Argumentación probabilística y revisión de creencias con aplicaciones a ciberseguridad

José N. Paredes **Marcelo A. Falappa** **Gerardo I. Simari**

Inst. de Cs. e Ing. de la Computación (Universidad Nacional del Sur–CONICET)
Av. Alem 1253, (8000) Bahía Blanca, Argentina
{jose.paredes,mfalappa,gis}@cs.uns.edu.ar

Resumen

Esta línea de investigación se centra en los aspectos algorítmicos y de representación de conocimiento y razonamiento asociados con los procesos de razonamiento dialéctico y dinámica del conocimiento bajo incertidumbre probabilística. La investigación es conducida por la aplicación de éstos en entornos relacionados con ciberseguridad y ciberguerra; dada la aplicabilidad de los resultados en datos provenientes del mundo real, la tratabilidad computacional es un tema central del proyecto.

Palabras clave: Representación de conocimiento y razonamiento, Argumentación, Revisión de creencias, Razonamiento bajo incertidumbre probabilística, Ciber seguridad.

1. Contexto

La presente línea de investigación se encuentra inserta en el marco del proyecto PGI 24/N035 "*Argumentación y Dinámica de Creencias para mejorar las capacidades de razonamiento y representación de conocimiento de Sistemas Multi-Agente*", financiado por la Universidad Nacional del Sur, y el proyecto PIP-CONICET 112-201101-01000 "*Combinación de Revisión de Creencias y Argumentación para mejorar las capacidades de Razonamiento y modelado de la Dinámica de Conocimiento en Sistemas Multi-agente*", financiado por el Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), ambos llevado a cabo dentro del Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur.

2. Introducción

A continuación haremos un breve relevamiento de las áreas más cercanas a esta línea de investigación, enfocándonos en los trabajos que comparten en mayor medida nuestros objetivos.

Sistemas argumentativos

Para los seres humanos, argumentar es una forma natural de encontrar una base segura para las creencias; es una manera de manejar racionalmente la información contradictoria con el fin de establecer qué es posible creer en el contexto de lo que ha sido ya aceptado. La actividad humana de discutir, y la propia naturaleza de una discusión, han sido objeto de intensa investigación en Filosofía desde la antigüedad, y la Lógica nació del esfuerzo realizado para estructurar la presentación de argumentos y su intercambio. El proceso de argumentación (tanto la humana como la computacional) refleja una forma de razonamiento en el cual tanto la conclusión como la forma de llegar a ella pueden ser cuestionadas; es decir, se modela un proceso dialéctico en el cual las partes proponen argumentos a favor de sus puntos de vista o en contra de los de la contraparte, y el proceso determina cuál es la posición ganadora.

En la literatura se han propuesto muchas formas de llevar a cabo este proceso, las cuales se dividen principalmente en abstractas [Dun95] y concretas [GS04]; las primeras se caracterizan por el análisis de los argumentos disponibles y la relación de ataque entre ellos, mientras que las últimas asumen la disponibilidad de la estructura interna de cada argumento. Esta área de estudio es de especial interés en el ámbito de la Inteligencia Artificial, principalmente porque permite razonar con información incompleta e incierta, y permite manejar inconsistencias en los sistemas basados en conocimiento. En las últimas dos décadas, la investigación en argumentación como mecanismo de razonamiento se ha expandido enormemente, y el campo está produciendo un gran variedad de resultados.

Es claro que este tipo de razonamiento puede ser útil para nuestros objetivos; como punto de partida para esta investigación se considerarán tanto los sistemas de argumentación abstracta como los sistemas concretos basados en reglas. En particular, se tomará como referente a tres sistemas concretos que han sido desarrollados en la última década: Defeasible Logic Programming (DeLP) [GS04], ABA [DKT09] y ASPIC+ [Pra10], pero también se tendrá en cuenta la aproximación abstracta, ya que en general resulta útil en las etapas iniciales de modelado cuando es necesario establecer qué propiedades debe cumplir el sistema.

Razonamiento bajo incertidumbre probabilística

Las herramientas para llevar a cabo un proceso de razonamiento probabilístico han sido estudiadas por mucho tiempo y en muchas disciplinas, dado su amplio campo de aplicación. Para nuestros objetivos, utilizaremos herramientas maduras que poseen implementaciones optimizadas y mantenidas periódicamente; dos ejemplos son las Redes Bayesianas [Pea14] y las Redes Lógicas de Markov [RD06], aunque se realizará un relevamiento completo del campo para identificar otras posibilidades.

Revisión de creencias

El problema que se estudia en el área de Revisión de Creencias (*Belief Revision*, en inglés) es el de cómo deben cambiar los estados epistémicos de un agente ante una nueva entrada epistémica; en otras palabras, cómo deben revisarse las creencias ante la presencia de información nueva que posiblemente contradice las creencias establecidas hasta el momento. Tradicionalmente, los estados epistémicos han tomado la forma o bien de conjuntos de creencias (conjuntos de fórmulas lógicas cerrados bajo consecuencia) [AGM85, Gär03] o bases de creencias (conjuntos no cerrados) [Han94, Han97].

Es evidente que el problema de la revisión de creencias aparece constantemente en el mundo real y, por lo tanto, cualquier sistema inteligente que funcione con datos del mismo deberá estar equipado con alguna forma de revisión. La metodología típica consiste en el desarrollo de operadores que toman la base de conocimiento actual y la entrada epistémica, y producen una nueva base de conocimiento que corresponde al resultado de revisar las creencias. Estos operadores en general se caracterizan por las propiedades que deben cumplir (expresadas en la forma de postulados); luego se proponen construcciones algorítmicas y se demuestra formalmente que las dos caracterizaciones son equivalentes; éste tipo de resultado lleva el nombre de teorema de representación.

De particular interés para esta línea de investigación es la capacidad de hacer revisión de creencias en entornos en los que se espera que haya intenciones de engañar al agente razonador (con información engañosa); por ejemplo, una persona o equipo que ejecuta un ciberataque típicamente intenta dejar pistas falsas para que los investigadores pierdan el rastro del verdadero responsable. A nuestro entender, no existen operadores de revisión de creencias que se hayan desarrollado con esta característica en mente. Por otra parte, también se explorarán operadores que sean especialmente propicios para funcionar con aspectos cuantitativos; el único trabajo en esta línea es la propuesta reciente de [SSF15].

3. Líneas de Investigación, Desarrollo e Innovación

Los últimos años han visto una verdadera explosión en el interés por el desarrollo y estudio de formalismos de argumentación probabilística. Si bien la mayor parte de esta atención ha estado enfocada hasta el momento en sistemas de argumentación abstracta [Hun12, Thi12]. existen también desarrollos con sistemas argumentativos concretos; el primer trabajo en esta línea fue el de [HKL00]; desde entonces, los trabajos más relevantes para este plan han sido los de Hunter [Hun13] y Shakarian, Simari y Falappa [SSF14].

La conexión entre la revisión de creencias y los sistemas argumentativos fue estudiada por primera vez en [Doy79]; en [FGKIS11] se realiza un relevamiento hasta la fecha de publicación y además se desarrollan operadores nuevos que explotan esta conexión. Los operadores basados en explicaciones de [FKIS02] son también relevantes a la combinación de estas dos áreas, dado que las explicaciones de por qué una revisión se hace de una manera particular pueden ser vistas como argumentos.

Por otra parte, la revisión de creencias probabilísticas también ha recibido algo de atención en la literatura, aunque menos que las discutidas anteriormente. Los trabajos funcionales en la línea de investigación más relevante para la nuestra — la que combina métodos de revisión clásicos con conocimiento probabilístico -– son los de [KIR04], que se centra en la fusión de información probabilística, y [CD05], que utiliza además evidencia incierta como entrada epistémica. Más recientemente, en [Dub11] se propone un tratamiento general que abarca la revisión y fusión en entornos tanto cualitativos como cuantitativos.

Por último, la única línea de investigación de la que tenemos conocimiento que combina las tres áreas es la iniciada recientemente por Shakarian, Simari, Falappa y otros [SSM$^+$15, SSF14, SSMP15, SSF15]. Estos trabajos plantean un formalismo general para realizar revisión de creencias en un lenguaje basado en DeLP extendido con presuposiciones [MGS12] y anotaciones probabilísticas. El modelo consiste de una base de conocimiento bipartita: el *modelo analítico* es un programa DeLP que incluye los datos y reglas que aplican al dominio en cuestión, mientras que el *modelo del entorno* es una base de conocimiento probabilística. Por último, una *función de etiquetado* relaciona ambas partes al asociar fórmulas compuestas por combinaciones de eventos básicos provenientes del modelo del entorno. En la Figura 1 se muestra un ejemplo de predicados que podrían utilizarse para modelar un dominio de ciberseguridad; una función de etiquetado podría por ejemplo asociar $motivo(X, Y)$ con el evento probabilístico $enConf(X, Y)$, denotando que se puede concluir que X tiene motivo de atacar a Y cuando éstos están en conflicto entre ellos.

Si bien los autores estudian operadores de revisión para este modelo, no han estudiado formas de hacerlos computacionalmente tratables, no se han enfocado en la revisión de información potencialmente producida con la intención de engañar, ni han estudiado su aplicación a datos provenientes del mundo real. En nuestra línea de investigación trabajaremos con este grupo en estos aspectos, no sólo enfocados en sus desarrollos anteriores, sino abriendo también la búsqueda a otros formalismos.

$\mathbf{P}_{EM}$:	$origIP(M, X)$	El malware M originó de una dirección de IP perteneciente a X.
	$malwOp(M, O)$	El malware M fue usado en la ciber-operación O.
	$malwPista(M, X)$	El malware M contiene una pista de que fue creado por X.
	$idiomaComp(M, I)$	El malware M fue compilado en un sistema que usa el idioma I.
	$idiomaNat(X, I)$	I es el idioma nativo de X.
	$enConf(X, Y)$	X e Y están en conflicto entre ellos.
	$numInstMCI(X, N)$	En el país X hay al menos N instituciones de primer nivel en matemática, ciencias e ingeniería.
	$infSisGob(X, M)$	Sistemas de gobierno de X fueron infectados con el malware M.
	$edadCapCib(X, N)$	X has tenido capacidades de ciber-guerra por N años o menos.
	$capCibGob(X)$	X tiene un laboratorio gubernamental de ciber-seguridad.
$\mathbf{P}_{AM}$:	$condOp(X, O)$	X condujo la ciber-operación O.
	$evidencia(X, O)$	Hay evidencia de que X condujo la ciber-operación O.
	$motivo(X, Y)$	X tiene un motivo para lanzar un ciber-ataque en contra de Y.
	$esCapaz(X, O)$	X es capaz de conducir una ciber-operación O.
	$objetivo(X, O)$	X fue el objetivo de la ciber-operación O.
	$invMCI(X)$	X tiene inversiones significativas en educación en matemática, ciencia e ingeniería.
	$expCib(X)$	X tiene experiencia en la conducción de ciber-operaciones.

Figura 1: Un ejemplo muy sencillo de un conjunto de predicados que podrían utilizarse para modelar un dominio de ciber-guerra. $\mathbf{P}_{EM}$ corresponden al modelo de entorno (probabilístico), mientras que $\mathbf{P}_{AM}$ contiene los predicados del modelo analítico (de argumentación).

4. Resultados y objetivos

El objetivo general de este plan de trabajo es investigar los aspectos tanto de representación de conocimiento como algorítmicos asociados con los procesos de razonamiento dialéctico y dinámica del conocimiento bajo incertidumbre probabilística. El enfoque particular será en su aplicación en entornos relacionados con ciberseguridad y ciberguerra; dado que se quieren aplicar los resultados en datos provenientes del mundo real, la tratabilidad computacional deberá ser tenida en cuenta a lo largo de todo el proyecto. Uno de los problemas principales que ocurre en ciberseguridad y ciberguerra es el llamado "problema de la atribución", que busca encontrar la parte culpable de un acto de interés en el ciberespacio (como, por ejemplo, un acceso no autorizado a una base de datos) [Tsa12, SSR13].

El objetivo particular posee dos partes principales:

1. El desarrollo de herramientas que combinen formalismos de argumentación estructurada con modelos probabilísticos, con especial enfoque en la capacidad de considerar múltiples fuentes de información que pueden contener información elaborada por enemigos con la meta de engañar al agente razonador; de ahora en más, nos referiremos a ésta como "información engañosa" (*deceptive information*, en inglés). Para ello, se evaluarán – entre otras opciones – diferentes aproximaciones a la revisión de creencias no priorizada. Este objetivo incluye el análisis de las propiedades teóricas de los resultados obtenidos, tales como su complejidad computacional en tiempo y espacio.

2. La evaluación experimental de las herramientas obtenidas en el objetivo (a); habrá dos enfoques particulares: (i) su capacidad para representar información que ocurre en escenarios del mundo real; y (ii) escalabilidad: que el sistema sea capaz de funcionar con una gran cantidad de información en un tiempo razonable; dado que la meta es resolver problemas complejos (tales como la decisión de quienes son responsables por un ataque); un "tiempo razonable", en este contexto, puede ser unas pocas horas. Resaltamos la diferencia entre este tipo de problemas y otros que suelen requerir una respuesta casi inmediata, como una consulta a una base de datos. Los dos objetivos particulares se influenciarán entre sí; el primero se enfoca en los aspectos teóricos, mientras que el segundo se encarga de probar cómo funcionan los desarrollos teóricos en la práctica.

Los pasos iniciales en este proyecto involucran el estudio profundo de sistemas argumentativos y modelos probabilísticos, que son las componentes básicas de los formalismos a desarrollar. A su vez, se estudian las diferentes formas de hacer revisión de creencias, identificando aproximaciones innovadoras que sean propicias para: (i) el modelado de problemas que incluyen información engañosa, y (ii) entornos en los cuales es necesario razonar con incertidumbre probabilística.

5. Formación de Recursos Humanos

Dentro de esta línea de investigación se lleva a cabo la tesis de Doctor en Ciencias de la Computación de José Paredes, bajo la dirección de Marcelo A. Falappa y Gerardo I. Simari, en desarrollo dentro del Laboratorio de Investigación y Desarrollo en Inteligencia Artificial (LIDIA) de la Universidad Nacional del Sur. Actualmente, el LIDIA cuenta con investigadores, becarios y estudiantes de posgrado trabajando intensamente en las áreas de Razonamiento bajo Incertidumbre e Inconsistencia, Web

Semántica, Razonamiento sobre Preferencias, Robótica Cognitiva, Argumentación Rebatible, Revisión de Creencias y Sistemas Multi-agente.

Los directores propuestos han llevado a cabo diferentes proyectos de investigación sobre Sistemas Argumentativos Basados en Reglas y su aplicación en áreas tales como Toma de Decisiones, Robótica Cognitiva y Web Semántica (entre otras). Dentro de estos proyectos se desarrolló una implementación conocida como DeLP (por sus siglas en Inglés de *Defeasible Logic Programming*)[GS04] y hoy cuenta con una versión disponible que permite el uso del sistema a través de la Web[1]. Se ha desarrollado también una versión llamada DeLP-Server que provee un servicio de razonamiento para sistemas multi-agentes donde diferentes agentes en host remotos pueden hacer uso de este servicio de razonamiento argumentativo [GRTS07]. Por último, el director propuesto posee vínculos estrechos con el grupo de investigación del Dr. Paulo Shakarian (Arizona State University, EE.UU.), experto no sólo en temas de ciberseguridad y ciberguerra (es exoficial del ejército estadounidense, incluyendo tareas de analista de inteligencia), sino también en formalismos de razonamiento bajo incertidumbre en general.

Referencias

[AGM85] Carlos E Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *The journal of symbolic logic*, 50(02):510–530, 1985.

[CD05] Hei Chan and Adnan Darwiche. On the revision of probabilistic beliefs using uncertain evidence. *Artificial Intelligence*, 163(1):67–90, 2005.

[DKT09] Phan Minh Dung, Robert A Kowalski, and Francesca Toni. Assumption-based argumentation. In *Argumentation in artificial intelligence*, pages 199–218. Springer, 2009.

[Doy79] Jon Doyle. A truth maintenance system. *Artificial intelligence*, 12(3):231–272, 1979.

[Dub11] Didier Dubois. Information fusion and revision in qualitative and quantitative settings. In *Proc. of ECSQARU*, pages 1–18. Springer, 2011.

[Dun95] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357, 1995.

[FGKIS11] Marcelo A Falappa, Alejandro J García, Gabriele Kern-Isberner, and Guillermo R Simari. On the evolving relation between belief revision and argumentation. *The Knowledge Engineering Review*, 26(01):35–43, 2011.

[FKIS02] Marcelo A Falappa, Gabriele Kern-Isberner, and Guillermo R Simari. Explanations, belief revision and defeasible reasoning. *Artificial Intelligence*, 141(1):1–28, 2002.

[Gär03] Peter Gärdenfors. *Belief revision*, volume 29. Cambridge University Press, 2003.

[GRTS07] Alejandro J García, Nicolás D Rotstein, Mariano Tucat, and Guillermo R Simari. An argumentative reasoning service for deliberative agents. In *Knowledge Science, Engineering and Management*, pages 128–139. Springer, 2007.

[GS04] Alejandro J García and Guillermo R Simari. Defeasible logic programming: An argumentative approach. *Theory and practice of logic programming*, 4(1+ 2):95–138, 2004.

[Han94] Sven Ove Hansson. Kernel contraction. *The Journal of Symbolic Logic*, 59(03):845–859, 1994.

[Han97] Sven Hansson. Semi-revision. *Journal of Applied Non-Classical Logics*, 7(1-2):151–175, 1997.

[HKL00] Rolf Haenni, Jürg Kohlas, and Norbert Lehmann. *Probabilistic argumentation systems*. Springer, 2000.

[Hun12] Anthony Hunter. Some foundations for probabilistic abstract argumentation. pages 117–128, 2012.

[Hun13] Anthony Hunter. A probabilistic approach to modelling uncertain logical arguments. *International Journal of Approximate Reasoning*, 54(1):47–81, 2013.

[KIR04] Gabriele Kern-Isberner and Wilhelm Rödder. Belief revision and information fusion on optimum entropy. *International Journal of Intelligent Systems*, 19(9):837–857, 2004.

[MGS12] Maria Vanina Martinez, Alejandro Javier García, and Guillermo Ricardo Simari. On the use of presumptions in structured defeasible reasoning. In *Proc. of COMMA 2012*, pages 185–196, 2012.

[Pea14] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 2014.

[Pra10] Henry Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1(2):93–124, 2010.

[1] `http://lidia.cs.uns.edu.ar/delp_client/`

[RD06] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine learning*, 62(1-2):107–136, 2006.

[SSF14] Paulo Shakarian, Gerardo I Simari, and Marcelo A Falappa. Belief revision in structured probabilistic argumentation. In *Proc. of FoIKS*, pages 324–343. Springer, 2014.

[SSF15] Gerardo I. Simari, Paulo Shakarian, and Marcelo A. Falappa. A quantitative approach to belief revision in structured probabilistic argumentation. *Annals of Mathematics and Artificial Intelligence*, pages 1–34, 2015.

[SSM+15] Paulo Shakarian, Gerardo I. Simari, Geoffrey Moores, Damon Paulo, Simon Parsons, Marcelo A. Falappa, and Ashkan Aleali. Belief revision in structured probabilistic argumentation: Model and application to cyber security. *Annals of Mathematics and Artificial Intelligence*, pages 1–43, 2015.

[SSMP15] Paulo Shakarian, Gerardo I Simari, Geoffrey Moores, and Simon Parsons. Cyber attribution: An argumentation-based approach. In *Cyber Warfare*, pages 151–171. Springer, 2015.

[SSR13] Paulo Shakarian, Jana Shakarian, and Andrew Ruef. *Introduction to cyber-warfare: A multidisciplinary approach*. Newnes, 2013.

[Thi12] Matthias Thimm. A probabilistic semantics for abstract argumentation. In *Proc. of ECAI*, pages 750–755, 2012.

[Tsa12] Nicholas Tsagourias. Cyber attacks, self-defence and the problem of attribution. *Journal of Conflict and Security Law*, pages 229–244, 2012.