

Improving Protocol Vulnerability Discovery via Semantic Interpretation of Textual Specifications

Dan Goldwasser (Purdue University), Cristina Nita-Rotaru (Northeastern University)

<http://nds2.ccs.neu.edu/nlp.html>



Objective

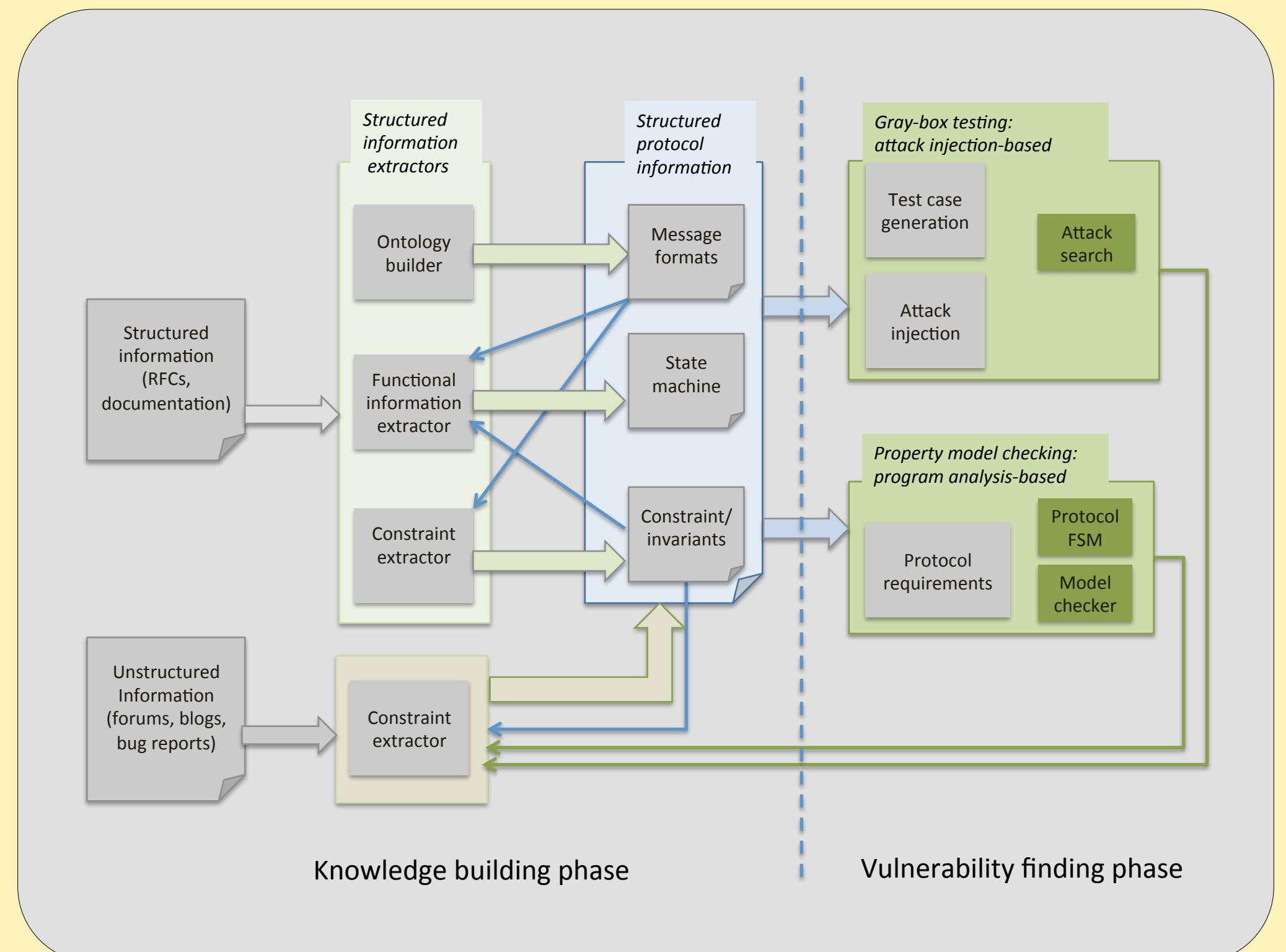
The goal of this project is to automatically analyze the information available in documentations in the form of textual specification and learn structured information about the protocol, such as message formats, protocols state machine, constraints, invariants and apply it to automated testing and model 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 documents 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 standard 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 a protocol description in RFC documents.
- Identify genuine entities from section headers:
 - Identify references to fields by learning string similarity and document structure
 - Differentiate between real and fake entities found in section headers.

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.
- Feature Extraction: Currently 13 features

Interested in meeting the PIs? Attach post-it note below!

