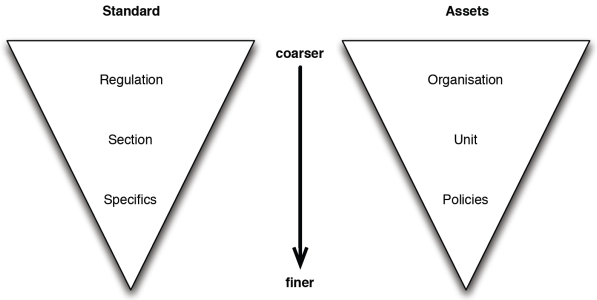

Figure 3: Different levels of granularity for the certification example. As before, coarser levels subsume results from the finer elements.

lation is implemented correctly (auditor finds no problems), whereas a win of the regulation means that it is not (auditor does find problems).

As explained above, the testers (auditors) now play on behalf of the regulation. This means that in the ranking each win of the auditor is counted as a win of the regulation. It is thus "regulation difficulty" rather than "tester ability" that is measured. Similar abstractions can of course be performed on the side of the organisation, e.g., playing and measuring on the country level, but that seems less natural when considering certification, where the question of interest is whether or not an organisation implements a certain standard or not.

It is interesting to consider what is measured in this setting. From the viewpoint of the asset, the organisation, not too much has changed; the resulting rank should still be read as effectiveness in fulfilling a certain regulation. From the viewpoint of the threat we now consider regulations, an abstract concept. This has two advantages: it increases the number of games that the rank will be based on, ensuring a higher reliability, and it provides a qualitative measure of the regulation.

What does a low score for a regulation mean? A low score results from many defeats, meaning that either the regulation (or parts hereof) is easy to implement and therefore loses against the organisation, or that the tester in question lacks the skills to detect that an organisation does not implement a regulation correctly. The latter case clearly is not related to the quality of the regulation. We assume that its impact will be negligible as different testers investigate its fulfilment. A high score for a regulation, on the other hand, means that no or only few organisations have been able or willing or interested in implementing it. This can either be on purpose, if designed to distinguish between different levels of fulfilling a regulation as is the case with ISO 15804 (The Common Criteria), or it can point at self-contradicting or nonsensical parts of a regulation.

If one would want to keep the tester in the game, then another option is to investigate how to realise rankings with three parties, where tester and regulation are distinguished; they play on the same side against the organisation, but have individual rankings, and the regulation's ranking is influenced differently by weak or strong testers.

5.2.3 Software Products

The same idea presented in the previous sections can be applied to estimate severity of software bugs and the resilience of software products they are found in. To realise this, we observe two events that are interpreted as games. The first event is the detection of a vulnerability. This game is won by the vulnerability over the software product the vulnerability is detected in, and their Elo scores are adjusted accordingly. The second event is the release of a bug fix. This game is won by the software product, and again the Elo scores are adjusted accordingly.

Instead of calculating ratings for individual vulnerabilities, they should be grouped in equivalence classes, where each class has a score of its own. When more patches against a vulnerability class become available, its rating will decrease. Software products that implement patches against a class of vulnerabilities at a later point than other products will still win the game against the vulnerability at that point, but will receive a smaller benefit, due to the already reduced rating of the vulnerability. Beyond this adaptation based on Elo scores, one could also add a second discount, based on how much time has passed since the vulnerability was detected, either overall, or in the software product in question.

The interpretation of scores for vulnerabilities is thus their distribution and yet untreated occurrences. The score for software products, on the other hand, measures the capability of its developers to counter and patch detected vulnerabilities fast. We are currently studying this particular example in more detail.

6. REQUIRED INFRASTRUCTURE

Based on the applications outlined above, we will discuss the infrastructure that is needed to support a collective effort in quantifying vulnerability. In particular, this concerns data availability and the use of the outcomes in risk management.

6.1 Willingness to Share Incident Data

First and foremost, if organisations are not willing to share incident data, not even in an anonymised form, then insufficient data will be available to calculate threat ratings and vulnerability levels. In order to make the Elo approach work, data on multiple threats against multiple organisations needs to be available, because the ratings can only be calculated in comparison. Therefore, more effort is needed to set up infrastructures for sharing incident data in a secure way, as well as (legal) incentives to use such infrastructures in practice. For this to work, sharing needs to be attractive both from an economic [14, 16, 23] and from a technical [21, 35] perspective.

The data issue is particularly pressing for the malicious case, where attackers may know about incidents that defenders are not aware of, and defenders may fail to report incidents to prevent similar attacks. Also, visibility of attacks may be reduced in cases of long-term exploitation rather than direct attacks, as well as when an attack fails already in its first stages. The information asymmetry in the malicious case and associated incentives for reporting need further study.

6.2 Data on Unsuccessful Threat Events

Secondly, it is not sufficient if only "successful" threats are reported. If we only obtain data about events where threat agents are successful, then the threat agents always win. In that case, there is no way to calculate their threat

capability. (Threat capabilities will just increase after every incident, but they will never decrease.) Therefore, organisations will need to have monitoring systems in place to identify such unsuccessful threat events, and will need to report these to the shared infrastructure. This holds for both accidental threats (storms against power lines) and malicious threats (cyber attacks). Thus, storms that do *not* destroy power lines and attackers that do *not* break into a system are essential to provide accurate and comparable ratings for threat capability and control strength.

6.3 Systematic Testing Efforts

As it may still be difficult to gather accurate operational data, testing efforts can help to provide input to the models. Small-scale experiments have been done in which penetration testers execute multi-step socio-technical attacks [6]. Such experiments could be extended in the context of Item Response Theory. Again, this will only help in establishing ratings if multiple threats and multiple organisations are tested against each other. For penetration testing in the context of cyber security and privacy, it has already been proposed to set up a public agency for this purpose [31]. With the current proposal to quantify vulnerability, the importance of such an institution has only increased.

To validate models estimated from data, one would need to separate training set from test set. This is typically easy with static ratings (Rasch), as these are calibrated on a training set by definition, and then used to measure other entities. For dynamic ratings (Elo), this is more difficult, as training and testing are combined. One could fix the ratings at a certain point in time, and then measure how well they perform with different entities. For example, one could establish Elo ratings for penetration testers based on an initial set of systems to attack, and then measure whether their performance on other systems matches their (fixed) rating.

6.4 Vulnerability, Threat Capability, and Risk

Although the main focus of this paper is on expressing vulnerability, we need to say something at this point on how to use logistic models of vulnerability in risk assessment. For reasons of simplicity, we assume that threat events cause a known amount of damage when "successful" against the controls, and no damage when unsuccessful.

We assume that risk is defined as the expected damage due to a specified type of threat within a specified time period. We then need models to express the following:

1. The damage caused by a successful threat event;
2. The expected number of threat events per time period (a fixed frequency if constant, or a density function if variable);
3. The distribution of the threat events over threat capability levels (relative frequencies for discrete threat capabilities, or a density function for continuous threat capabilities);
4. The likelihood of success for threat agents of specified threat capability (i.e., the logistic vulnerability function).

These four models can be combined to estimate the expected level of damage within a given time frame, as they will provide both the expected frequencies of threat events of particular capabilities (2 and 3 combined), as well as the expected damage for a single threat event of particular capability (1 and 4 combined). In this paper, we assume the first three models as given, and concentrate on the fourth, but it is important to have the overall picture in mind to see how the approach can be applied.

Typically, a countermeasure will reduce the vulnerability of an asset to a threat. (It may also reduce impact rather than vulnerability, but that case is not relevant here.) By implementing countermeasures, one will increase the rating of the defence (a different type of system will have a different rating), thereby reducing the vulnerability. As the countermeasure was not present in the system before, the increased control strength can only be estimated based on the actual effect of the countermeasure in different systems. A somewhat ironic observation is that less information may be available on the best countermeasures, as these measures may prevent attacks from taking place at all (attackers may divert their efforts elsewhere).

This is the central explanation that determines how the Elo-approach accounts for the success of countermeasures. However, if a countermeasure is added where no attacks are expected, it still won't reduce the risk, even if the vulnerability is reduced. Therefore, the Elo-rating alone cannot provide a full risk management approach.

Especially for malicious threats, estimation of threat event frequencies and associated threat capabilities is a big problem in itself. Malicious attackers will typically have knowledge of the vulnerability of a system, and adapt their behaviour (and therefore the number of threat events and their threat capability) to this knowledge. We do not address these issues in this paper.

In future research, we wish to investigate whether the logistic model can also be applied to represent the damage directly, i.e., interpreting the result as expected damage rather than probability of damage.

7. OPEN QUESTIONS

As the proposed approach to quantify vulnerability is new, there are many open problems to be studied in the research community. In the following we discuss some open questions within the proposed research paradigm, based on the applications outlined above.

7.1 Granularity

As discussed before, the identity of threats and controls is not as clear-cut as the identity of children and math problems. Therefore, the definition of populations of threats and controls is not straightforward. Robust classification approaches for threats and controls are needed to support logistic models of vulnerability. These models will most likely not be based on binary classifications, but rather on similarity measures between entities and their classes. The granularity of the classes plays an important role here. With small classes, the measurements are specific, but there are only few measurements per class, leading to large statistical uncertainties. Especially if attackers stop their malicious activities after a single successful (or unsuccessful) attack, not enough information will be available. With large classes, there are more measurements per class, but as the classes are larger, the predictions for future events may not be as precise. Further research is needed to identify the optimal population definitions for the application of Item Response

Theory to the particular context of security risk management.

7.2 Polytomous Rasch Models

The models we discussed assume that there is a binary variable to express winning or losing a game. In case of threats, this is not always as obvious as when children try to solve math problems. Therefore, we need clear definitions of what it means for a threat to be successful.

A question for further research is whether it is possible to use a different scale than binary for success. In the standard Rasch model, failure is represented by 0 and success by 1. However, threats may inflict different amounts of damage upon assets, and when the amount of damage is considered, intermediate values would be needed. For example, the outcome of a hack may be user access, administrator access, etc., which can be associated with different success values. This would lead to a so-called polytomous Rasch model. In such a model, it can be expressed that threat agents with lower threat capability will cause lower impact. This scenario would be similar to a child obtaining a lower score on a math task.

7.3 Mapping Latent Scale and Existing Scales

The models discussed in this paper will produce a latent scale on which both threat capabilities and control strengths are rated. In principle, all threats and all controls would then be rated on the same scale.

However, for particular types of threats, such as storms, commonly used scales may already exist. For prediction purposes, it may be useful to map these scales to the latent scale, for example to predict the effects of a storm with a particular wind speed. The scales will need to be fitted to one another to make this possible.

7.4 Graceful Degradation

As discussed before, data is needed to make the approach work, but getting data on security incidents is not always easy. When there is a lack of empirical data, can we gracefully degrade to other information sources, such as expert judgement? Can we then use the Rasch model of expert risk judgement presented in [11] in combination with our Rasch model of vulnerability? This would allow factoring in expert judgement where not enough sample data is available, without relying too much on subjective risk attitudes.

In such a model, there would be three classes of entities: threat agents (with threat capabilities), controls (with control strength), and experts (with risk bias). We would want to estimate all parameters based on actual events, as well as on judgements of the experts on the vulnerability.

7.5 Multi-Step Attacks

Another important question is how this approach would work if a malicious attack consists of more than one step, i.e., a multi-step attack [27]. One would then need to assess the vulnerability for each of the steps, and for the overall attacks. Also, the measure of success may be available only for the attack as a whole, requiring a distribution of the resulting control strength rating update over the different components involved, similar to update rules distributing reinforcements to the different connections in neural networks. The threat agent may consist of multiple components too: multiple attackers may cooperate and perform part of the steps each. This would require a distribution of the update on the Threat Capability side as well. Log analysis may be helpful in judging which steps of an attack have been executed successfully. The success level of an attack (like in polytomous Rasch models) may then be expressed in terms of the number of successful steps relative to the total number of steps (when known).

7.6 Including the Time Factor

For threat events as well as for problem solving, time may be relevant for judging skills or threat capabilities. If an attacker with limited computation power takes a week to crack a password, and an attacker with extensive computation power achieves the same result in an hour, then we may want the time to be reflected in the "score". As the Math Garden already takes time into account in the update rules, it may provide inspiration for similar attempts in the security setting, although the time scales for solving a math problem are obviously different than for launching a cyber attack.

One may generalise time spent into a more abstract notion of resources, where skill plus resources would determine the threat capability of attacks. In this context, resources can also be seen as an indication of motivation, as more motivated attackers will spend more resources on an attack. However, resources spent depend not only on motivation, but also on the resources available to the attacker. Whether other relevant aspects of motivation should be included as well needs to be determined. Also, the relation between resources spent by the attacker and likelihood of success may be quite different depending on the type of attack, e.g. brute force password cracking versus picking a lock versus social engineering. With brute force cracking, more computing time will lead to a higher chance of success. With lock picking, skill probably plays a bigger role than time spent, and with social engineering, spending more time may actually *reduce* the likelihood of success, as the activities may raise suspicion.

7.7 Preventing Arbitrary Rating Fluctuations

In the Math Garden case, it can be reasonably assumed that there are no rapid major changes in the populations of children and problems that cause major variations in the ratings. However, in case of threats, a very successful virus (say), may suddenly become very *un*successful after a patching round. In the hypothetical case that all virus attacks would be monitored, the first organisations that the virus would attack after the patch would win against a very highly ranked opponent (in terms of threat capability), and would therefore see a major increase in their control strength. The threat capability of the virus would decrease accordingly. Consequently, organisations that are attacked only later would see a smaller increase in their control strength. This would lead to a rather arbitrary difference in control strength values among organisations, depending on the order in which the virus would attack them after the patching round. This problem could potentially be addressed by investigating different means of determining the update (K) factor, for example by employing Kalman filters [18].

Another issue in this context is how to handle after-the-fact discovery of an attack. When an attack has already happened, but has not been identified yet, the rating of a

system may lag behind with respect to the actual situation. If the attack is then discovered, it is possible in theory to reduce the rating retrospectively. However, this would require recomputing *all* ratings, as other ratings may also have been updated based on the inaccurate rating. This is especially problematic if ratings would be based on different sources (e.g. certification versus actual attacks), which is therefore not recommended.

7.8 Simulations and Experiments

Simulations with artificial input data, based on different sets of assumptions, can provide additional information on properties of the models. For example, given the arms race between attackers and defenders in a malicious context, one would expect ratings of both attackers and defenders to increase gradually over time. Such hypotheses could be tested in simulations. Also, simulations allow for sensitivity analysis, i.e. establishing the sensitivity of vulnerability or risk values to variations in ratings, or vice versa.

Furthermore, it would be possible to set up experiments in which hackers play against previously prepared systems with certain known weaknesses, e.g., virtual machines in the cloud. Such experiments could provide initial skill levels (threat capability) of penetration testers, and would also provide data on how ratings would converge or fluctuate in practice, depending on different settings of the system. In a socio-technical setting, the experiments by Dimkov et al. on laptop theft [6] could serve as an inspiration for evaluating the approach empirically. However, contrary to these experiments, penetration testers would need to play multiple scenarios to get meaningful data. In this case, both the skills of the testers and the quality of the scenarios would influence success.

7.9 Triangular Games

In this paper, three settings have come up where three rather than two types of entities interact:

1. An expert (risk bias) predicts the vulnerability of an asset (control strength) to a threat (threat capability);
2. Auditors / penetration testers (testing skill) judge the compliance of an organisation (compliance capability) with a regulation (regulation difficulty).
3. Penetration testers (testing skill) use a scenario (scenario quality) on an organisation (control strength).

In such situations, one would want to estimate all three latent abilities based on the outcome of events. Typically, a combination of two of the entities will win or lose against the third, and update rules should reflect this by distributing the won/lost rating points over the winning/losing entities. For example, if a penetration test succeeds, then the combination of tester and scenario wins against the organisation, and they should share the points. The right kind of models (in particular update rules) for such settings still need to be determined, and simulations should be executed to study their properties.

8. CONCLUSIONS

This paper is part of an initiative to quantify security risks. We have identified several research topics in this area, and focused on the question how to measure the vulnerability of a system to a particular threat. Above, we have discussed an innovative framework to define and measure this vulnerability, combining the Open Group Risk Taxonomy with the Math Garden rating system, based on Item Response Theory. The approach employs Rasch models and Elo ratings to quantify vulnerability, which provides a justifiable relation between threat capability, control strength, and vulnerability. Furthermore, the explicit calculation of threat capability and control strength yields the advantage that predictions can be made about the "success" of future threat events. This is similar to predicting the success of a child in solving a math problem, without having solved the same problem before. We have outlined how this approach could work in security risk management practices, and we have identified open problems for future research within this new framework.

In future work, we will – in addition to the questions outlined above – deal in more depth with quantification of threat event frequencies, as well as the aggregation of vulnerability to particular threats into an overall measure of system vulnerability. Combining these ideas, we can extend the foundations provided by The Open Group into a complete quantitative security risk management paradigm. However, the application of Item Response Theory to security by itself already provides many new directions of study, and may inspire both modelling and empirical research in the field.

Acknowledgements

The research of the first author is supported by financial assistance of the European Commission in the context of the SESAME project. The views expressed herein are those of the authors and can therefore in no way be taken to reflect the official position of the European Commission. The authors wish to thank (in alphabetical order) Raoul Grasman and Sharon Klinkenberg for a helpful brainstorm session, and the NSPW participants, the reviewers, and our shepherd Rainer Böhme for excellent comments.

9. REFERENCES

[1] R. Anderson and T. Moore. The economics of information security. *Science*, 314(5799):610–613, 2006.

[2] A. Birnbaum. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, editors, *Statistical theories of mental test scores*, chapter 17–20, pages 397–479. Addison-Wesley, Reading, MA, 1968.

[3] S. Bleikertz, M. Schunter, C. W. Probst, D. Pendarakis, and K. Eriksson. Security audits of multi-tier virtual infrastructures in public infrastructure clouds. In *Proceedings of the 2010 ACM workshop on Cloud computing security workshop*, CCSW '10, pages 93–102, New York, NY, USA, 2010. ACM.

[4] R. Böhme. A comparison of market approaches to software vulnerability disclosure. In G. Müller, editor, *Emerging Trends in Information and Communication Security*, volume 3995 of *Lecture Notes in Computer Science*, pages 298–311. Springer Berlin / Heidelberg, 2006.

[5] R. Böhme and M. Félegyházi. Optimal information security investment with penetration testing. In

T. Alpcan, L. Buttyán, and J. Baras, editors, *Decision and Game Theory for Security*, volume 6442 of *Lecture Notes in Computer Science*, pages 21–37. Springer Berlin / Heidelberg, 2010. 10.1007/978-3-642-17197-0_2.

[6] T. Dimkov, W. Pieters, and P. H. Hartel. Laptop theft: a case study on effectiveness of security mechanisms in open organizations. In *Proceedings of the 17th ACM Conference on Computer and Communications Security (CCS), Chicago, Illinois, US*, pages 666–668. ACM, October 2010.

[7] T. Dimkov, W. Pieters, and P. H. Hartel. Portunes: representing attack scenarios spanning through the physical, digital and social domain. In *Proceedings of the Joint Workshop on Automated Reasoning for Security Protocol Analysis and Issues in the Theory of Security (ARSPA-WITS'10). Revised Selected Papers, Paphos, Cyprus*, volume 6186 of *Lecture Notes in Computer Science*, pages 112–129, Berlin, March 2010. Springer Verlag.

[8] A. Elo. *The rating of Chessplayers, Past and present.* Arco Publishers, New York, 1978.

[9] World football Elo ratings. Available at `http://en.wikipedia.org/wiki/World_Football_Elo_Ratings`. Last accessed March 2012.

[10] Elo rating system. Available at `http://en.wikipedia.org/wiki/Elo_rating_system`. Last accessed March 2012.

[11] S. Figini, R. S. Kenett, and S. Salini. Optimal scaling for risk assessment: merging of operational and financial data. *Quality and Reliability Engineering International*, 26(8):887–897, 2010.

[12] D. Florencio and C. Herley. Sex, lies and cyber-crime surveys. Technical Report MSR-TR-2011-75, Microsoft Research, June 2011.

[13] S. Foley and W. Fitzgerald. An approach to security policy configuration using semantic threat graphs. In Ehud Gudes and Jaideep Vaidya, editors, *Data and Applications Security XXIII*, volume 5645 of *Lecture Notes in Computer Science*, pages 33–48. Springer Berlin / Heidelberg, 2009. 10.1007/978-3-642-03007-9_3.

[14] E. Gal-Or and A. Ghose. The economic incentives for sharing security information. *Information Systems Research*, 16(2):186–208, 2005.

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

[16] K. Hausken. Information sharing among firms and cyber attacks. *Journal of Accounting and Public Policy*, 26(6):639–688, 2007.

[17] C. Ioannidis, D. Pym, and J. Williams. Information security trade-offs and optimal patching policies. *European Journal of Operational Research*, 216(2):434 – 444, 2012.

[18] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

[19] S. Klinkenberg, M. Straatemeier, and H. L. J. van der Maas. Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Comput. Educ.*, 57:1813–1824, September 2011.

[20] B. Kordy, S. Mauw, S. Radomirović, and P. Schweitzer. Foundations of attack–defense trees. In *Formal Aspects of Security and Trust, 7th International Workshop, FAST 2010*, volume 6561 of *Lecture Notes in Computer Science*, pages 80–95. Springer, 2011.

[21] P. Lincoln, P. Porras, and V. Shmatikov. Privacy-preserving sharing and correction of security alerts. In *Proceedings of the 13th conference on USENIX Security Symposium - Volume 13*, SSYM'04, pages 17–17, Berkeley, CA, USA, 2004. USENIX Association.

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

[23] D. Liu, Y. Ji, and V. Mookerjee. Knowledge sharing and investment decisions in information security. *Decision Support Systems*, 52(1):95–107, 2011.

[24] M. S. Lund, B. Solhaug, and K. Stølen. *Model-Driven Risk Analysis: The CORAS Approach.* Springer, 2011.

[25] S. Mauw and M. Oostdijk. Foundations of attack trees. In D. Won and S. Kim, editors, *Proc. 8th Annual International Conference on Information Security and Cryptology, ICISC'05*, volume 3935 of *Lecture Notes in Computer Science*, pages 186–198. Springer, 2006.

[26] P. Mell, K. Scarfone, and S. Romanosky. Common vulnerability scoring system. *Security & Privacy, IEEE*, 4(6):85–89, 2006.

[27] 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 *Proceedings of the 24th Annual ACM Symposium on Applied Computing, SAC'2009, Honolulu, Hawaii, USA*, pages 66–73, New York, March 2009. ACM.

[28] J. Pamula, S. Jajodia, P. Ammann, and V. Swarup. A weakest-adversary security metric for network configuration security analysis. In *Proceedings of the 2nd ACM workshop on Quality of protection*, QoP '06, pages 31–38, New York, NY, USA, 2006. ACM.

[29] W. Pieters. Representing humans in system security models: An actor-network approach. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, 2(1):75–92, 2011.

[30] C. W. Probst and R. R. Hansen. An extensible analysable system model. *Information security technical report*, 13(4):235–246, 2008.

[31] C. W. Probst, M. A. Sasse, W. Pieters, T. Dimkov, E. Luysterborg, and M. Arnaud. Privacy penetration testing: How to establish trust in your cloud provider. In S. Gutwirth, R. Leenes, P. De Hert, and Y. Poullet, editors, *European Data Protection: In Good Health?*, pages 251–265. Springer Netherlands, 2012.

[32] G. Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests.* MESA Press, 1960.

[33] J. Rowe, K. Levitt, S. Parsons, E. Sklar, A. Applebaum, and S. Jalal. Argumentation logic to assist in security administration. In *Proceedings of the 2012 New Security Paradigms Workshop (NSPW).* ACM, 2012.

[34] B. Schneier. Attack trees: Modeling security threats. *Dr. Dobb's journal*, 24(12):21–29, December 1999.

[35] A. Slagell and W. Yurcik. Sharing computer network logs for security and privacy: a motivation for new methodologies of anonymization. In *Security and Privacy for Emerging Areas in Communication Networks, 2005. Workshop of the 1st International Conference on*, pages 80–89, September 2005.

[36] K. Wauters, P. Desmet, and W. Van Den Noortgate. Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers & Education*, 58:1183–1193, 2012.

[37] M. Yearworth, B. Monahan, and D. Pym. Predictive modelling for security operations economics. In *Proc. I3P Workshop on the Economics of Securing the Information Infrastructure*, 2006.