USING THUMBPRINTS TO TRACE INTRUDERS

STUART STANIFORD-CHEN
stanifor@cs.ucdavis.edu

May 7, 1995

This document is a short overview of the work we are doing to develop and apply thumbprinting to the problem of tracing intruders across the network.

1 The Problem

At present, it is extremely difficult to trace intruders across the internet to their point of origin. Several factors go into this. The main reason is that many sites have poor security and intruders are able to amass a collection of accounts on such sites. They then use these to mask their origin by logging into a series of them in turn before attacking a target site. We refer to such a sequence of logins as an extended connection, or a connection chain.

The result of this is that it is extremely difficult to hold intruders accountable for their actions.

2 Previous Work

In the DIDS project, UC Davis developed a distributed intrusion detection system. One of its main functions was to unravel the connection chains. When a user logged from one machine to another in the DIDS domain, the system accounted activity on both hosts to a single NID (network identifier). This was done by examining the audit trails on the hosts in question.

A problem for DIDS was that if a connection chain disappeared outside the domain of operation of DIDS but then re-entered it, it became impossible to correlate those two pieces of the chain.

To solve this problem, the designers proposed thumbprinting.

3 Thumbprinting

The idea behind thumbprinting is to notice that the content of the connections at different points in the chain are the same - if an *ls* command is given by the attacker, those two characters (and their echoes) traverse the entire length of the chain. Similarly, the target machine's response travels through every connection involved. Thus, if small summaries of the content of the connections were kept at two points, they could later be compared to establish whether or not those two formed links in the same chain. One can think of such a summary as a checksum of the data in the connection (or some piece of it). Existing checksum algorithms are not suitable however. In practice these summaries are formed by passive monitoring of network traffic.

Possible uses of this technology include:

1) In the DIDS context, thumbprinting could allow the system to relate activity which went outside the domain but then re-entered. This might be important when an inside attacker was seeking to disguise himself as an outsider.

2) Thumbprinting systems could be placed at the places were a network for some site touched other networks. This would allow the administrators of that site to determine whenever their systems were being used as a pass-through site. (Bob Pallasek of Livermore suggested this application).

3) Law enforcement could use this technology at a variety of places they thought likely in order to track down exactly where an attacker was coming from as part of an investigation.

4) This technology could be a useful component in a general internet tracing system (akin to the trap-and-trace facility provided by the phone networks).

The main vulnerability of the technology is in its basic design principle. It assumes that the content of the connections at all points in the chain are the same. Hence, if that content is encrypted differently in each connection, the method fails. In a world with routine encryption of connections, the method would thus be useless. Such a world seems to be a long time coming however. Hackers could also develop tools to help them encrypt their connections themselves. At present, this could be detected and might itself be cause for suspicion.

4 Work to date

We developed criteria which a thumbprinting technology has to satisfy. Besides the obvious ones of small size and easy calculation and comparison, we noted that thumbprints need to be robust. This means that when the data from which the thumbprint is formed changes slightly, the thumbprint itself should change only slightly. This is necessary because passive monitoring systems do not reliably capture every packet.

Another necessary property is additivity. We thumbprint on small time intervals (tens of seconds). Because of propagation delays, data which appears in one time interval in one place can be in a different time interval in a second place. This can ruin the thumbprint comparison for both time intervals In this case, it is useful to combine the thumbprints for the two time intervals into a single thumbprint which will still be valid.

The thumbprints we use are based on linear combinations of the frequencies with which particular characters appear in the connection content. We choose the optimal linear combinations using a statistical method known as principal component analysis. The resulting thumbprints have all the desired properties. We developed an algorithm for comparing the thumbprints and recombining them additively where necessary. We use this to assign a statistical significance level to any particular comparison - when the comparison is very significant, it is very probable that the two thumbprints are based on the same content.

We have built a system to record these thumbprints, and we did experiments in which we recorded activity that was going from our systems to remote sites in Georgia and in England and back again. We then compared the thumbprints at the beginning and end of the chain. These trials were very successful and showed that the system was capable of comparing connections with very few errors.

5 Ongoing work

We are still doing experiments to optimize our techniques and to characterize under what circumstances they work well, and where they have difficulty. We are also developing a research prototype system capable of doing actual tracing of intruders using this technology. This work will be completed in the next few months.

Ultimately, we plan to incorporate this technology into the wide-area intrusion detection system we are developing for ARPA. We are also looking for people to collaborate with who could benefit from using this technology in the field. We would like to get feedback on what is needed in a realistic setting. Possible partners include the Air Force Office of Information Warfare, and Lawrence Livermore Laboratories.