Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*

StreamSmart

CHALLENGE

Machines think in terms of numbers. Humans excel at communication using language. Cyber defenders face the daunting challenge of processing a massive stream of these "numbers;" finding and ranking problems to investigate; and finally rationalizing their actions to colleagues, leadership, and customers. What if a machine could rationalize its recommendations or summarize the data in English, so they are easy to understand and act upon for human users? With this grand vision in mind, we focus on the extreme-scale stream processing aspects of this challenge and seek to develop a scalable, automated hypothesis generation system for massive cyber data streams.

Streaming hypothesis generation with human-in-the-loop to provide explanations for decisions instead of simply threat indicators of anomalous activity.

CURRENT PRACTICE

Systems such as IBM's Watson showcase the advances made in bridging the human-machine gap, especially in terms of language. However, Watson-like systems excel at conversational settings, where a textual problem specification is provided. These systems are not ready

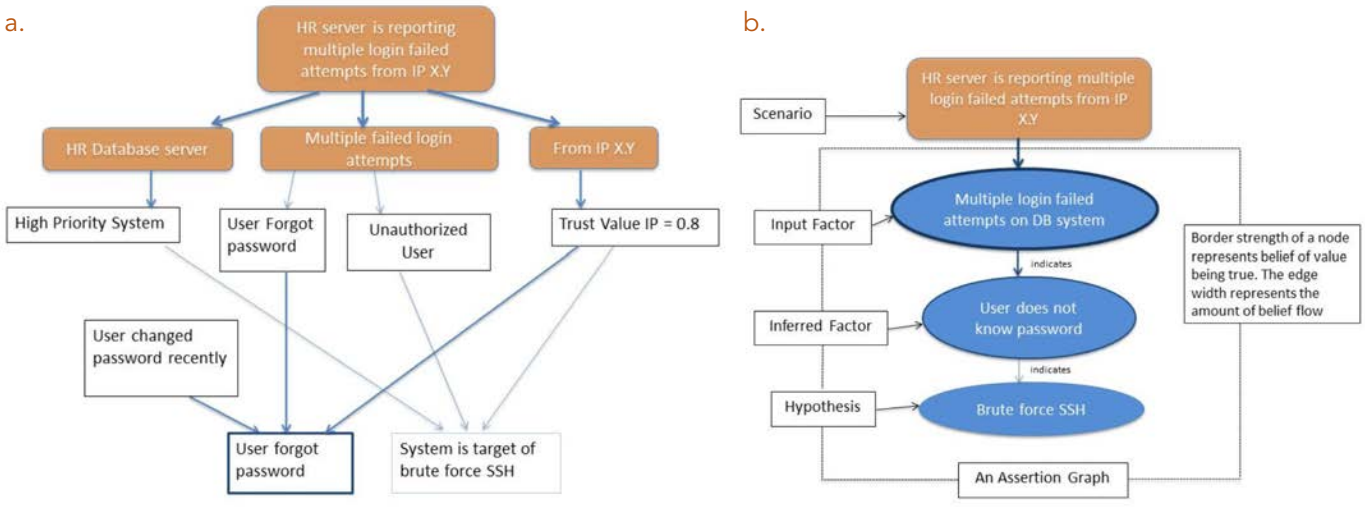

Figure 1a. Example of analyzing a scenario through a series of questions and sub-questions; b. Illustration of hypothesis generation.

for deployment into a full-streaming fashion, where a machine monitors millions of packets flowing every minute, autonomously reasons about attacks or failures in the data, and produces an English explanation for the user.

TECHNICAL APPROACH

Our approach can be divided into three steps.

Problem Summarization. The process we are trying to emulate here is similar to a medical diagnosis, so it must begin with a summarization of the problem.

Take as an example, a medical diagnosis query: a 32-year-old woman with type 1 diabetes mellitus has progressive renal failure. Her hemoglobin concentration is 9 g/dL. A blood smear shows normochromic normocytic cells. What is the problem?

A cyber equivalent of this may be: User A's account had 10 incorrect login events in the past hour. User A changed the password yesterday. Prior to that, User A's account had two critical alerts raised in past 72 hours. What is the problem?

Question Generation. Given a problem summary, the next step will involve question generation. Figure 1a provides an example of questions asked for the cyber example. The next step involves forming sub-questions, shown in the third row with "High Priority System," "User Forgot password," etc. While the boxes display the answers, observe that the resulting what/why/where type of question differs in each case. We will mine the text descriptions associated with various events to learn what questions to ask in the context of a particular event.

Hypothesis Generation. Given the fragments of evidence, we will need to find a coherent chain that binds them together. We will extract a knowledge graph that contains indicative relations between various events (i.e., how the presence of an event or satisfaction of a property indicates the likelihood of another). Finally, we will generate hypotheses via a graph search algorithm that will find high-scoring paths in the knowledge graph that maximally covers the observed events (Figure 1b).

IMPACT

Prioritization and interpretability are two key challenges that cyber defenders face everyday. By orders of magnitude, the daily number of anomalies or attacks on a cyber infrastructure typically outgrows the number of incidents that humans can respond to. Postmortem analysis of most security breaches reveals that most attacks are detected by installed tools but are overlooked by the human experts. Therefore, capabilities such as StreamSmart that extend an algorithm from merely producing a decision to providing an explanation behind the decision will have a disruptive impact for cybersecurity, critical infrastructure, and financial systems.

Contacts

Sutanay Choudhury
Principal Investigator
(509) 375-3978
sutanay.choudhury@pnnl.gov

John R. Johnson
Program Director
(509) 375-2651
John.Johnson@pnnl.gov

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965