# Identifying Reconnaissance Activity: A Strategy for Network Defence

## Abstract

Over recent years there has been a massive increase in the need to build stronger and more effective defensive systems in many contexts due to the amplified threat of terrorism. Network security is no exception to this increased need to secure systems against attack. One area within network security which has received a heightened interest is the correlation of reconnaissance activities, rather than merely blocking source addresses without further investigation. This paper will examine the work to date and detail how various researchers have approached the correlation of network scan activity. After the summary of existing research this paper will then detail new work we are undertaking in this field, using clustering techniques in conjunction with a peer to peer network, to correlate port scan activity in real-time.

**Keywords**

Data Correlation, Attack Detection, Network Scans

## INTRODUCTION

Security as a topic in general has increased in prominence over recent years in reaction to the threat of terrorism. In the space of several years there has been a proliferation of laws, alerts and suggested “do’s” and “don’ts” in an effort to increase security. The bulk of these changes, and indeed the terrorist threat, are well removed from the realms of computers and networks. However, there is still a valid concern about the risk of terrorists using networks and the computers connected to them as a tool with which to commit or more likely assist an attack.

Cyber terrorism is not at present a large concern. It is far easier to get a job at a major power plant and then switch it off or blow it up manually, then to hack in and affect its operation remotely (Schneier 2000). The threat is nothing like what is often portrayed in the media or television programs such as the 24 where agent Jack Baur battles terrorists who ambiguously “hack the internet”. However, it can only be realistically seen as a problem that will grow with time. Since World War 2 we have witnessed a rapid increase in the number of computer systems and networks and in their level of interconnectivity, and there is no cause to believe that this will not continue. As this continued growth occurs it is a logical conclusion that activities carried out by terrorist groups will increase.

Mitnick writes in The Art of Intrusion (Mitnick & Simon 2005) about how two young American hackers were contacted by a person calling himself Khalid, who claimed to be a member of the Harkat-ul-Mujahideen organisation. Since 1995 that organisation was listed as a terrorist group by the U.S State Department for its links to Osama Bin Laden. Khalid offered them money for Boeing plane specifications and asked another hacker for maps of U.S government networks. Some of his requests were filled, and some of the payments were sent. Eventually the FBI swooped on the hackers, although not until after the Boeing plans were delivered, and a plane was hijacked in India with Khalid claming some involvement. Khalid still has not been apprehended.

Closer to the Australian security context, convicted Bali bomber Imam Samudra has written a book from his gaol cell entitled "Me Against the Terrorist!” (Samudra 2004). In the book he highlights the value of cyber crime in the pursuit of Jihad in the aptly entitled chapter “Hacking, Why Not?”. Samudra concisely details a how-to-guide of cyber crime, focusing on the theft of personal information to gain finances to support more traditional terrorist activities. The online hacking resources he describes are all, with the exception of one, non-Muslim sources of information (Sipress 2004). He, like Khalid, draws strongly upon the existing, easily accessible, knowledgebase found in western target countries.

While these threats can motivate an increased fervour to secure systems, the threat of system breaches across networks will be likely to remain predominately the traditional perpetrators (or their techniques) for the foreseeable future. The Khalid attack, for example, was carried out by American teenagers. Therefore, there is little need to re-focus network security in response to terrorism, but instead we should continue to strengthen the systems we have, defending against the same threats that we have faced for the last few decades.

This paper will examine data mining in the context of defence, and specifically the network context. The paper will then examine the existing scan correlation systems, before discussing the proposed work.

### Date Mining in Defence

One of the most talked about domestic countermeasures to terrorism and related activities over the last few years has centered on the use of electronic profiling in airports and borders. Electronic profiling employs the use of data mining techniques to attempt to draw together various pieces of data to extract a probability that a given person is a terrorist, and not a harmless traveller detecting a virtual "needle in a haystack". Likewise data mining methods have been proposed, and more than likely implemented, to examine intercepted email traffic, in conjunction with the already automated mining powers of phone call interception provided by systems such as Echelon (Schneier 2000). Data mining of terror-related website activity is also a tool that is being employed (Shapira et al. 2003).

Data mining in the network security context is in many ways very similar to airline passenger profiling. There is again a "needle in a haystack" situation, with the vast majority of traffic and activity being totally benign. It is at this point that scan correlation systems (IDS) can be useful, attempting to mine the useful information out of the masses of audit log data and intercepted network traffic.

## INTRUSION DETECTION AND SCAN CORRELATION

Over the past 20 years since Denning (1987) first proposed an automated real-time intrusion detection system, the monitoring of systems and network activity has enjoyed a rapid evolution from the manual monitoring of audit log files during the 1980's' through to the complex multi-agent AI driven autonomous detection systems present today. The systems that are currently deployed frequently involve amalgamating data from multiple sources and extracting the relevant data to match an attack signature, or behaviour which does not fit a profile of normal activity. The techniques and methods with which this is done have grown in complexity and effectiveness, now detecting attacks that have multiple components sometimes lasting days and coming from multiple source IP addresses. It is these challenges that the relatively new area of scan correlation (previously largely ignored until after a more overt measure was taken) now also faces. As such, researchers in this area can draw upon the previous work completed in Intrusion Detection.

### Usefulness of Scans

Every IP address gets scanned. There are a finite number of IP addresses; it is the way that network addressing was designed. To re-visit the often used analogy of a hacker being like a burglar breaking into a computer instead of a home. Imagine a burglar who has access to an address book of every house in the world, and can find out a few details such as whether anyone lives there and what the alarm system is probably running without leaving the relative safety of his own home. This is why every computer gets scanned. Hackers (whether motivated by terrorism or not) have an address book of possible locations, and with a handful of scans that last only a few seconds they can discover whether a computer is at the address and what services it is running.

There are many readily available tools which allow for various automated scans to be completed at the click of a button. The most commonly used probing utility which is used both by system administrators and malicious users alike is called nmap (fyodor 2005). Nmap allows for a wide range of different scans to be completed over various IP ranges or lengths of time. It comes equipped with over of 700 different recognisable operating system finger prints to match to scan results. Figure 1 depicts a scan that was observed in our audit logs of a user who scanned each host in an entire Class C address space twice in the space of 60 seconds using a tool like nmap. Furthermore, it also allows for various scans that guarantee anonymity such as idle scans. An idle scan involves using a second computer as an intermediary to hide the identity of the scanner, without actually having control of the second computer (Zalewski 2005).

## SCAN CORRELATION SYSTEMS TO DATE

### Heuristic Threshold Based Systems

During 2004 two Scan correlation systems were built and published, one by two of the authors of this paper (), and a second by Jung et al (2004). Both of these systems employ a scanner detection technique that makes use of threshold based heuristic.

Jung et al's system focused not only on detecting scanners, but also classifying benign scans correctly to avoid false positive classifications from scan activity. The algorithm they implemented was based on the mathematical technique called Sequential Hypotheses Testing described by Wald (1947). Each IP's activity was then classified as being either benign or malicious, depending on which hypotheses threshold was reached.

The work carried out by author et al () was focused differently then the Jung system. The correlation that occurred in the system was spread across multiple data sources, with the aim of detecting scanners who were interested in more then a single gateway upon a network. Source IP activity was monitored, and those users who

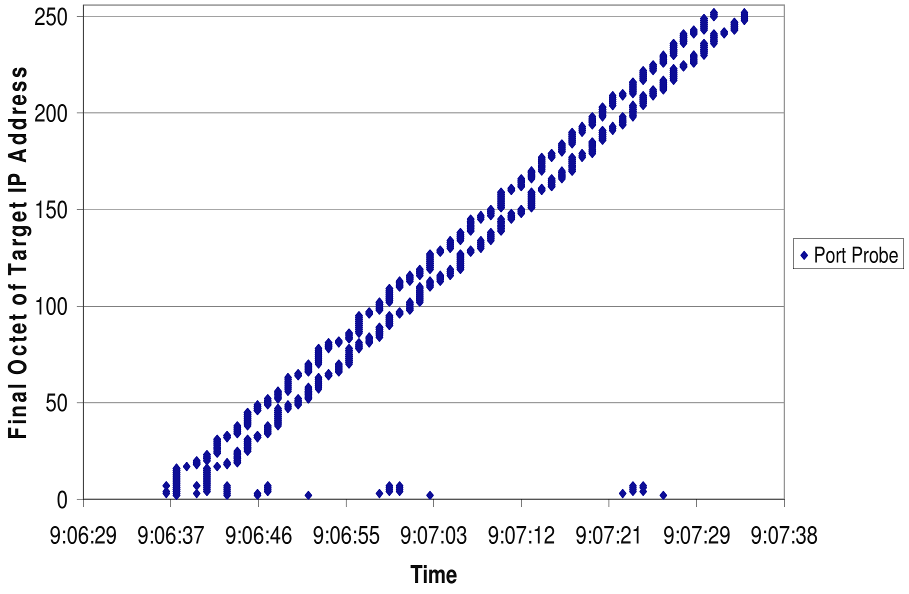

**Figure 1. An observed Host scan across an entire Class C address space.**

did not access multiple gateways before a given threshold were deemed to be not going to access more then a single gateway and could therefore be ignored. Those who were accessing multiple gateways were then blocks across all the monitored gateways, even those they had not yet attempted to access for a period of time.

Both of these systems demonstrate that effective countermeasures can be put in place to hinder scan activities usefulness as a reconnaissance tool.

### Spade and Spice: Simulated Annealing

The third system to be described here is the one closest to the proposed work in this paper. The previous two systems have produced results that will be worth comparing to ours, but are limited in scope and overall functionality. This third system was being built with Defense Advanced Research Projects Agency (DARPA) funding in the U.S by the Silicon Defense group; however, in the fallout from 9/11, lost their funding due to its classification status.

The work that had been proposed, and partially implemented, by Stainford et al (2002 ) was intended to correlate scan activity and would not only classify the user as a scanner, but also be able to link it to past scan activity. The goal of such a connection would be to link users who operate over a long time under multiple source IP addresses to avoid detection. This is precisely what is required in a system that identifies and tracks reconnaissance activity.

The proposed system involved Spade (sensor) feeding events into Spice (correlation engine), along with an anomaly score generated based on the source IPs activity (the negative log of the probability of the event occurring). Spice then places the event in a graph noting the various properties of the event such as source IP, target IP, target and source port and time. The location at which the item is placed into the graph is found through the use of the search algorithm Simulated Annealing. Prior to each event being added, the graph is searched to locate the best location for it, placing it near events of a similar nature. It is here that past scans and events can then be correlated to determine whether a source IP has been changed. It could also be used to decide on the correct response to the user’s activity. However, as mentioned above, Spice was never implemented.

## PROPOSED WORK

The proposed system builds upon the threshold based system previously made by the authors conceptually, but will be re-implemented from scratch using similar system architecture as that of the Spice and Spade system discussed above. The goals of our system are as follows:

- to correlate data from multiple sources to preserve network context,
- to present no single point of failure
- to transport the needed data between locations efficiently,

- to detect which scans are malicious in nature, and also whether they are multiple gateway based attacks,
- to detect users who change source IP addresses over time.

The goals fall into two main categories: correlation and data distribution.

### Correlation

The correlation that has occurred in the two completed systems is relatively simple, making use of basic thresholds. Certainly they are very useful, but the proposed system will be designed to achieve much more. It should match multiple source IPs to a single scanner. Such a correlation process requires much more storage and computation than a threshold-based system would.

As already stated, the system design takes much of its inspiration from the Spice. However, we do not use the Simulated Annealing algorithm used in Spice. The system needs to run as fast as possible, operating as close to real-time as is computationally feasible. Therefore instead of opting for the partially random graph based Simulated Annealing algorithm, our implementation will make use of a modified version of the CLASSIT algorithm.

CLASSIT (Gennari, Langley & Fisher 1989) is a conceptual clustering algorithm, which operates incrementally building knowledge over time. As items are added to the model, groups are formed through the clustering of similar items together based on their traits. CLASSIT is largely an extension of an existing model called COBWEB (Fisher 1987 ) which dealt with nominal attributes, while CLASSIT can handle real values. In the case of CLASSIT the items are stored in a hierarchical tree. Each clustered group is then deemed a concept, which may then be used to classify those instances. Gennari et al (Gennari, Langley & Fisher 1989 ) define the goals of incremental conceptual clustering in the following way:

> "- Given: a sequential presentation of instances and their associated descriptions;
>
> - Find: clusterings that group those instances in categories;
>
> - Find: an intentional definition for each category that summaries its instances;
>
> - Find: a hierarchical organisation for those categories."

In our system conceptual clustering is intended to occur to group scanners with similar behaviour pattern together. The traits that will be clustered upon will include such things as target and source ports and addresses, timing of scans and frequency of activity. The intention is that such groups may then represent which scans are malicious in nature, and which are not, while also categorising those in these two broad groups into sub groups such as multipoint scanners, and possibly those who change source IP between attacks.

As the source IP will only be one possible trait, and probably not the main trait for the classification, updating a specific user's activity does pose a sizeable problem. The items within the hierarchy are sorted by trait, and any updating that is to occur to a given instance when its source IP has scanned the system again, would require an extensive search to locate that so that it can be updated. As a result, there will be a second data structure in the system which will be purely sorted based on source IP address, and which will store a pointer to the actual instance containing the trait information in the larger CLASSIT tree.

### Distributed System

The two completed scan systems listed above were centralised points of analysis, and even the unimplemented Spice system only discussed the further work of going distributed, and was initially only planned to be centralised in nature. Centralised systems have two main concerns regarding their design. Firstly, the central controlling component acts as a single point of failure within the system; if it goes down then you are left unprotected. There is no redundancy possible, as can be present within a distributed system. Secondly, they tend not to scale well, given that they are operating on a finite network size. Therefore, the system we are proposing will be distributed, with an instance of the correlation engine running at each gateway upon the network.

The main challenge with opting for a distributed system is being able to transport all the data around the network in an efficient manner that doesn't actually accelerate the negative effects of a denial of service attack or similar attack. Various IDS have worked on this issue and developed methods of overcoming this risk; however, as some recent work in the area of security policy systems make use of peer-to-peer architecture (Janakiraman, Waldvogel & Zhang 2003 ). We are planning on investigating a peer-to-per system to transport the audit data around the network between hosts, as it is a highly active research area focusing on efficiently sharing data between hosts.

## FURTHER WORK

The scalability of the proposed system is one of the crucial features that make dictate whether or not the system will be ultimately useful in a live context. If that limit is reached to easily then a more intelligent data distribution system will be needed. Another serious concern with distributed security systems is the integrity of the nodes spread throughout the network. If one of the correlation engines on the network is compromised it would be possible for it to send false reports to other engines. To combat such a risk, a mobile agent system could be implemented to monitor the hosts, running integrity checks on the correlation engines.

## CONCLUSION

Data mining techniques have grown in importance within the defence context over the past few years. They are used to identify possible terrorists and to monitor terrorist linked websites. They are also used in network defence systems. In the bulk of these situations there are masses of data needing to be analysed to extract the few vital pieces of information that could later be used to thwart an attack. Network defence is a classic example of this and is of increased importance, with the possibility of terrorist-motivated network intrusion.

This paper has outlined the existing scan correlation systems, before explaining a proposed system that will make use of incremental conceptual clustering to correlate scan activity across multiple network gateways. The amalgamation of data from multiple sources through the use of a peer-to-peer network will preserve network context, facilitating accurate knowledge acquisition. The system aims to be able to detect malicious users who use complicated scanning techniques in an attempt to mask their identity.

The system proposed is fundamentally an anomaly based system, clustering behaviour into groups in an attempt to extract those users posing a threat. The techniques and models used are applicable in other defence contexts when attempting to locate instances of behaviour whose intent is malicious, while attempting to appear benign on initial analysis.

## REFERENCES:

Denning, D 1987, 'An Intrusion-Detection Model', IEEE Transactions on Software Engineering, vol. SE 13, no. 2, pp. 222-32.

Fisher, DH 1987 'Knowledge Acquisition Via Incremental Conceptual Clustering ', Mach. Learn. , vol. 2 no. 2 pp. 139-72

fyodor 2005, Nmap, <http://www.insecure.org/nmap/>.

Gennari, JH, Langley, P & Fisher, D 1989 'Models of incremental concept formation ', Artif. Intell. , vol. 40 no. 1-3 pp. 11-61

Janakiraman, R, Waldvogel, M & Zhang, Q 2003 'Indra: A peer-to-peer approach to network intrusion detection and prevention ', in Proceedings of the Twelfth International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises IEEE Computer Society, p. 226

Jung, J, Paxson, V, Berger, AW & Balakrishnan, H 2004, 'Fast Portscan Detection Using Sequential Hypothesis Testing', paper presented to IEEE Symposium on Security and Privacy, May 9-12.

Mitnick, KD & Simon, WL 2005, The Art of Intrusion: The Real Stories Behind the Exploits of Hackers, Intruders & Deceivers, John Wiley & Sons, New York.

Samudra, I 2004, Aku melawan teroris! (Me Against the Terrorist!), Jazeera.

Schneier, B 2000, Secrets and lies : digital security in a networked world, John Wiley, New York ; Chichester.

Shapira, B, Last, M, Elovici, Y, Zaafrany, O & Kandel, A 2003, 'Using Data Mining for Detecting Terror-Related Activities on the Web', paper presented to 2nd European Conference on Information Warfare and Security (ECIW 2003), University of Reading, June 30 - July 1.

Sipress, A 2004, 'An Indonesian's Prison Memoir Takes Holy War Into Cyberspace: In Sign of New Threat, Militant Offers Tips on Credit Card Fraud', Washington Post Foreign Service, Tuesday, December 14, p. A19.

Staniford, S, Hoagland, JA & McAlerney, JM 2002 'Practical automated detection of stealthy portscans ', J. Comput. Secur. , vol. 10 no. 1-2 pp. 105-36

Wald, A 1947, Sequential Analysis, John Wiley and Sons, New York.

Zalewski, M 2005, Silence on the Wire: A Field Guide to Passive Reconnaissance and Indirect Attacks, No Starch Press.

## COPYRIGHT