Chapter 7.16
Data Confidentiality on the Semantic Web:
Is There an Inference Problem?

Csilla Farkas
University of South Carolina, USA

ABSTRACT

This chapter investigates the threat of unwanted Semantic Web inferences. We survey the current efforts to detect and remove unwanted inferences, identify research gaps, and recommend future research directions. We begin with a brief overview of Semantic Web technologies and reasoning methods, followed by a description of the inference problem in traditional databases. In the context of the Semantic Web, we study two types of inferences: (1) entailments defined by the formal semantics of the Resource Description Framework (RDF) and the RDF Schema (RDFS) and (2) inferences supported by semantic languages like the Web Ontology Language (OWL). We compare the Semantic Web inferences to the inferences studied in traditional databases. We show that the inference problem exists on the Semantic Web and that existing security methods do not fully prevent indirect data disclosure via inference channels.

INTRODUCTION

The emergence of standardized languages, such as the eXtensible Markup Language (W3C, 2004a), the Resource Description Framework (W3C, 2004b), and the Web Ontology Language (W3C, 2004c), supports automated data management. These languages provide simple syntax and precise semantics that are understandable to both humans and machines. The envisioned Semantic Web (Berners-Lee, Hendler, & Lassila, 2001; Hendler, Berners-Lee, & Miller, 2002) and the applications taking advantage of the Semantic Web will be built upon these languages. A necessary requirement for these future applications is to provide information security and privacy.

Existing security solutions for the Web target specific areas like trust management, secure Web services, access control models for XML, and Web privacy (see Thuraisingham, 2002 for an overview). A promising new research trend aims to incorporate semantics in security models like semantic-aware access control and policy

specification. Although the number of research and development efforts to provide security for the Semantic Web is increasing, only a few researchers consider the inference problem in this context (Farkas & Jajodia, 2002).

Inferences over semantically enhanced data and metadata play a fundamental role on the Semantic Web. Indirect disclosures resulting from the inference capabilities of the Semantic Web are similar to the inference problem studied in statistical and relational databases (Farkas & Jajodia, 2002; Jajodia & Meadows, 1995). However, the characteristics of these two environments differ from the perspectives of (1) data completeness, (2) scope of data control, (3) data models, (4) amount of data (scalability), and (5) data quality. These characteristics affect not only the detection of indirect data accesses but also the applicable removal methods. For example, in traditional databases, removal of an inference channel is usually performed by limiting accesses to data used to derive unwanted inference. However, in the open and decentralized environment of the Semantic Web, some of the data yielding unwanted inferences may be outside of the protected domain. In this case, removal of the inference channel may not be possible by limiting data accesses. New approaches like leakage of misleading information need to be considered.

The goal of this chapter is to evaluate the risk of unwanted inferences in the context of the Semantic Web. Our claim is that the risk of such inferences has increased due to large-scale, semantically enhanced, and automated data processing (Stoica & Farkas, 2002, 2004; Frakas & Stoica, 2003). We compare the inference threat on the Semantic Web to the inference problem studied in traditional databases. We study two types of inferences: (1) entailments defined by the formal semantics of the Resource Description Framework (RDF) and the RDF Schema (RDFS) and (2) inferences supported by semantic languages like the Web Ontology Language (OWL). Indirect data disclosure may occur due to the existence of replicated data with inconsistent security classification and inferences that disclose disallowed data or data associations. Existing access control models that are applicable to Semantic Web data and metadata do not prevent such indirect disclosures, thus are unable to protect against inference-based attacks. Since inferences are considered a fundamental activity on the Semantic Web, we believe that it is necessary to consider their impact on security.

The organization of the chapter is as follows. The Semantic Web section contains a brief overview of the Semantic Web technologies. The Database Inference Problem section presents the inference problem in traditional databases. The Inferences Problem on the Semantic Web section discusses possible inference threats on the Semantic Web, including RDF-based inferences and ontology-driven inferences. The Security Analysis and Future Trends section lists the distinguishing characteristics of the Semantic Web inference problem, outlines prevention methods, and identifies future research areas. The last section, a summary, concludes the chapter.

THE SEMANTIC WEB

This section gives a brief overview of the Semantic Web and Web inference engines. For detailed description of the related concepts, the reader should consult the Web pages of the World Wide Web Consortium (http://www.w3c.org) and the Semantic Web Community Portal (http://www.semanticweb.org).

XML, RDF, and Ontology Languages

The eXtensible Markup Language (XML), XML schema, Resource Description Framework (RDF), and RDF schema are the basic components of the Semantic Web. XML (W3C, 2004a) separates data content from its presentation. XML syntax supports interoperation between heterogeneous domains. Recent research considers XML from