XMeta: a Bayesian approach for computer forensics

Thomas DUVAL (DGA/CELAR, Supélec) thomas.duval@supelec.fr
Bernard JOUGA (Supélec) bernard.jouga@supelec.fr
Laurent ROGER (DGA/CELAR) laurent.roger@dga.defense.gouv.fr
Supélec: Avenue de la boulaie, BP 81127 F-35511 CESSON SEVIGNE FRANCE
DGA / CELAR: route de Laillé, F-35170 BRUZ FRANCE

November 22, 2004

keywords: *Computer forensics, digital evidence, computer crime, Bayesian networks.*

Introduction

"Is the attacker still inside ?", "Where did he go ?", "How did he break in ?", etc. Even if information systems are more and more secured, computer criminality increases [1], and the need for forensics analysis increases too. Computer Forensic helps administrators and investigators to find what happened, how it has been done and who has broken in. In this paper, we propose a system which helps investigators in their inquiries by giving them some assumptions about the crime scene. The first section presents the main issues of computer forensics. The second section summarizes the use of Bayesian networks in expert systems, then in the third section we present our proposal. Endly, we conclude with future works.

1 Computer Forensics

When arriving on the *crime scene*, the investigators first job is to collect the maximum of information about the attacked organization and the compromised systems. The objectives are to learn about the organization activity, the technical architecture, the reported damages, the involved persons, and to collect and preserve technical materials. One of the first essential tasks is to record volatile system variables and to make images of (non-)removable media. After the first observations, the investigators come back to the lab with the collected devices that need deeper analysis. Important elements are: presence of specific data or softwares, configuration files, log files on hosts communication and security equipments, hidden or deleted files, etc.

Computers being connected with networks in an open environment, the investigators must then jump from one compromised host to another, in order to traceback the attack. Besides these technical investigations, other information such as the context of the accident and testimonies are also important.

Endly, investigators must write a report with all the gathered observations and evidence. This report must be clear enough to be understood by lawyers (they may not understand some technical details), but also complete in order to accurately record the steps of the investigation.

2 Bayesian networks

Bayesian networks (BN) have been already widely used in expert systems (SpamAssassin [2]) and for some forensics applications ([3]) BN are directed acyclic graphs where nodes are variables and links are causal connections weigthed with conditional probabilities. A Bayesian Network can model situations of which the perception is partial. For example, an administrator notices that his Web Server W has been defaced. There are only two causes (in this example) for this situation:

- the attacker used a exploit E,
- the attacker stole the administrator's password S.

A well-formed Bayesian network will propose an answer in a probabilist format: There is X chance in a hundred to perform the exploit E on a Web Server W, the probability value been given by the Bayes Theorem:

$$P(E \mid W) = \frac{P(E,W)}{P(E)}$$

with:

- $P(E \mid W)$ the probability to perform the exploit E knowing that we have a Web Server W,
- $P(E,W)$ the probability to have an exploit E and a Web server W,
- $P(E)$ the probability to have an exploit E.

$P(E,W)$ and $P(E)$ can be estimated thanks to a knowledge base recording the results of past inquiries.

The Bayesian approach has been chosen for three reasons: 1) it can model complex situations, 2) it offers a non-binary (probabilistic) logic, 3) it is resistant when data are unknown or unsure.

The construction of a Bayesian Network follows three steps: [a] construction of a causal graph, [b] construction of probability tables associated to nodes, [c] propagation of probabilities in the network. At this step, the network is ready to be used by investigators.

3 Our proposal

When arriving on a host, an investigator enters the host configuration (**TH** -Targeted Hardwares- and **TS** -Targeted Softwares-) figure 1, and the observed losses

(**RD** -Reported Damages-). These information are used to dynamically build a Bayesian network from the knowledge database of previous cases. Inference on this network produces some clues about the compromission (**GA** -Generic Attacks-, **AA** -Additional Actions-, **SA** -Source Address-[1]), and the best forensic technics to use (**IT** -Investigation Technics-). When the investigator checks a particular attack or action, he can set it into the system which change its values thanks to this new fact.

When an host has been entirely checked:

1. the source attack address is local, then the investigation is achieved,

2. the source attack address is not local (internal or external) and then the investigators could eventually "jump" to a new host if accessible (ie. in the same organisation). In this case, a new investigation plan is created for this host.

Investigators can create as much investigation plans as needed (multiple links are allowed).

4 Testbed

A testbed, XMeta, has been developped, using "Bayesian Tools in Java" [4] for inference and a Python based GUI. The initialization of the inference system should be done from previous investigations, but at this time we miss real data. In a first step, we used the ICAT vulnerability database [5]. This database gives us relations between softwares (more than 6000), attacks, actions and losses. The only facts that cannot be extracted from ICAT are investigation technics. So, the version of our testbed only gives the most vulnerable softwares and attacks and actions of the attacker(s).

For example here is a (fictive) case where confidential information has been stolen from an workstation belonging to a secretary. The compromised host was a Linux Debian system (with no patch), the inference system has been initialized with Debian softwares (kernel, libc, Windowmaker, OpenSSH, etc.) and confidentiality (for ICAT loss type). Our system answers that the source of the attack was probably local to the machine (it was true) and the most vulnerable softwares were XFree86, the Linux libc and the kernel (the real compromised software was the kernel). Endly, the attacker used an exploit to become root and to copy the stolen file, this attack was in the 7th position in the answer (over 21 possible attacks). In the case of a real investigation, all the attacks should have been checked. Depending on the context, some of them could be quickly put away (for example, a DOS attack is non sense for a data theft).

After this case study, the database of our inference system has been extended with new records (in particular with investigation technics) and it gives now better results.

[1] the address can be local: the same host, internal: the same network or external

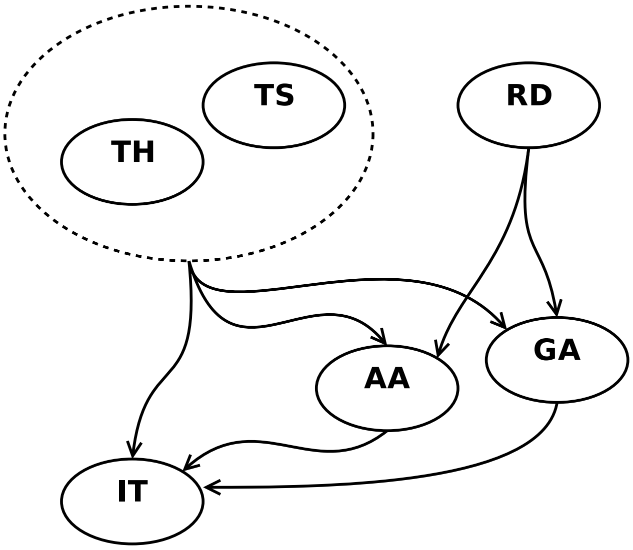

Figure 1: An investigation plan

Conclusion and future works

XMeta can help investigators to find some clues about a particular case. But an effort must be done to increase our database in order to have more accurate results (in particular for investigation technics). We work also on an other way to make assumptions more accurate, with the creation of a criminal profile which can help to find the preferred softwares of an attacker or his preferred attacks and actions on a system.

References

[1] CCIPS. Computer Crime and Intellectual Property Section of the Criminal Division of the US Department of Justice. http://cybercrime.gov.

[2] Spamassassin. http://spamassassin.apache.org. An extensible email filter used to identify spam.

[3] T.S. Levitt and K.B. Laskey. Computational inference for evidential reasoning in support of judicial proof. *Cardozo Law Review*, 2004.

[4] William H. Hsu. Bnj - bayesian network tools in java. http://bndev.sourceforge.net.

[5] Icat vulnerability database . http://icat.nist.gov/icat.cfm.