IRM PRESS
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.irm-press.com
ITB11700

This chapter appears in the book, *Web and Information Security*
edited by Elena Ferrari and Bhavani Thuraisingham © 2006, Idea Group Inc.

Chapter IV

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