Data has become such a critical resource in many organizations that efficiently accessing the data, sharing it, extracting information from it, and making use of that information has become an urgent need. Hence, it is necessary to protect data and information from unauthorized access as well as from malicious corruption. In recent years, many efforts have researched security informatics, which looks at how to integrate security technologies and information technologies.¹

The advent of the Web in the mid-1990s resulted in an even greater demand for effectively managing data, information, and knowledge. There is now so much data on the Web that managing it with conventional tools is becoming almost impossible. New tools and techniques are needed to effectively manage this data. Therefore, Tim Berners Lee conceived the Semantic Web to provide interoperability as well as to ensure machine-understandable webpages.² Furthermore, as information becomes unmanageable, we need data mining techniques to analyze the information and extract the nuggets often previously unknown.

Our discussion here focuses on this pressing need and how researchers might integrate security technologies with Semantic Web and data mining information technologies. Along this line, we also discuss inference control and policy specifications as well as privacy-preserving data mining techniques.

Semantic Web and Security

Figure 5 illustrates how the Semantic Web and security technologies can benefit from one another. Languages such as XML and the Resource Description Framework (RDF) are being used extensively to represent documents, but appropriate policies have to be enforced on these documents.³ Research on securing XML and XML schemas is looking at how to control access to various portions of the document for reading, browsing, and modifications. To secure RDF, however, we must secure XML, which RDF is based on, and we need security for semantic interpretations. For example, we might want to classify the statements “Tim Berners Lee is the inventor of the Semantic Web” or “H.C. Chen coined the term security informatics.”

//To help readers follow along, the entire Trends & Controversies dept. uses a continuous figure numbering scheme. That is, your first figure is Figure 5 because there are other figures by other authors before yours.//

Figure 5. Intersection of the Semantic Web and Security.

Another Semantic Web security application is in the area of ontology alignment, which helps identify matching concepts across ontologies. We are investigating an approach based on the path difference among concepts in the ontology combined with the approximate privacy-preserving matching technique.³ Path difference consists of comparing the path leading from each concept to the root of the ontology. The more similar these paths are, the more semantically similar the concepts. Because we can compactly code //rewording correct?// these paths as strings, by adopting, for example, some numeric identifiers for concepts, we can replace the corresponding values in the original records. We can then match these modified records by using the approximate privacy-preserving matching technique.³ This matching technique privately computes the distances between pairs of records and returns only the records with a distance that is below a certain threshold.

Semantic Web technologies can also be applied for security problems. The significant advantage of Semantic Web technologies is their representational and reasoning power. Due to the representational power, these technologies can be used not only for representing the data, but also the security policies. For example, XML-based languages such as XACML (eXtensible Access Control Markup Language) have been used to specify
various types of security policies. Researchers have proposed extensions for XACML for finer-grained access control. Furthermore, languages such as RDF are also being explored for representing the policies. Reasoning engines such as Jena and Pellet are also being explored for representing and reasoning about Semantic Web-based policy specifications and determining whether there are security violations through inference. For example, Jena manages RDF graphs, and Pellet reasons with RDF graphs. We have applied these technologies to develop an inference controller.

Data Mining and Security

Figure 6 shows how data mining and security technologies can benefit from each other as well. For example, researchers are applying data mining to problems such as intrusion detection and auditing. Anomaly detection techniques could be used to detect unusual patterns and behaviors. Link analysis could help trace viruses to the perpetrators. Classification might be used to group various cyberattacks and then use the profiles to detect an attack when it occurs. Prediction might help determine potential future attacks using information learned about terrorists through email and phone conversations. Data mining can even be used to analyze weblogs and audit trails.

Data mining can also benefit from security. First, the data mining tasks must be assured. For example, we must ensure that the data mining tasks are not maliciously corrupted. Of more concern are the inference, aggregation, and privacy problems with respect to data mining. For example, data mining gives us associations between entities that are not visible to humans but that might be highly sensitive or private. Data mining might also result in nuggets that the users can utilize to infer highly sensitive information.

With respect to privacy and confidentiality, the challenge is determining how to protect the privacy or confidentiality of the individual data while giving out results. Several privacy-preserving data mining approaches have been proposed to address this. In one approach, the individual data values are perturbed or random values are introduced. The goal is to mine the perturbed data but still get the original results. In another approach called multiparty computation, each party knows its own inputs and the results of mining, but they do not know anything about their partners’ data.

Intersection of Semantic Web, Data Mining, and Security

Figure 7 illustrates the intersection of data mining, security, and the Semantic Web. A document might be represented in XML or RDF. To ensure that Semantic Web mining does not result in security violations via inferences, we need to apply privacy-preserving data mining for XML and RDF data. Semantic Web technologies might also be used to specify policies. Policy consistency can be determined using data mining techniques. That is, we can apply association rule mining techniques to eliminate redundant and inconsistent policies specified in XML or RDF.

Figure 7. Intersection of the Semantic Web, Security, and Data Mining. To protect against security violations via inferences requires applying privacy-preserving data mining for XML and RDF data.

Much work is necessary to prepare the way for the integration of the Semantic Web, data mining, and security. Although some of the technologies such as privacy-preserving data mining are fairly mature, research in areas such as privacy-preserving ontology alignment is just beginning. Privacy-preserving third-party data publication and privacy-preserving information integrations are also fruitful areas of research in security informatics.

Acknowledgments

We thank Elisa Bertino, Elena Ferrari, and Barbara Carminati for collaborating with us on the topics we discussed here. This work was partially supported by grants from the US Air Force Office of Scientific Research. We thank Robert Herklotz for his support.
References


*Bhavani Thuraisingham is the Louis Beecherl Jr., I Distinguished Professor of Computer Science and Director of the Cyber Security Research Center at the University of Texas at Dallas. Contact him at bhavani.thuraisingham@utdallas.edu.*

*Latifur Khan is Associate Professor of Computer Science and Director of the Data Mining/Database Laboratory at the University of Texas at Dallas. Contact him at latifur.khan@utdallas.edu.*

*Murat Kantarcioglu is Assistant Professor and Director of the Data Security and Privacy Laboratory at the University of Texas at Dallas. Contact him at //muratk@utdallas.edu.*