

Ontology-based generation of Bayesian networks

Stefan Fenz, A Min Tjoa
Institute of Software Technology and Interactive Systems
Vienna University of Technology &
Secure Business Austria
Vienna, Austria
{fenz, tjoa}@ifs.tuwien.ac.at

Marcus Hudec
Institute of Scientific Computing
University of Vienna
Vienna, Austria
marcus.hudec@univie.ac.at

Abstract

Bayesian networks are indispensable for determining the probability of events which are influenced by various components. Bayesian probabilities encode degrees of belief about certain events and a dynamic knowledge body is used to strengthen, update, or weaken these assumptions. The creation of Bayesian networks requires at least three challenging tasks: (i) the determination of relevant influence factors, (ii) the determination of relationships between the identified influence factors, and (iii) the calculation of the conditional probability tables for each node in the Bayesian network. Based on existing domain ontologies, we propose a method for the ontology-based generation of Bayesian networks. The ontology is used to provide the necessary knowledge about relevant influence factors, their relationships, their weights, and the scale which represents potential states of the identified influence factors. The developed method enables, based on existing ontologies, the semi-automatic generation and alternation of Bayesian networks.

1. Introduction

According to Neapolitan [11], Bayesian networks are 'graphical structures for representing the probabilistic relationships among a large number of variables and doing probabilistic inference with those variables'. According to Heckerman et al. [6], the Bayesian probability of an event x is a person's degree of belief in the very event. Therefore, the frequentistic probability concept (e.g. the probability that a coin lands heads or tails, based on a large number of tries tending towards infinity) is distinguished from the Bayesian (personal) probability, in which probabilities encode degrees of belief about certain events and data is used to strengthen, update, or weaken these assumptions (e.g. the degree of belief that the coin will land heads at the next

throw) [14]. These useful characteristics are applied to several real-world applications (cf. [11, 7, 17, 10]).

The following challenges arise at the generation of Bayesian networks: (i) identification of variables which are relevant to the considered domain, (ii) identification of relationships between the identified variables, (iii) determination of a scale which represents the states of the identified variables, and (iv) creation of conditional probability tables for each variable. Referring to the ontology definition by Neches, ontologies are a potential solution to address the stated challenges: *an ontology defines the basic terms and relations comprising the vocabulary of a topic area as well as the rules for combining terms and relations to define extensions to the vocabulary* [12].

Based on existing domain ontologies, this paper proposes a method for the ontology-based generation of Bayesian networks, by (i) using ontology concepts to create the nodes of the Bayesian network, (ii) using ontology relations to link the Bayesian network nodes, and (iii) exploiting the ontological knowledge base to support the conditional probability table calculation for each node. The method is demonstrated on the example of the threat probability determination, which uses a security ontology as the underlying formal knowledge base. Therefore, the first part of the paper reviews related work, introduces the security ontology and the goals of the threat probability determination, whereas the second part describes the applied method for the ontology-based generation of Bayesian networks.

2. Related Work

Sadeghi et al. [15] use the descriptive information captured in an ontology to construct Bayesian decision models. The authors assume that the underlying ontology is specifically created to model the considered decision problem and that one hypothesis can explain the states of evidence observed in the domain. While the authors describe their overall approach - (i) build the ontology, (ii) implement the

Bayesian network model, (iii) acquire the knowledge, and (iv) build the knowledge base - no detailed description regarding the ontology-based derivation of the Bayesian network is given.

Zheng et al. [19] encode Clinical Practice Guidelines into an ontology that contains uncertainty features, and propose an algorithm which uses these features to construct the structure and the conditional probability tables of Bayesian networks. The proposed approach requires the creation of an ontology providing the properties necessary for the Bayesian network generation and does not consider the usage of already existing domain ontologies.

Devitt et al. [2] propose the following phases for automating the Bayesian network construction by ontologies: (i) identification of the variables of interest, (ii) specification of the values these variables can take, (iii) definition of the relations between the variables, and (iv) assignment of a conditional probability distribution. The main difference between our approach and the approach by Devitt et al. is that we construct the Bayesian network directly from existing domain ontologies and do not require any Bayesian network-specific ontology extensions.

3. The Security Ontology

The security ontology [4, 3, 5] was proposed based on the security relationship model described in the National Institute of Standards and Technology Special Publication 800-12 [13]. Figure 1 shows the high-level concepts and corresponding relations of the ontology. A threat gives rise to follow-up threats, represents a potential danger to the organization's assets and affects specific security attributes (e.g. confidentiality, integrity, and/or availability) as soon as it exploits a vulnerability in the form of a physical, technical, or administrative weakness, and it causes damage to certain assets. Additionally each threat is described by potential threat origins (human or natural origin) and threat sources (accidental or deliberate source). For each vulnerability a severity value and the asset on which the vulnerability could be exploited is assigned. Controls have to be implemented to mitigate an identified vulnerability and to protect the respective assets by preventive, corrective, deterrent, recovery, or detective measures (control type). Each control is implemented as asset concept, or as combinations thereof. Controls are derived from and correspond to best-practice and information security standard controls (e.g. the German IT Grundschutz Manual [1] and ISO/IEC 27001 [8]) to ensure the incorporation of widely accepted knowledge. The controls are modeled on a highly granular level and are thus reusable for different standards. When implementing the controls, a compliance with various information security standards is implicit. To enrich the knowledge model with concrete information security knowledge

the German IT Grundschutz Manual has been superimposed on the security ontology and more than 500 information security concepts and 600 corresponding formal axioms have been integrated into the ontological knowledge base (cf. [5]). The coded ontology follows the OWL-DL (W3C Web Ontology Language) [18] standard and ensures that the knowledge is represented in a standardized and formal form.

Since the security ontology provides detailed knowledge about threat, vulnerability, and control dependencies, this knowledge can be utilized to build up the Bayesian network for the threat probability determination. Figure 2 gives an overview of the connections between the proposed Bayesian threat probability determination and the security ontology. The following section provides an in-depth description of the proposed method and shows how the method has been applied to create the Bayesian threat probability determination network.

4. Ontology-based Generation of Bayesian Networks

In the following we aim at describing the ontology-based generation of Bayesian networks by using the security ontology, which provides the foundation to enrich the Bayesian network with concrete knowledge. The main phases of the proposed method are:

1. **Concepts** → **nodes**. Those ontology concepts, which are relevant to the considered problem and should be represented in the Bayesian calculation schema are selected to establish the nodes of the Bayesian network.
2. **Relations** → **links**. Ontology relations starting and ending between the selected concepts are used to establish the links between the Bayesian network nodes.
3. **Axioms** → **node scales and weights**. Scale- and weight-relevant axioms are used to determine potential states and weights of the Bayesian network nodes.
4. **Instances** → **findings**. Instances of concepts which are represented by the Bayesian network's leaf nodes are used to derive and enter concrete findings in the Bayesian network.

4.1 Concepts → nodes

Based on an existing ontology, the analyst has to select those concepts which are relevant to the considered problem and should be represented by Bayesian network nodes.

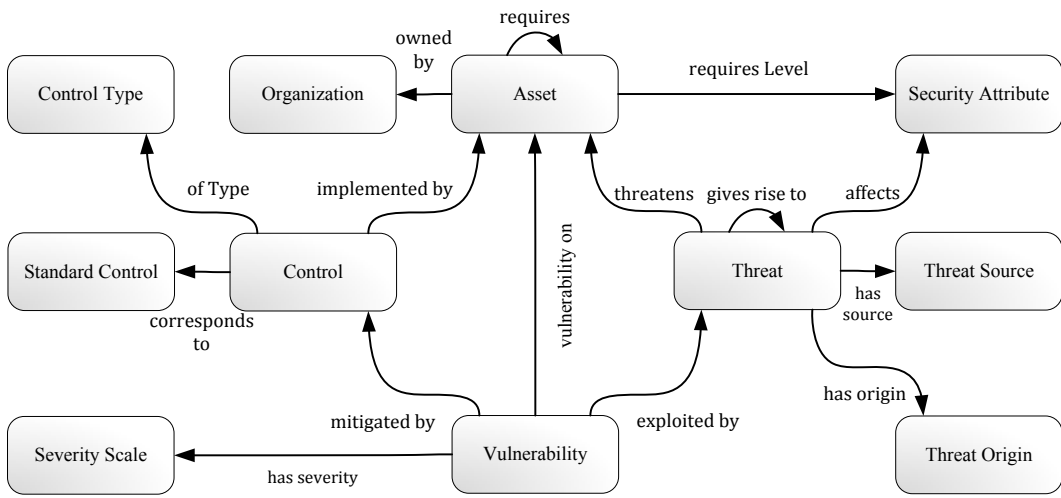


Figure 1. Security relationships

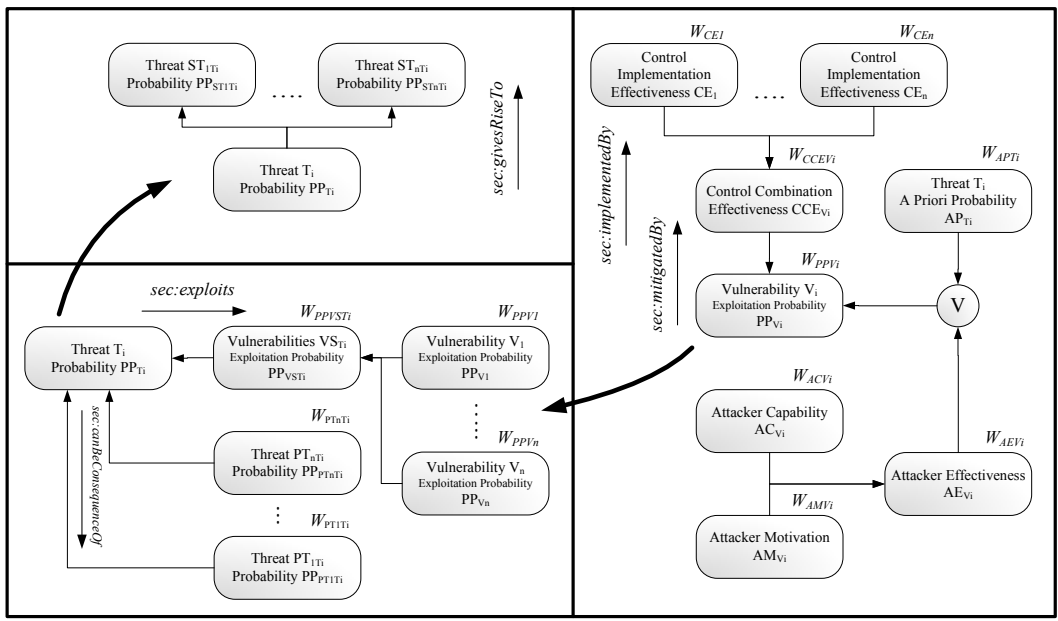


Figure 2. Utilizing the security ontology for the Bayesian threat probability determination

Concept	Node
sec:Threat	PP_{T_i}
-	PP_{VST_i}
sec:Vulnerability	PP_{V_i}
sec:Control	CCE_{V_i}
sec:Asset	CE_i
sec:Attacker	AE_{V_i}
sec:AttackerMotivation	AM_{V_i}
sec:AttackerCapability	AC_{V_i}
sec:APrioriProbability	AP_{T_i}

Table 1. Concept - node mapping

The objective of the exemplary Bayesian network, which is built to demonstrate the proposed ontology-based generation method, is to determine the probability of threats taking various influence factors into account. Therefore, the following factors have been identified: (1) predecessor threats ($PT_{1T_i}, \dots, PT_{nT_i}$) influence the considered threat (T_i) which influences its successor threats ($ST_{1T_i}, \dots, ST_{nT_i}$); therefore dependencies amongst a given threat set T had to be considered (see the upper left section in Figure 2), (2) according to Stoneburner et al. [16], each threat (T_i) requires one or more vulnerabilities (V_1, \dots, V_n) to become effective; thus the existence of unmitigated vulnerabilities significantly influences the threat probability (see the lower left section in Figure 2), (3) controls can be used to mitigate identified vulnerabilities, while the mitigation depends on the effectiveness of a potential control combination (CCE_{V_i}) which again depends on the actual effectiveness of the control implementations which are used in this combination (CE_1, \dots, CE_n), and (4) (a) in the case of deliberate threat sources, the vulnerability exploitation probability (PP_{V_i}) is determined by the effectiveness of a potential attacker (AE_{V_i}) which is again determined by the motivation (AM_{V_i}) and the capabilities (AC_{V_i}) of the attacker as stated in ISO 27005 [9], (b) in the case of accidental threat sources and/or natural threat origins, the vulnerability exploitation probability (PP_{V_i}) is determined by the a priori probability (AP_{T_i}) of the corresponding threat (T_i) (see the right section of Figure 2). The result for each threat probability is represented as a distribution of the chosen rating scale (e.g. high, medium, and low). Figure 2 shows the proposed model for determining threat probabilities by taking the aforementioned factors into consideration.

Table 1 shows the mapping between selected concepts and their representation in the Bayesian network. Since PP_{VST_i} is an intermediate node - i.e. a node which is introduced to reduce the calculation complexity by aggregating the states of its parents - it is not represented in and extracted from the security ontology. Note that the current mapping is just concerned with the selection of high-level concepts (e.g. sec:Threat PP_{T_i} and sec:Vulnerability PP_{V_i}). Concrete representatives of the selected high-level concepts are derived by querying their sub-concepts (e.g. sec:Fire $\rightarrow PP_{T_{Fire}}$ and sec:NoMalwareProtection $\rightarrow PP_{V_{NoMaPr}}$).

Concept I	Relation	Concept II
sec:Threat	sec:giveRiseTo	sec:Threat
sec:Threat	sec:canBeConsequenceOf	sec:Threat
sec:Threat	sec:exploits	sec:Vulnerability
sec:Vulnerability	sec:mitigatedBy	sec:Control
sec:Control	implementedBy	sec:Asset
sec:Threat	sec:hasProbability	sec:APrioriProbability
sec:Attacker	sec:hasMotivation	sec:AttackerMotivation
sec:Attacker	sec:hasCapability	sec:AttackerCapability

Table 2. Concept relations

Node	Parents
PP_{T_i}	$\{PP_{VST_i}, PP_{T_i}\}$
PP_{VST_i}	$\{PP_{V_i}\}$
PP_{V_i}	$\{CCE_{V_i}, (AE_{V_i} AP_{T_i})\}$
CCE_{V_i}	$\{CE_i\}$
CE_i	$\{\}$
AE_{V_i}	$\{AM_{V_i}, AC_{V_i}\}$
AM_{V_i}	$\{\}$
AC_{V_i}	$\{\}$
AP_{T_i}	$\{\}$

Table 3. High-level node links

4.2 Relations \rightarrow links

Subsequent to the node creation, the dependencies among the nodes have to be set up. As ontological relations already model these dependencies they are used to create the links between the Bayesian network nodes. For each relation of each concept selected in the previous phase it has to be checked if the considered relation starts and ends at a concept which has been selected in the previous phase. If this is true for the considered relation it can be used to connect the corresponding Bayesian network nodes. While the potential relations can be derived automatically from the ontology, the link direction (i.e. the determination of parent and child nodes) requires the human interpretation of the ontological relation. In contrast to the node creation phase, the link creation phase only considers the high-level concepts selected in the previous phase.

Table 2 shows each relation which exists between the security ontology concepts selected in the previous phase. Since each node in the Bayesian network belongs to one selected high-level concept which is represented in Table 2 each node can be linked to appropriate parent or child nodes. Table 3 shows the final relationships in the established Bayesian network.

4.3 Axioms \rightarrow node scales and weights

While the previous two phases set up the nodes and their links, the current phase is concerned with equipping each node with an appropriate scale (i.e. potential states) and (if applicable) its weight in the context of its siblings to support the conditional probability table generation.

The determination of potential node scales requires the definition of the ontology concept which represents the scale concept. In the context of the threat probability determination the concept `sec:Scale` provides with its sub-concepts several node scales (e.g. a three-point Likert scale). Now, the formal axioms of each high-level concept defined in the first phase can be queried regarding statements including `sec:Scale` sub-concepts. If such a concept has been found the considered node can be equipped with the given scale (e.g. high, medium, and low).

After setting up the nodes, node links, and node scales the conditional probability tables (CPTs) for each node have to be calculated to finalize the Bayesian network structure. A mathematical function is required which describes how the state of each node is influenced by the state of its parent nodes. Currently, this knowledge is not provided by the security ontology. What is provided is the weight of each node in the context of its siblings. In the case of vulnerabilities the security ontology provides a severity rating for each vulnerability (e.g. high, medium, and low). As the vulnerabilities vector $PP_{V_{S_{T_i}}}$ is determined by single vulnerabilities and their weights, the weight of each vulnerability which influences the intermediate vulnerabilities vector $PP_{V_{S_{T_i}}}$ was determined. Since the security ontology provides a severity rating S_{V_i} for each vulnerability (high (3), medium (2), and low(1)), a numerical weight $W_{PP_{V_i}}$ for each vulnerability can be determined by dividing the severity of the considered vulnerability by the severity sum of all vulnerabilities relevant to the threat:

$$W_{PP_{V_i}} = \frac{S_{V_i}}{\sum_{j=1}^n S_{V_j}} \quad (1)$$

Although, node-specific mathematical functions are not provided by the security ontology, globally defined functions incorporating the derived node weights can be used for the CPT calculation.

4.4 Instances → findings

The previous phases have created a Bayesian network including relevant nodes, node links, node scales, and node weights supporting the CPT calculation. The current phase exploits the ontological knowledge base to provide concrete findings to the Bayesian network.

First, potential input nodes have to be defined. The proposed Bayesian threat probability determination requires input at the leaf nodes AM_{V_i} (attacker motivation), AC_{V_i} (attacker capability), CE_i (control implementation effectiveness), and AP_{T_i} (a priori threat probability). Concrete instances of the corresponding ontology concepts are used to provide concrete findings to the Bayesian network.

5. Example

Figure 3 shows an exemplary application of the generated Bayesian network. The depicted example uses the generated Bayesian network to determine the probability for the Break In threat by incorporating the exploitation probability of connected vulnerabilities (No Entrance Control, No Intrusion Alarm System, No Secure Doors, and No Secure Windows). Since the Break In threat has a deliberate threat origin its vulnerabilities are connected to a specific attacker profile, which is represented in the ontology by the instance `sec:AttackerSBA`. Since the modeled capability is set to `sec:high` and the motivation is set to `sec:medium`, these findings can be entered into the Bayesian nodes AC_{SBA} and AM_{SBA} representing the concrete attacker profile in the Bayesian network. The second component which influences the vulnerability exploitation probability and therefore the final threat probability are the effectiveness values of available control implementations. By changing the control combination the exploitation probability of the corresponding vulnerability and subsequently the probability of the corresponding threat alter to maintain the consistency of the Bayesian network. Substituting the medium-effective safety windows by highly effective ones decreased the Break In probability from the range 18-51% to 15-49%. Please note that these ranges are derived from the distribution of the Bayesian node which represents the Break In probability.

6. Conclusion

At the creation of Bayesian networks analysts are confronted amongst others with the following challenges: (i) What are the relevant influence factors for my problem?, (ii) How do these influence factors relate to each other?, (iii) What are potential states of the identified influence factors?, and (iv) What is the weight of each influence factor in the context of its siblings? To address these challenges, this paper has proposed an ontology-based method for the generation of Bayesian networks and has shown its applicability by an use case in the field of threat probability determination.

The proposed method (i) enables the semi-automatic creation of Bayesian networks by using existing ontologies, (ii) reduces the complexity of modeling Bayesian networks by using high-level concepts and relations to integrate relevant sub-concepts into the Bayesian network, and (iii) provides by the usage of ontologies the possibility of easily maintaining the underlying knowledge body of Bayesian networks. The limitations of the proposed method are: (i) functions for calculating conditional probability tables are not provided by the ontology and have to be modeled externally, and (ii) human intervention is still necessary if the ontology

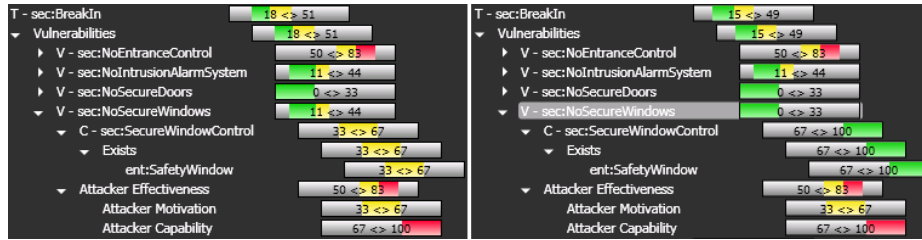


Figure 3. Application of the generated Bayesian network

provides a knowledge model which does not exactly fit the domain of interest. Further research has to address these limitations to provide a more efficient method for creating and updating Bayesian networks.

7 Acknowledgments

This work was supported by grants of the Austrian Government's FIT-IT Research Initiative on Trust in IT Systems under the contract 813701 and was performed at the Research Center Secure Business Austria funded by the Federal Ministry of Economics and Labor of the Republic of Austria (BMWA) and the City of Vienna.

References

- [1] BSI. IT Grundschutz Manual, 2004.
- [2] A. Devitt, B. Danev, and K. Matusikova. Constructing bayesian networks automatically using ontologies. In *Proceedings of Second Workshop on Formal Ontologies Meets Industry (FOMI 2006)*, 2006.
- [3] A. Ekelhart, S. Fenz, G. Goluch, and E. Weippl. Ontological mapping of common criteria's security assurance requirements. In H. Venter, M. Eloff, L. Labuschagne, J. Eloff, and R. von Solms, editors, *New Approaches for Security, Privacy and Trust in Complex Environments, Proceedings of the IFIP TC 11 22nd International Information Security Conference, IFIPSEC2007, May 14-16*, volume 232/2007 of *IFIP International Federation for Information Processing*, pages 85–95, Sandton, South Africa, May 2007. International Federation for Information Processing. 978-0-387-72366-2.
- [4] A. Ekelhart, S. Fenz, M. Klemen, and E. Weippl. Security ontologies: Improving quantitative risk analysis. In *Proceedings of the 40th Hawaii International Conference on System Sciences, HICSS2007*, pages 156–162, Los Alamitos, CA, USA, January 2007. IEEE Computer Society. 0-7695-2755-8.
- [5] S. Fenz. *Ontology- and Bayesian-based information security risk management*. PhD thesis, Vienna University of Technology, October 2008.
- [6] D. Heckerman. A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Advanced Technology Division, Redmond, WA 98052, November 1996.
- [7] D. Heckerman, J. Breese, and K. Rommelse. Troubleshooting under uncertainty. Technical Report MSR-TR-94-07, Microsoft Research, Redmond, Washington, January 1994.
- [8] ISO/IEC. ISO/IEC 27001:2005, Information technology - Security techniques - Information security management systems - Requirements, 2005.
- [9] ISO/IEC. ISO/IEC 27005:2007, Information technology - Security techniques - Information security risk management, November 2007.
- [10] R. Kennett, K. Korb, and A. Nicholson. Seabreeze prediction using Bayesian networks. In *PAKDD '01: Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 148–153, London, UK, 2001. Springer-Verlag.
- [11] R. Neapolitan. *Learning Bayesian networks*. Prentice Hall, 2003.
- [12] R. Neches, R. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator, and W. R. Swartout. Enabling technology for knowledge sharing. *AI Mag.*, 12(3):36–56, 1991.
- [13] NIST. An Introduction to Computer Security - The NIST Handbook. Technical report, NIST (National Institute of Standards and Technology), October 1995. Special Publication 800-12.
- [14] J. Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, March 2000.
- [15] S. Sadeghi, A. Barzi, and J. Smith. Ontology driven construction of a knowledgebase for bayesian decision models based on umls. *Stud Health Technol Inform*, 116:223–228, 2005.
- [16] G. Stoneburner, A. Goguen, and A. Feringa. Risk management guide for information technology systems. NIST Special Publication 800-30, National Institute of Standards and Technology (NIST), Gaithersburg, MD 20899-8930, July 2002.
- [17] J. G. Torres-Toledano and L. E. Sucar. Bayesian networks for reliability analysis of complex systems. In *IBERAMIA '98: Proceedings of the 6th Ibero-American Conference on AI*, pages 195–206, London, UK, 1998. Springer-Verlag.
- [18] W3C. OWL - web ontology language, February 2004.
- [19] H.-T. Zheng, B.-Y. Kang, and H.-G. Kim. An ontology-based bayesian network approach for representing uncertainty in clinical practice guidelines. In F. Bobillo, P. C. G. da Costa, C. d'Amato, N. Fanizzi, F. Fung, T. Lukasiewicz, T. Martin, M. Nickles, Y. Peng, M. Pool, P. Smrz, and P. Vojs, editors, *URSW*, volume 327 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2007.