The Center for Cyber Defenders

Expanding computer security knowledge

SAND2019-8210D

PHALANX

Packet Heuristic and Learning Network Examiner

Ryan Adams
University of New Mexico

James Gallagher
Northeastern University

Daniel Krolopp
Purdue University

Project Mentor: John Mulder, Org. 5820

Problem Statement

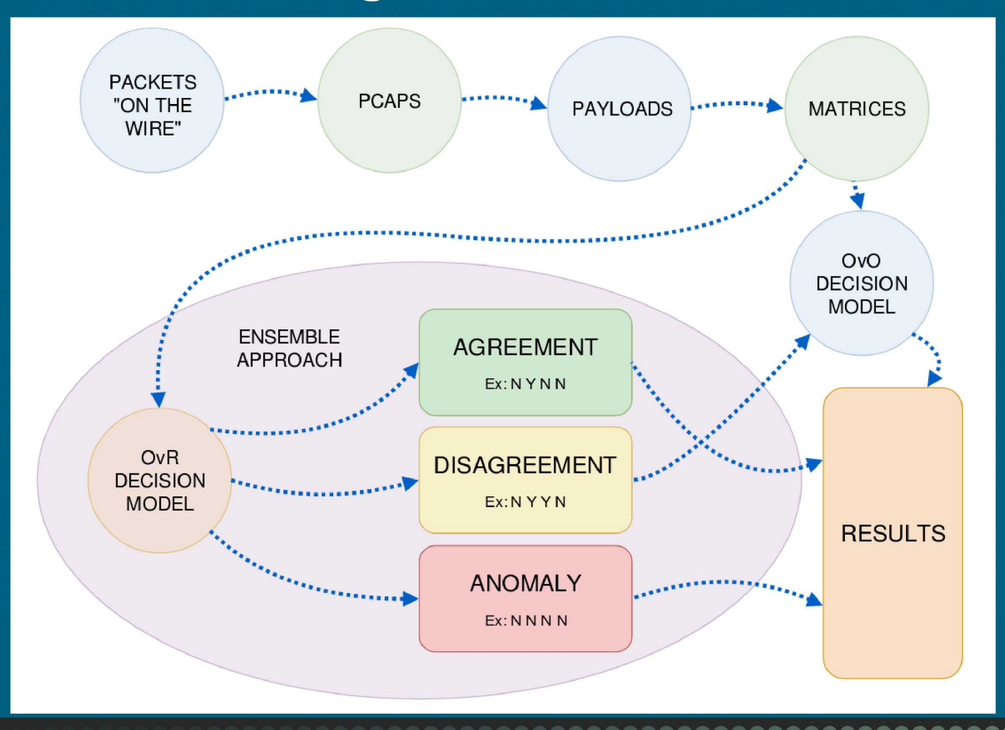
Current technologies, such as Wireshark, rely primarily on predetermined port numbers to identify protocols. However, the address space for protocols is large and varies depending on system configuration making port-specific identification ineffective.

Port Ranges	
System Ports	0 – 1023
User Ports	1024 – 49151
Dynamic and/or private	49152 – 65535
* Internet Assigned Numbers Authority (IANA)	

Most network services use well-known ports identified by IANA, but this is largely convention and not enforced. This can cause difficulty for those tasked with assessing and protecting networks.

Objectives and Approach

- Parse pcaps with known protocols on known ports as truth and extract payloads
- Break payloads into n-grams for use as features in machine learning algorithms
- Train multiple machine learning models
  - One-vs-one (OvO) Packet Identification: Classifying a packet to a known protocol
  - Binary one-vs-rest (OvR) Voting: Yes/No classification voting heuristic
- Use an ensemble approach to identify the protocol, use an OvO model in case of OvR disagreement



Working Environment

- 20,000 data points, 279 protocols
- Primary data source: PCAPS
- Classifiers from SKLearn:

  NB_Bernoulli, NB_multinomial, Decision_tree, Ada_boost, SGD, Random_Forest, MLP, SVM
- Optimized Hyperparameter Tuning using GridsearchCV
- Additionally we are experimenting with Keras to do deep learning classification

Results

Using AdaBoost we managed to achieve an accuracy of 99.90% when identifying Modbus packets against three protocols. We are investigating whether or not this will scale with more protocols.

Confusion Matrix	**Not Modbus**	**Modbus**
Not Modbus	4239	11
Modbus	8	15093

By building a large, curated set of test data and testing a wide range of algorithms and hyper-parameters, we are attempting to provide a formal basis for the eventual tool that will be used in production.

Impact and Benefits

Quickly detecting protocols will allow users to flag anomalous (unallowed or blacklisted protocols) as well as potentially malformed or unrecognizable packets.

If successful, PHALANX will increase the security of Industrial Control Systems used in our nation's critical infrastructure.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND2019-XXXXC.