CERIAS

The Center for Education and Research in Information Assurance and Security

Improving Protocol Vulnerability Discovery via Semantic Interpretation of Textual Specifications

Samuel Jero, Maria Leonor Pacheco, Dan Goldwasser, Cristina Nita-Rotaru
Purdue University and Northeastern University

Objective

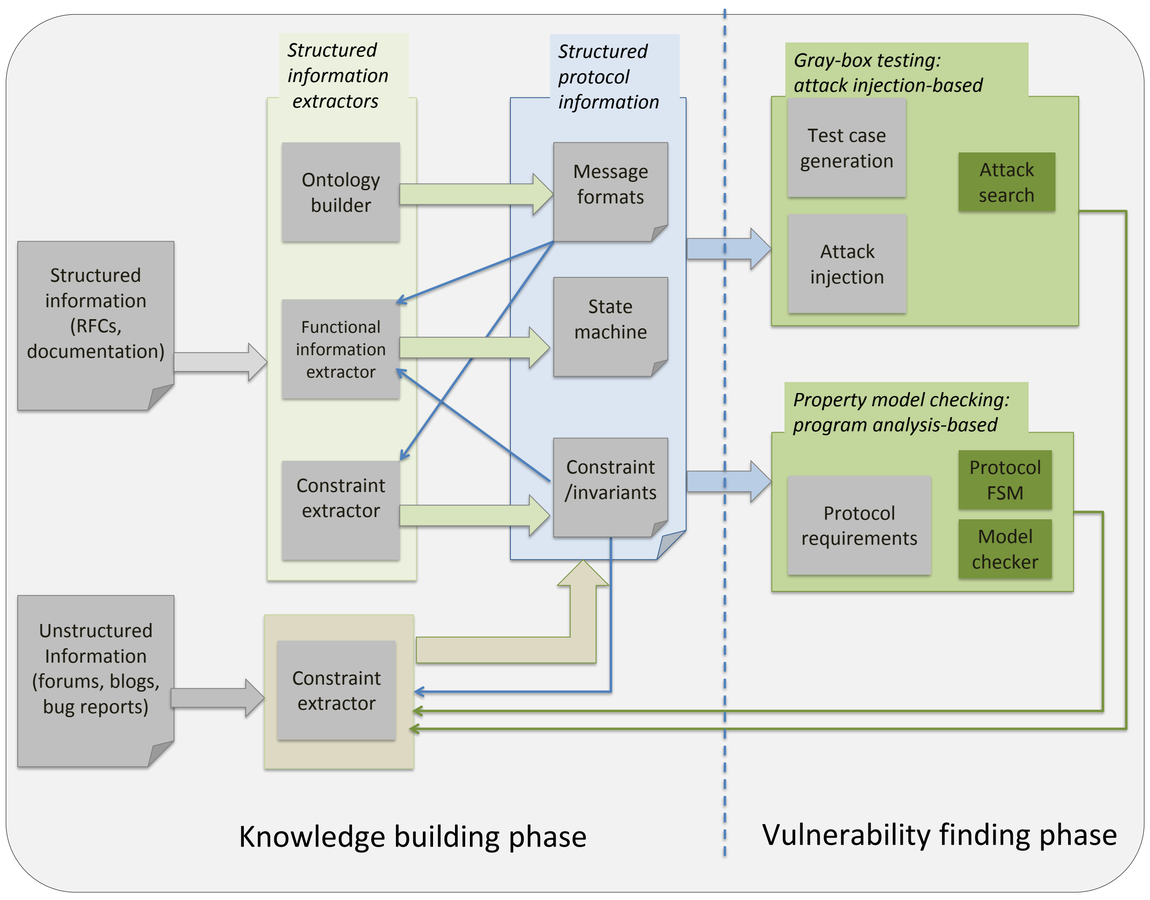
The goal of this project is to automatically analyze the information available in textual specification documents and learn structured information about the protocol, such as message formats, protocol state machines, constraints, invariants and then apply this to automated testing and model checking for protocols

Problem

- Protocol testing requires detailed and structured information such as message formats, state machine, constraints, etc
- Such information is derived manually by people with different levels of expertise, and represents a challenging, time consuming and error-prone task

Opportunities

- Untapped resource of information in the form of RFCs and documentation for many protocols, as well as numerous forums, blogs, and bug reports
- Growing interest in NLP, by the security and privacy research communities, and several applications utilizing NLP tools were constructed



Approach

- Extract structured information from protocol documents using semantic interpretation NLP techniques
- Validate our approach against RFCs, Internet drafts, and other standards documents
- Use the extracted information to close the loop and completely automate protocol testing
- Leverage an existing tool for attacks finding in transport protocols

Current Work

- Focus on RFCs for DCCP, TCP, IP, IPv6
- Bootstrap an entity and relation classifier by abstracting over the specific symbols appearing in the protocol description in RFC documents
- Identify relationships between fields:
  - Identify references to fields by learning string similarity and document structure
  - Two stage classification to detect first any relationship and then specific types of relationships

Details

- Document Preparation: Each document is parsed into a hierarchical structure of sections, which have headers, bodies, and subsections. Annotations are extracted from headers and bodies and tables and images are separated out. The output is a hierarchical structure of sections
- Entity Identification: Creates a list of all the Entities (packet fields) in the document by parsing section headers
- Entity Extraction: Extracts references to Entities using a classifier
- Relationship Extraction: Use a set of key words and build features checking for similarity between chunks and keywords

CERIAS®

PURDUE UNIVERSITY