How to Use Experience in Cyber Analysis: An Analytical Reasoning Support System

Chen Zhong, Deepak S. Kirubakaran, John Yen, Peng Liu
The Pennsylvania State University
State College, PA USA
{czz111, dok5181, jyen, pliu}@ist.psu.edu

Steve Hutchinson, Hasan Cam
U.S. Army Research Laboratory
Adelphi, MD USA
{Steve.E.Hutchinson.ctr, hasan.cam.civ}@mail.mil

***Abstract*—Cyber analysis is a difficult task for analysts due to huge amounts of noise-abundant monitoring data and increasing complexity of the reasoning tasks. Therefore, experience from experts can provide guidance for analysts' analytical reasoning and contribute to training. Despite its great potential benefits, experience has not been effectively leveraged in the existing reasoning support systems due to the difficulty of elicitation and reuse. To fill the gap, we propose an experience-aided reasoning support system which can automatically capture experts' experience and subsequently guide the novices' reasoning in a step-by-step manner. Drawing on cognitive theory, we model experience as a reasoning process involving "actions", "observations", and "hypotheses". Computability and adaptability are the comparative advantages of this model: the "hypotheses" capture analysts' internal mental reasoning as a black box, while the "actions" and "observations" formally representing the external context and analysts' evidence exploration activities. This paper demonstrates how this system, built on this experience model, can capture and utilize experience effectively.**

Keywords—Experience-aided; Analytical Reasoning

I. INTRODUCTION

As cyber-attack has become a main threat for organizations, cyber analysis is critical to defense against attacks. Analysts need to monitor the audit/alert data generated in an enterprise network, analyze the evidence and draw a conclusion on the current network status. This process is called ***analytical reasoning***. It's goal is to know whether there are malicious activities that have happened in a network, how these attacks are carried out, and what could happen in the future.

One major challenge for analysts is the noise-abundant data. Many Intrusion Detection Systems (IDS) have been developed to help analysts. Most IDSs can monitor the activities of the hosts in a network, examine the traffic and report alerts to analysts. In addition to IDS alerts, typical monitoring data sources includes firewall logs, vulnerability reports, packet dumps, virus reports, system logs and inner-network fileserver logs [1]. The collected data that come from the various sources are overwhelming; just one IDS device could report thousands of alerts per day. Compared to the large amount of data, the data processing capability of a human is quite limited. Besides, IDS devices tend to falsely generate alerts to report benign or regular activities, or fail to capture malicious activities. Another challenge is the increasing complexity of the tasks, especially when multistep attack has become a growing trend. The severity and subtle nature of multistep attacks makes it more difficult to analyze the noisy data, connect the dots and make judgments under tight time pressure.

Due to these difficulties, cyber analysis places high demands on each analyst's capability for data processing and analytical reasoning. On one hand, an efficient reasoning support system is urgently needed to assist analysts in evidence exploration, information correlation, hypothesis maintenance, and reasoning. On the other hand, experience should be fully leveraged in cyber analysis. Expert analysts perform much better than novices because they are well experienced in sense making in attack detection while ignoring irrelevant evidence. However, most of such experience remains untapped, due to the difficulty of eliciting, capturing, sharing and transferring experience knowledge. It's the so-called "knowledge engineering bottleneck". For this reason, few existing tools can utilize experience effectively to facilitate cyber analysis. Analysts typically conduct their analysis tasks as a solitary duty which greatly impedes efficient collaboration. Analyst training also turns out to be a long and arduous process often accomplished only through on-the-job experience.

We propose an experience-aided reasoning support system for cyber analysis. The main motivations for such a system are: (1) to capture and represent experience from experts; (2) to provide novice analysts with step-by-step guidance using the captured experience; (3) to enable analysts to effectively communicate with others to benefit from other analysts' experience. The contribution of this work is mainly two-fold:

- We model experience as a reasoning process involving action, observation and hypothesis. The model makes experience capturing and reusing computational and well adapted to analysts' reasoning which is highly uncertain due to the dynamic cyber environment.
- An experience-aided analytical reasoning support system is developed based on this model to capture experience and provide sequential guidance to analysts.

II. ROOTS IN LITERATURE

An experience-aided reasoning support system must provide two important functions: (1) it must provide a computational representation and mechanism for experience elicitation, utilization, management, and sharing; and (2) it must provide a means to transfer individual experiences (during analytical reasoning) from existing models to build a sharable experience repository for training, communication and reference. We review the research literature in knowledge engineering and analytical reasoning for some comparable works that support these functions.

Supported by ARO Grant W911NF-09-1-0525 (MURI).

A. *Knowledge Engineering for Cyber Analysis*

Logic-based models are widely used to represent experts' knowledge and preferences. One crucial problem in cyber analysis is alert correlation given that IDS alerts are redundant and noisy. Most cyber analysis tools use rules and logical patterns to help analysts group, verify or invalidate alerts [2, 3]. Logical attack graphs can also be generated by logic reasoning based on specified rules [4]. Given a network with known vulnerabilities, a logical attack graph can be easily generated, presenting all possible cyber-attack paths [4]. An edge between two nodes represents a "caused-by" relationship between two vulnerability exploits. By using rules to represent experience-based knowledge, Chen et al. [5] point out that relaxing the conditions of the rules is critical to utilize experience efficiently. However, pattern-based representations are inherently inflexible and many patterns may require exceptions. They also require knowledge to be highly formalized and structured. These limitations reduce the effectiveness of such tools and approaches and render them of little use for cyber analysis.

B. *Experience in Analytical Reasoning*

Research in cognitive science has shown that the human has limited working memory and information processing capabilities [6]. Typically, the large amount of data generated by existing cyber-attack detection tools far exceeds the analysts' cognitive capabilities. Grounded in perceptual and cognitive theory, many visual analytical tools have been developed to facilitate sense-making. Sense-making is the theoretical foundation to achieve understanding from the use of analytical reasoning. It involves information seeking, observation analysis, insight development and result production [7]. Although it's known that experience plays an important role in sense-making, there is not a clear definition of experience in the literature.

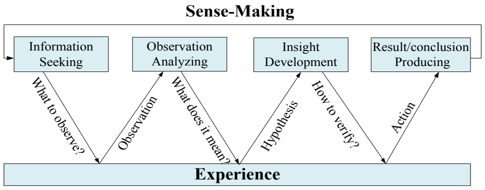

Fig. 1. The role of experience in analytical reasoning process

In the context of cyber analysis, we show that experience facilitates an analyst's sense-making by providing guidance through its four processes illustrated in Fig. 1. These guiding questions include: (1) what data source to look into? (2) what is the implied by the evidence? and (3) how to verify the hypothesis? Unfortunately, most current logic-based representation methods are often unable to capture this rich meaning of experience.

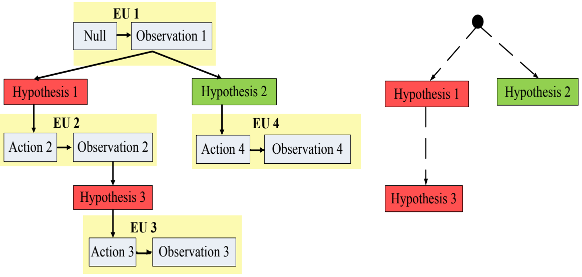

Fig. 3. An example of E-Tree and its H-Tree

III. The Experience-Aided Reasoning Support System

The computing world of cyber analysis consists of three dimensions: analyst, task, and time. This world is used to represent experience and knowledge. A point in this world is the triple: $\boldsymbol{P=(a_m,\ t_n,\ T_t)}$, which refers to the sense-making actions performed by analyst $\boldsymbol{a_m}$ in task $\boldsymbol{t_n}$ at time $\boldsymbol{T_t}$. The upper right of Fig. 2 describes the reasoning process of analyst $\boldsymbol{a}$ *while* performing task $\boldsymbol{t}$ from time $\boldsymbol{T_1}$ to $\boldsymbol{T_2}$.

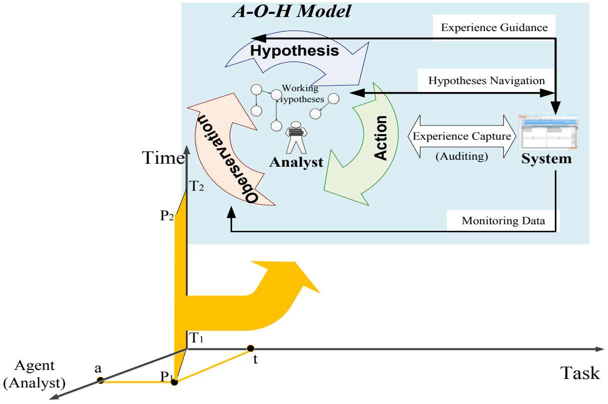

Fig. 2. An analyst works with the experience-aied reasoning support system

A. *The "A-O-H" Experience Model*

We model experience as an analytical reasoning process involving three key cognitive constructs: "Action", "Observation" and "Hypothesis" (***the "A-O-H" model*** in Fig. 2). "Actions" refer to analysts' evidence exploration activities; "observations" refer to the observed data/alerts considered relevant by the analyst; "hypotheses" represent the analyst's awareness and assumptions in a certain situation. These three constructs iterate and form reasoning cycles: the initial trigger could be a suspicious observation (e.g. an IDS alert, denied accesses in firewall logs). This observation may result in new or updated hypotheses (all the hypotheses maintained by an analyst are called "**working hypotheses**"); each hypothesis could trigger further actions to verify or invalidate it (e.g. looking into the vulnerability reports for a web server to check whether an attack on this server could succeed). New actions will lead to new observations; thereby, another "A-O-H" cycle begins. The loop ends when a conclusion is drawn.

B. *Computational Representation of Experience*

For the purpose of computing, "actions" and "observations" in the "A-O-H" model have a structured representation because they are explicit facts. However, considering analysts' mental reasoning is complicated, we keep the A-O-H representation as a "black box" to ensure the adaptability of the model by allowing, "hypotheses" to be represented by free text. We combine each action with its resulting observation into a pair, called an "**Experience Unit (EU)**", to denote the external activities and related contexts. An "**E-Tree**" is constructed to represent the reasoning process by connecting the external reasoning ("EUs") with the internal reasoning ("hypotheses"). The branches connecting an EU with a set of hypotheses illustrate that these disjunctive hypotheses are created in the light of this EU's observation. In order to emphasize the mental reasoning, we further extract the hypotheses from the E-tree to form an "**H-Tree**". A "H-Tree" provides an analyst with a clear hypothesis navigation. An example is shown in Fig. 3 (on the left).

C. *Working Schema of the System*

For ease of data monitoring, the system integrates multiple common monitoring data sources gathered by existing tools, including IDS alerts, vulnerability reports, packet dumps, anti-virus reports, port scanner reports, system logs, inner-network database and fileserver logs. The system works closely with the analyst by providing experience capturing, hypothesis navigation and experience guidance. The working schema of the system is shown in the upper right corner of Fig. 2.

1) Experience Capturing

If the system is informed that the current user is an authorized expert, it captures the analytical reasoning process of the analyst in a non-intrusive way. Each time the analyst examines any data source and specifies the entries of interest, a "EU" will be generated to record the action and its related observation. Once the analyst has a new thought, he/she can create a "hypothesis" and describe it in free-text to the system. This new "hypothesis" will be automatically linked to its corresponding "EU" to construct the E-Tree and H-Tree.

2) Hypotheses Navigation

The H-Tree maintains the relationships among working hypotheses, as well as their contents. In this way, it prevents the analyst from getting lost in his/her thought progression. Since each hypothesis is created under a particular context (i.e. its ancestor "EU"s), whenever an analyst selects one hypothesis to work on, the system can automatically transfer the current context to the corresponding context of the selected hypothesis. Analysts can also easily manipulate any working hypothesis, for example, by changing the content or marking it as True or False.

3) Experience Guidance

Maintaining an experience base, the system can provide analysts with timely made-to-measure guidance. The system keeps track of the analyst's observations and actions. In the light of a current observation, it retrieves previous experiences with similar observation(s) and presents them to the analyst for reference. Usually, given an observation, analysts make hypotheses based on their intuition or inference. Our system provides them with multiple options, because they can learn what other experts did, namely how they made hypotheses facing a similar situation in the past. They may agree and follow it, or depart from the system-provided 'guidance' and make their own decisions. Whenever any action is taken that results in a new observation, the system will recall relevant experience based on the new context and provide updated suggestions.

IV. Case Study

We conducted a pilot study to test whether the system works well and help cyber analysis. The system is implemented by 5120 lines of C# code. Fig. 4 shows its interface. Analysts are categorized by their expertise. Only the experts' experience is captured. The experience base for this pilot contains 31 pieces of experience that were elicited from two researchers from in our cyber-security lab. The test bed is shown on the right below of Fig. 4 and contains a web server (MS IIS), a mail server (MS Exchange Server), a DNS server (Linux) in the DMZ, five workstation PCs (Win XP SP3), a database (Oracle/Linux) and a file server (Linux) in the internal network. Two IDSs (Snort) are deployed in both the DMZ and internal networks. Eight sets of monitoring data are collected by launching two multistep attacks (on right below in Fig. 4) on the test bed [1]. The attack chains are: (1) Scenario 1: PC2->Mailserver->PC5; (2) Scenario 2: Webserver->PC3->Database. We recruited two graduate students from the Penn State security lab as the subjects, called Subject 1 and Subject 2. Given the system integrating the data, both subjects were asked to detect the attack chain in both scenarios. According to our pre-task survey, Subject 1 has experience in mail server attack analysis while Subject 2 is more familiar with database attacks. Therefore, we let the system capture Subject 1's experience in Scenario 1 and Subject 2's experience in Scenario 2. While the subjects were performing these tasks, we also manually recorded their reasoning processes using a "think-aloud" method. We repeated the study three times (with at least one week intervening time interval).

Compared with the reasoning process recorded by "think-aloud", the experience automatically captured by system (represented as "E-Tree" and "H-Tree") reflected the analyst's real reasoning process. The average number of nodes ("EU"s and "hypotheses") in the E-Trees is 27.67 for scenario 1 and 18.67 for scenario 2. According to the post-task questionnaire, our system can efficiently capture experts' experience without disrupting their analytical reasoning process. Regarding whether past experiences can guide current analytical reasoning, most of the suggested experiences (E-Trees) were found to be helpful to allow them to make decisions. In addition, the context-based experience retrieval was also found to be efficient in time. The results of this pilot study will help us prepare future, large-scale experiments.

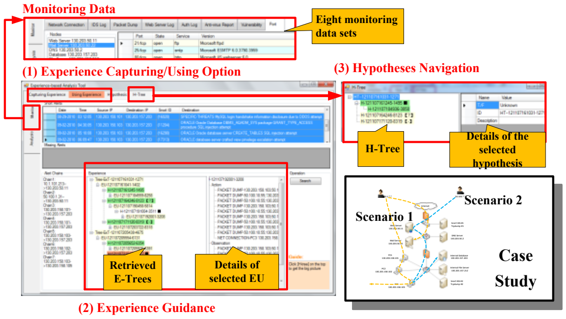

Fig. 4. The experience-aided reasoning support system with three key functions. The attack chains are explained in Case Study section.

References

[1] Giacobe, N. A. (2012). Data fusion in cyber security: first order entity extraction from common cyber data. In Proc. of SPIE Vol (Vol. 8408).

[2] Morin, B., Mé, L., Debar, H., & Ducassé, M. (2009). A logic-based model to support alert correlation in intrusion detection. Information Fusion, 10(4), 285-299.

[3] Tabia, K., Benferhat, S., Leray, P., & Mé, L. (2011). Alert correlation in intrusion detection: Combining AI-based approaches for exploiting security operators' knowledge and preferences. In the third IJCAI-11 Workshop on Intelligent Security (SECART-11), (pp. 42-49).

[4] Ou, X., Boyer, W. F., & McQueen, M. A. (2006). A scalable approach to attack graph generation. In Proc. of the 13th ACM CCS (pp. 336-345).

[5] Chen, P. C., Liu, P., Yen, J., & Mullen, T. (2012). Experience-based cyber situation recognition using relaxable logic patterns. In CogSIMA, IEEE International Multi-Disciplinary Conference on (pp. 243-250).

[6] Yen, J., McNeese, M., Mullen, T., Hall, D., Fan, X., & Liu, P. (2010). Rpd-based hypothesis reasoning for cyber situation awareness. Cyber Situational Awareness, 39-49.

[7] Pirolli, P., & Card, S. (2005). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In Proc. of International Conference on Intelligence Analysis.

[1] The data set is collected by S.Oh published in http://yenlab5.ist.psu.edu/cybersa/.