From Data to Knowledge Engineering for Cybersecurity

Gerardo I. Simari

Department of Computer Science and Engineering, Universidad Nacional del Sur (UNS), Argentina
Institute for Computer Science and Engineering (UNS–CONICET), Argentina
Arizona State University, USA
gis@cs.uns.edu.ar

Abstract

Data present in a wide array of platforms that are part of today's information systems lies at the foundation of many decision making processes, as we have now come to depend on social media, videos, news, forums, chats, ads, maps, and many other data sources for our daily lives. In this article, we first discuss how such data sources are involved in threats to systems' integrity, and then how they can be leveraged along with knowledge-based tools to tackle a set of challenges in the cybersecurity domain. Finally, we present a brief discussion of our roadmap for research and development in the near future to address the set of ever-evolving cyber threats that our systems face every day.

1 Introduction

Ever since the onset of the so-called "Information Age", data present in a wide range of platforms has played an important role in much of the global population's decision making processes. One important example of such platforms is social media, which empowers users to share information and knowledge in a quick and simple way. People increasingly get information and news from social media, and base their daily activities on what they find there. A particularly sensitive domain in which data analysis is playing an increasingly important part is cybersecurity, which is generally loosely defined as the set of practices designed to defend computer systems from "malicious" attacks—here, *malicious* refers to the point of view of the systems' owners or administrators.

The rest of this article is organized into three parts: first, we discuss a set of challenges faced by cybersecurity practitioners—given the ever-evolving landscape of threats to computer systems worldwide, we do not aim for this list to be complete; second, we briefly mention several AI-based tools that in our opinion can be used as building blocks towards effective approaches to address these challenges; finally, we provide a discussion of our roadmap for combining data-driven and knowledge-based AI theory and practice in the near future, and briefly present some ideas already being developed in this direction.

In the following, we provide brief discussions on various topics that are relevant to tackling real-world cybersecurity problems; given the lack of space and forum in which this work is presented (IJCAI *Early Career Spotlight*), we generally give references to our own work and refer the interested readers to references therein.

2 Modern Cybersecurity Challenges

In this broad discussion of challenges related to cybersecurity, we take the broad stance that issues related to *malicious social interactions* and *manipulation of information or views* is to some degree related to the pursuit of the integrity of information systems—a proper discussion of this stance is outside the scope of this paper, but pondering the effects of such threats in the daily lives of the majority of the world's population should suffice to consider this view to be reasonable.

We will divide the discussion into two parts: in the first we describe areas in which we have carried out research and development towards addressing some challenge; then, we briefly mention other relevant areas and problems.

Cyber Attribution. Finding who is responsible for a cyber attack (or any cyber event of interest) is commonly referred to as *attribution*. This process presents several unique problems—chief among them are that the technical artifacts produced by cyber attacks are difficult to understand, and it is easy (and quite useful) for a malicious actor to carry out actions designed to deceive their opponents. See [Nunes *et al.*, 2018b] for a discussion of several AI-based approaches towards solving this family of problems.

Predicting Specific Attacks. This refers to evidence-based predictions that certain software products, platforms, or vendors are at risk of being the target of a cyber attack. In [Nunes *et al.*, 2018a] and [Almukaynizi *et al.*, 2018] we inspect discussions by hackers in forums and marketplaces in the Deep Web towards this end, leveraging information present in the National Vulnerability Database (NVD)[1].

Adversarial Deduplication. The classical problem in the Databases literature of deciding that multiple virtual objects (such as tuples in a personnel table) actually map to the same real-world entity is called *deduplication* or *entity resolution*. This problem has been effectively addressed and solved for many real-world cases, with the underlying assumption that the situation arose due to involuntary errors such as typos,

[1] https://nvd.nist.gov/

imperfect OCR, or as the result of data integration efforts. However, in most cybersecurity scenarios, this kind of mapping arises as the result of malicious intentions—consider for example the hackers behind the accounts used in the previous paragraph, who wish to remain anonymous and might decide to create multiple accounts to escape law enforcement. In [Paredes *et al.*, 2018b] we take some first steps towards applying machine learning classifiers to text obtained from the same kind of posts described above. In [Paredes *et al.*, 2018a] and Section 4 we discuss how such basic tools can be leveraged in the (semi-)automatic derivation of hypotheses to help security analysts in their tasks.

Managing Information Flow. There are many consequences that arise from the fact that there are now many different ways in which information flows virtually freely among people and organizations—most of the problems in the next category are rooted precisely in this cause. Adequately managing and reasoning about the processes by which information finds its way from one actor to another is therefore a fundamental tool. In [Gallo *et al.*, 2017] we present the concept of *Network Knowledge Bases* (NKBs), which is a model based on multi-layer complex networks that allows to represent the individual beliefs of each node, as well as multiple attributes of nodes and their relations. The NKB model is designed to represent information from multiple social platforms in a single structure, and thus holds all necessary information to address problems related to the management of information flow.

Other Challenges

There are many other challenges related to cybersecurity that have been addressed with AI tools—both Machine Learning and Knowledge Representation-based—that we only have space to briefly mention here; the following discussion is thus only a sample and not meant to be complete.

Consider the task of *prioritizing patching activities*; this is very important since many cyber attacks are successfully carried out by exploiting publicly known vulnerabilities (that can be found, for instance, in the NVD) and for which patches have already been developed—one salient example of this is the Heartbleed bug, which affected specific versions of the OpenSSL cryptography library used in many implementations of the TLS protocol (used in many forms of secure communications, such as Web browsing). Since enterprise systems typically consist of hundreds of subsystems, and there is no "one button solution" to apply all patches, prioritizing the to-do list of patches in a way that reflects real threats is the basis of minimizing unnecessary risk.

Another high-impact task is *identifying bots/botnets*, which are autonomous systems that can range from simple scripts to sophisticated intelligent agents and play a major role in many cyber activities such as stock market trading, customer care, and of course social media platforms. Deciding whether or not an online user is being controlled by a person or a *bot* is essential to deterring or hindering malicious activities (more on this below). *Botnets* are networks of bots organized towards achieving a specific goal, such as stealing CPU cycles from infected hosts to carry out cryptocurrency mining activities, or disseminate fake news.

Sometimes related to bots and adversarial deduplication, *sock puppets* are identities used in service of online deception, such as business promotion or favorable reviews. They can be effectively used, for instance, to cause people to believe that certain information is true (related to fake news and manipulated discussions, described below), believe that certain candidates or views have a larger following, or to avoid restrictions such as bans due to improper behavior. A particular use of sock puppets is called *Sybil attack*; in this kind of malicious interaction, an attacker creates multiple identities to achieve greater influence and mislead the reputation system of a peer-to-peer network.

It is hard to overestimate the importance of *fake news*; the events surrounding the US, Argentine, and Brazilian presidential elections, as well as the Brexit vote in the UK—between 2015 and 2018—uncovered the vulnerability of modern society's psyche. Since then, the propagation of false (or twisted versions of the facts) through social media designed to favor a certain candidate/outcome in an election has proved effective throughout the globe. This practice is now commonly referred to as *fake news*, and is part of the greater *post truth* phenomenon, in which the actual degree of truth behind pieces of information has lost importance. Detecting and curtailing such practices is another cornerstone of preserving the (data) integrity of information systems.

Finally, *other forms of manipulation* may involve actions taken towards manipulating people or online discussions that take other forms not mentioned above, such as phishing, social engineering, trolling, or other kinds of bullying. Having proper systems in place to detect, deter, or otherwise avoid this kind of behavior is instrumental in preserving the overall integrity of our systems.

3 AI Tools for Cybersecurity

We now present a set of building blocks that in our view will be fundamental in the development of effective approaches in the cybersecurity realm. Since work involving these tools often combine more than one, we will briefly present each and then discuss how they have already been used towards solving specific problems.

3.1 Building Blocks

Argumentation. This area of study seeks to model decision-making and reasoning based on the way humans carry out the process (when done in an organized and effective way, such as in legal trials) [Rahwan et al., 2009]. Approaches are essentially divided into *abstract* and *structured*—in the former, arguments are indivisible elements and only the relations among them are modeled (such as attack and support); on the other hand, in the latter arguments are sets of elements that can be analyzed, and relationships between them can arise from relationships between specific elements (such as one conclusion attacking another). Structured argumentation-based reasoning has, among others, two main strengths: it can be easily integrated into *human-in-the-loop* systems (where human users can intervene in different ways, such as detecting and removing erroneous information sources, cf. Section 4) and results are accompanied by struc-

tures (such as dialectical trees) that can be used in the derivation of *explanations* to support conclusions.

Extended Logic-based Languages. There have been many developments in this direction; one of the most significant for reasoning about unknown objects (i.e., through so-called *value invention*) in a flexible and scalable manner is Datalog+/– [Calì *et al.*, 2012; Simari *et al.*, 2017], a family of ontological languages that arose from the Databases community as a more powerful (and, to some, readable) alternative to Description Logics. Since its initial development, there have been several extensions (discussed below) incorporating different capabilities that enhance its expressive power (such as preferences, probabilities, and temporal reasoning).

Preferences. Reasoning about how certain types of objects (such as query answers, domain elements, cyber threats, etc.) can be *ordered* is a fundamental capability when solving certain kinds of problems. One of the main computational hurdles to overcome in this domain is that the number of ways in which n objects can be ordered is $n!$, and the factorial function grows faster than any exponential (with a constant base). However, significant progress has been made towards incorporating this capability into practical tools.

Inconsistency Tolerance and Belief Revision. It is very common for databases and inferences made from richer knowledge bases to fail to preserve logical consistency; this happens because data integrity may not be actively enforced, as the result of data integration, or any number of other possible reasons. However, simply throwing out any data that is involved in some form of conflict is not recommended, so many inconsistency-tolerant semantics have been developed to both perform query answering or data cleaning. *Belief Revision* is an area of study that is closely related to this effort, in which the main problem studied is how an "epistemic input"—a piece of information posed as a potential addition to a knowledge base—should be incorporated (or not). Clearly, both efforts play a central role in dealing with constantly evolving data streams that are very likely to contain inconsistencies and can also be designed to be deceptive.

Reasoning about Networked Data. Closely related to the challenge discussed in the previous section regarding managing information flow, tools developed to support reasoning about networks is crucial in addressing many real-world problems. This involves effectively representing rich interconnected heterogenous information, modeling (and manipulating) cascades, node centrality, among other issues that incorporate others already appearing on this list, such as uncertainty, temporal reasoning, and computational tractability.

Data-driven Models. Machine Learning-based tools have recently received much attention due to their success in tackling certain real-world problems such as image classification, game playing, and natural language understanding. Though their success has certainly been impressive, we believe that they only represent a piece of the bigger picture, in much the same way that purely logic-based approaches will never be able to solve complex problems on their own. Below we discuss several successes in the general endeavor of combining developments born out of two often conflicting views of how AI should be developed.

3.2 Towards AI & Cybersecurity

Work mostly done in the past decade or so using these "basic" building blocks (actually, entire areas of study) with direct or indirect applications to cybersecurity include efforts to model cyber attribution problems with probabilistic argumentation [Shakarian *et al.*, 2015; Nunes *et al.*, 2016], belief revision operations with this model [Shakarian *et al.*, 2016], and predicting at-risk systems using hybrid structured argumentation and data-driven tools [Nunes *et al.*, 2018a].

Much work has also been carried out with respect to extending ontology languages for reasoning under uncertainty, such as [Lukasiewicz *et al.*, 2012a; Gottlob *et al.*, 2013], incorporating preferences [Lukasiewicz *et al.*, 2013], and inconsistency-tolerance [Lukasiewicz *et al.*, 2012b; Lukasiewicz *et al.*, 2015], Other logic-based tools include effectively using deception for cyber defense using probabilistic logic [Jajodia *et al.*, 2017], logic programming formalisms to model multiple complex cascades in networks [Shakarian *et al.*, 2013], probabilistic temporal reasoning [Shakarian *et al.*, 2011] that was later applied to predict enterprise-targeted cyber attacks [Almukaynizi *et al.*, 2018], and flexible models to represent and tractably reason with probabilistic preferences [Lukasiewicz *et al.*, 2014; Lukasiewicz *et al.*, 2016].

Finally, other works we can mention here are Inconsistency Management Policies (IMPs) [Martinez *et al.*, 2014], Belief Revision operators in NKBs [Gallo *et al.*, 2017], graph-theoretic models for defending moving targets [Dickerson *et al.*, 2010], and logic-based formalisms for reasoning about degrees of fulfillment [Simari *et al.*, 2008] (which could be used, for instance, to determine degrees of compliance to security policies).

There is clearly much more work to be done; we refer the interested reader to a recent overview in this same forum of how many of these approaches can be leveraged in the development of decision-support systems [Martinez, 2017].

4 Discussion and Outlook

Even though the AI & Cybersecurity literature has seen an explosion of developments in recent years, as observed in the works mentioned in the previous two sections (and references therein), many of the successes have been "one-off" efforts towards solving specific problems. In this section, we provide an overview of an approach that we believe can be used to incorporate such individual successes towards next-generation systems that provide effective decision support in cybersecurity domains. The ideas presented here are summarized from recently developed work (currently under review).

Consider the setup described in Figure 1, where we have a bipartite knowledge base comprised of: (i) An *Ontological KB*, which contains both basic data that is ingested from multiple, constantly updated, information sources such as social media platforms, Dark and Deep Web sources, news sites, NVD, etc., as well as logic-based rules that encode the knowledge available about the domain in question. Datalog+/– is a natural candidate to be used as the basis in the design of a language for this module. (ii) A *Network KB/Diffusion Model*,

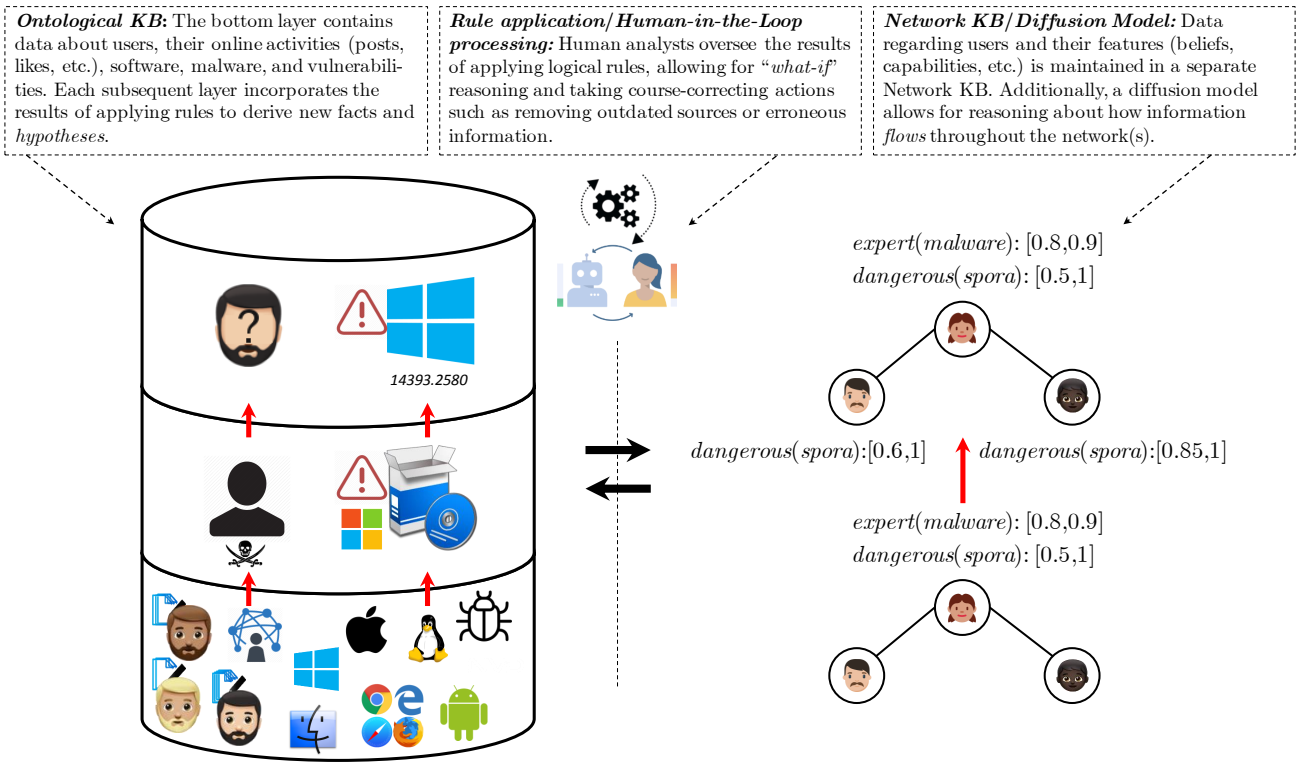

Figure 1: A framework to support semi-automated reasoning processes that combine ontological reasoning, network diffusion models, and HITL processing (*not shown*: multiple data sources that are constantly being updated). Value invention in ontology languages like Datalog+/– allows to pose hypotheses based on information that is currently available, which can then be updated as new information arrives.

which contains information regarding actors of interest, how they are related to each other, and rules that model information diffusion processes. The work of [Shakarian *et al.*, 2013] has the necessary requirements to be used in this module.

There is interaction between both parts since the ontological KB can use the information in the network KB to derive new knowledge. The crucial capability of *value invention*—inferring the existence of an object with certain characteristics without having explicit knowledge of an object that fits the bill—boasted by ontology languages is crucial in what we call *automatic generation of hypotheses*. For example, if we are observing posts on hacker forums mentioning a specific vulnerability X (mentioned using a CVE number and thus linked to other info in the NVD), and others requesting to purchase an exploit for this same vulnerability, we can infer the existence of both an actor wishing to carry out an attack on software Y with vulnerability X, and a potential victim whose systems infrastructure includes software Y.

Implementing such a system involves overcoming several formidable hurdles, including: (i) *Data and knowledge engineering*: Domain experts and automatic tools are required in order to derive ontological and diffusion rules. It is thus necessary to acquire adequate datasets and experts that are familiar with the specific problem or set of problems being tackled. (ii) *Data and knowledge integration*: Data inputs come from different sources, and they must be integrated in a common schema; though this is a classical problem in Databases and Ontology-mediated Query Answering, for which many tools have been developed, addressing it in combination with the following point makes it especially difficult to overcome effectively in this kind of domain. (iii) *Stream reasoning*: The reasoning tasks necessary for solving problems in cybersecurity domains use large amounts of frequently updated data that often cannot be stored at once, leading to problems similar to classical view maintenance or stream processing in Databases but with an additional reasoning layer; this is commonly known as *stream reasoning*. The main challenge lies in striking an adequate balance between the computational complexity of the tasks in the reasoning layer and the frequency with which updates must be handled. (iv) *Decidability and tractability*: Adding expressive power to ontology languages can quickly make them intractable or undecidable; however, it is possible to set constraints to rein in this power—this is the main idea behind the Datalog+/– family ("+" for expressive power and "–" for constraints). Since our approach requires extended reasoning algorithms that are more general than the classical ones such as the *chase*, we can expect to have at least the same kind of problems as the ones mentioned above. Thus, we must study conditions that either guarantee tractable complexity or obtain good approximations. Finally, the kind of *interactions* allowed between the Ontological and Network KBs are key in this task.

This roadmap will guide our medium-term R&D agenda towards developing hybrid KR-ML tools for cybersecurity.

Acknowledgments

This work was supported by Universidad Nacional del Sur (UNS), CONICET, and EU H2020 MSCA grant No. 690974 for project "MIREL". We thank CYR3CON (https://www.cyr3con.ai/) for providing access to their datasets.

References

[Almukaynizi *et al.*, 2018] Mohammed Almukaynizi, Ericsson Marin, Eric Nunes, Paulo Shakarian, Gerardo I. Simari, Dipsy Kapoor, and Timothy Siedlecki. DARKMENTION: A deployed system to predict enterprise-targeted external cyberattacks. In *Proc. IEEE ISI*, pages 31–36, 2018.

[Calì *et al.*, 2012] Andrea Calì, Georg Gottlob, and Thomas Lukasiewicz. A general Datalog-based framework for tractable query answering over ontologies. *J. Web Semant.*, 14:57–83, 2012.

[Dickerson *et al.*, 2010] John P. Dickerson, Gerardo I. Simari, V. S. Subrahmanian, and Sarit Kraus. A graph-theoretic approach to protect static and moving targets from adversaries. In *Proc. AAMAS*, pages 299–306, 2010.

[Gallo *et al.*, 2017] Fabio R. Gallo, Gerardo I. Simari, Maria Vanina Martinez, Marcelo Falappa, and Natalia Abad Santos. Reasoning about sentiment and knowledge diffusion in social networks. *IEEE Internet Comput.*, 21(6):8–17, 2017.

[Gottlob *et al.*, 2013] Georg Gottlob, Thomas Lukasiewicz, Maria Vanina Martinez, and Gerardo I. Simari. Query answering under probabilistic uncertainty in Datalog+/– ontologies. *AMAI*, 69(1):37–72, 2013.

[Jajodia *et al.*, 2017] Sushil Jajodia, Noseong Park, Fabio Pierazzi, Andrea Pugliese, Edoardo Serra, Gerardo I. Simari, and V. S. Subrahmanian. A probabilistic logic of cyber deception. *IEEE TIFS*, 12(11):2532–2544, 2017.

[Lukasiewicz *et al.*, 2012a] Thomas Lukasiewicz, Maria Vanina Martinez, Giorgio Orsi, and Gerardo I. Simari. Heuristic ranking in tightly coupled probabilistic description logics. In *Proc. UAI*, pages 554–563, 2012.

[Lukasiewicz *et al.*, 2012b] Thomas Lukasiewicz, Maria Vanina Martinez, and Gerardo I. Simari. Inconsistency handling in Datalog+/- ontologies. In *Proc. ECAI 2012*, pages 558–563, 2012.

[Lukasiewicz *et al.*, 2013] Thomas Lukasiewicz, Maria Vanina Martinez, and Gerardo I. Simari. Preference-based query answering in Datalog+/– ontologies. In *Proc. IJCAI*, pages 1017–1023, 2013.

[Lukasiewicz *et al.*, 2014] Thomas Lukasiewicz, Maria Vanina Martinez, and Gerardo I. Simari. Probabilistic preference logic networks. In *Proc. ECAI*, pages 561–566, 2014.

[Lukasiewicz *et al.*, 2015] Thomas Lukasiewicz, Maria Vanina Martinez, Andreas Pieris, and Gerardo I. Simari. From classical to consistent query answering under existential rules. In *Proc. AAAI*, pages 1546–1552, 2015.

[Lukasiewicz *et al.*, 2016] Thomas Lukasiewicz, Maria Vanina Martinez, David Poole, and Gerardo I. Simari. Probabilistic models over weighted orderings: Fixed-parameter tractable variable elimination. In *Proc. KR*, pages 494–504, 2016.

[Martinez *et al.*, 2014] Maria Vanina Martinez, Francesco Parisi, Andrea Pugliese, Gerardo I. Simari, and V. S. Subrahmanian. Policy-based inconsistency management in relational databases. *IJAR*, 55(2):501–528, 2014.

[Martinez, 2017] Maria Vanina Martinez. Knowledge engineering for intelligent decision support. In *Proc. IJCAI*, pages 5131–5135, 2017.

[Nunes *et al.*, 2016] Eric Nunes, Paulo Shakarian, and Gerardo I. Simari. Toward argumentation-based cyber attribution. In *Proc. AICS@AAAI*, pages 177–184, 2016.

[Nunes *et al.*, 2018a] Eric Nunes, Paulo Shakarian, and Gerardo I. Simari. At-risk system identification via analysis of discussions on the darkweb. In *Proc. APWG eCrime*, pages 1–12, 2018.

[Nunes *et al.*, 2018b] Eric Nunes, Paulo Shakarian, Gerardo I. Simari, and Andrew Ruef. *Artificial Intelligence Tools for Cyber Attribution*. Springer, 1st edition, 2018.

[Paredes *et al.*, 2018a] José Paredes, Maria Vanina Martinez, Gerardo I. Simari, and Marcelo Falappa. Leveraging probabilistic existential rules for adversarial deduplication. In *Proc. PRUV@IJCAR 2018*, 2018.

[Paredes *et al.*, 2018b] José Paredes, Gerardo I. Simari, Maria Vanina Martinez, and Marcelo Falappa. First steps towards data-driven adversarial deduplication. *Information*, 9(8):189, 2018.

[Rahwan et al., 2009] Iyad Rahwan et al. *Argumentation in Artificial Intelligence*, volume 47. Springer, 2009.

[Shakarian *et al.*, 2011] Paulo Shakarian, Austin Parker, Gerardo I. Simari, and VS Subrahmanian. Annotated probabilistic temporal logic. *TOCL*, 12(2):14:1–14:44, 2011.

[Shakarian *et al.*, 2013] Paulo Shakarian, Gerardo I. Simari, and Devon Callahan. Reasoning about complex networks: A logic programming approach. *TPLP*, 13(4-5), 2013.

[Shakarian *et al.*, 2015] Paulo Shakarian, Gerardo I. Simari, Geoffrey Moores, and Simon Parsons. Cyber attribution: An argumentation-based approach. In *Cyber Warfare – Building the Scientific Foundation*, pages 151–171. 2015.

[Shakarian *et al.*, 2016] Paulo Shakarian, Gerardo I. Simari, Geoffrey Moores, Damon Paulo, Simon Parsons, Marcelo Falappa, and Ashkan Aleali. Belief revision in structured probabilistic argumentation: Model and application to cyber security. *AMAI*, 78(3-4):259–301, 2016.

[Simari *et al.*, 2008] Gerardo I. Simari, Matthias Broecheler, V. S. Subrahmanian, and Sarit Kraus. Promises kept, promises broken: An axiomatic and quantitative treatment of fulfillment. In *Proc. KR*, pages 59–69, 2008.

[Simari *et al.*, 2017] Gerardo I. Simari, Cristian Molinaro, Maria Vanina Martinez, Thomas Lukasiewicz, and Livia Predoiu. *Ontology-Based Data Access Leveraging Subjective Reports*. Springer, 2017.