

VISUAL BEHAVIOR CHARACTERIZATION FOR INTRUSION DETECTION IN LARGE SCALE SYSTEMS

ROBERT F. ERBACHER

*Department of Computer Science, LI 67A
University at Albany - SUNY, Albany, NY 12222
erbacher@cs.albany.edu*

ABSTRACT

This work focuses on the visual representation of relations towards aiding the exploration and analysis of network intrusions. Fundamentally, the visual representations aid an analyst in comprehending the activity of individuals incorporated within the data set. Their actions are represented visually using a node and link metaphor. The visualization aids the analyst in identifying the complex interactions intrinsic to identifying the overall goal of an individual, i.e., the individuals true behavior. Such analyses are becoming critical with the continuing growth of the Internet and the corresponding growth of hackers and attempted intrusions. This is complicated by the fact that hackers, in general, will attempt to hide their activities from analysis; thus increasing the complexity of the analysis needed to identify their actions, particularly when a successful intrusion has occurred.

Key Words: Information Visualization, Intrusion Detection, Computer Security

1. INTRODUCTION

Our work focusses on the exploration of network and system data to identify network intrusions. Identifying an intrusion requires the identification of complex interrelated events. Rarely will a single event identify an intrusion. A single event may provide an initial warning of a user's activities but will not identify their goal. Limiting the analysis to individual events will lead to many false alarms; as many controversial acts are examined for possible intrusions. It also leads to many missed intrusions, since hackers will attempt to cloak their activities within normal activity to escape detection.

We are not concerned with trivial attacks by script kiddies as they are easily thwarted by other means, e.g., portsentry [1]. Our concern is focused on the advanced hacker who truly has an understanding of what they are doing, can gain entry to a system, and can avoid detection using new and adaptive means. It is this need to analyze adaptive behavior that is our goal in the analysis of

network and system data. Since analysis of individual actions is insufficient in the analysis process we must represent all events of all users, showing temporal as well as parameter based relationships. From these details the overall activities can be viewed and analyzed for abnormality. Thus, the visual representation, through the representation of parameters, aids the analyst in characterizing the behavior of individuals according to the degree of acceptability of the activity and the potential for future unacceptable behavior.

2. PREVIOUS WORK

The principal body of work related to network intrusion is from the information exploration shoot-out, organized by Georges G. Grinstein and supported by the National Institute of Standards and Technology (NIST) [2]. In this project, researchers were given access to a data set consisting of a network intrusion. The idea was to identify which researcher's techniques were effective at identifying the intrusion. The driving philosophy was that little work has been done to compare visualization techniques in a formal setting. Perceptual studies have been done to identify characteristics of the human visual system [3] that should be used as a basis for the development of visualization techniques, but little has been done to actually compare and contrast visualization techniques. There is no body of literature that identifies what visualization techniques work better on a given data set.

Most previous work involving visualization related to networks has emphasized graphics that depict network performance and bandwidth usage [4, 5, 6, 7], even down to the router [6], individual packets [8], and individual e-mail messages [9]. The techniques developed for these purposes do not provide sufficient detail or handle sufficient numbers of nodes and attributes in combination for our needs. The work by Eick et al. [9] strictly deals with e-mail and subsequently resolves many fewer nodes and attributes than is needed for intrusion detection.

Becker et al. [10] discuss the SeeNet environment that provides linkmaps for visually representing the

amount of data being sent between two network nodes. Livelink [11] is an environment for visualizing and measuring the web. By probing web accesses they gather statistics as to the number of hits web sites are receiving. Heydon et al. [12] discuss the application of visual languages to security specification. These tools are representative of most general network visualization tools. They do not adequately represent the interrelationships between systems and accesses both spatially and temporally to identify any cohesive attack on a system. They also do not provide sufficient visual attributes to identify the activity occurring in sufficient detail, as will be accomplished by our approach.

3. VISUALIZATION TECHNIQUES

The current network analysis problem is with too much information and inefficient analysis techniques. Analysts must filter what information they collect and analyze daily. The information they generally collect is much less than the information that is actually available. Only when a problem is detected will additional information be collected. This leads to situations where intrusions may not be detected for quite some time. User level applications are not reported in these logs due to the volume of information that would be generated. Also, network traffic is generally not collected due to its volume. All of this additional information, while useful,

would take too long to analyze with conventional techniques.

The problem with analyzing log files results from the fact that reading textual information is perceptually a serial process. Interpretation of graphical images, on the other hand, is perceptually a parallel process. Forcing the analyst to use textual information, therefore, slows the analysis process substantially in comparison to the use of graphics. An additional advantage of imagery is that more *concepts* can be presented in a single image. Thus, rather than observing individual reports or report summaries, it is possible to observe a single image that embodies the same information. This will reduce the amount of mental context switching required by users, making system assessment both easier and more efficient.

Visually, we must provide informative and perceptually based techniques that allow the analyst to examine the activities in the computing environment as a whole and quickly identify activities that require further investigation. Figure 1 shows log information for a one hour period of time in a lightly loaded environment. About 40 messages were generated. Heavily used environments will generate hundreds of messages each hour. Notice that with this textual information it is impossible to correlate the individual messages to derive a greater sense of the activities being performed by each user. It is critical that the analyst be able to see, quickly, the interactions of each user on the system, between

```

Jan 9 12:15:12 visualizer-s.cs.albany.edu xinetd[899]: START: pop3 pid=28097 from=169.226.2.54
Jan 9 12:15:12 visualizer-s.cs.albany.edu xinetd[28097]: USERID: pop3 WIN32 : Analyst
Jan 9 12:16:31 broomstick.cs.albany.edu in.telnetd[16593]: connect from root@cs.albany.edu
Jan 9 12:16:31 cs.albany.edu in.telnetd[16593]: connect from root@cs.albany.edu
Jan 9 12:22:29 visualizer-s.cs.albany.edu CROND[28100]: (root) CMD ( /sbin/rmmod -as)
Jan 9 12:25:31 broomstick.cs.albany.edu in.telnetd[16628]: connect from cdial20.infoblvd.net
Jan 9 12:25:31 cs.albany.edu in.telnetd[16628]: connect from cdial20.infoblvd.net
Jan 9 12:26:02 cs.albany.edu named[25266]: dangling CNAME pointer (google.lb.google.com)
Jan 9 12:29:45 cs.albany.edu in.telnetd[16654]: connect from Workstation72.ctg.albany.edu
Jan 9 12:29:51 von.cs.albany.edu in.rlogind[5625]: connect from pb@broomstick.cs.albany.edu
Jan 9 12:30:13 visualizer-s.cs.albany.edu xinetd[899]: START: pop3 pid=28101 from=169.226.2.54
Jan 9 12:30:13 visualizer-s.cs.albany.edu xinetd[28101]: USERID: pop3 WIN32 : Analyst
Jan 9 12:31:30 cs.albany.edu named[25266]: Cleaned cache of 799 RRs
Jan 9 12:31:30 cs.albany.edu named[25266]: NSTATS 979061490 977153081 Unknown=6 A=393521 NS=3
CNAME=98 SOA=9575 PTR=73966 MX=15120 TXT=10 AAAA=42 AXFR=32 ANY=12019
Jan 9 12:31:30 cs.albany.edu named[25266]: XSTATS 979061490 977153081 RR=198301 RNXD=66697
RFwdR=150932 RDupR=302 RFail=619 RFErr=0 RErr=17 RAXFR=32 RLame=16943 ROpts=0 SsysQ=23483 SAns=373313
SFwdQ=131146 SDupQ=30183 SErr=0 RQ=504450 RIQ=0 RFwdQ=131146 RDupQ=2489 RTCP=1069 SFwdR=150932
SFail=3460 SFErr=0 SNAAns=68541 SNXD=241409
Jan 9 12:32:28 visualizer-s.cs.albany.edu CROND[28103]: (root) CMD ( /sbin/rmmod -as)
Jan 9 12:34:07 karp.cs.albany.edu in.telnetd[27063]: connect from nas-70-57.albany.navipath.net
Jan 9 12:34:17 cs.albany.edu named[25266]: dangling CNAME pointer (gd25.doubleclick.net)
Jan 9 12:42:29 visualizer-s.cs.albany.edu CROND[28105]: (root) CMD ( /sbin/rmmod -as)
Jan 9 12:45:12 visualizer-s.cs.albany.edu xinetd[899]: START: pop3 pid=28106 from=169.226.2.54
Jan 9 12:45:12 visualizer-s.cs.albany.edu xinetd[28106]: USERID: pop3 WIN32 : Analyst
Jan 9 12:52:29 visualizer-s.cs.albany.edu CROND[28108]: (root) CMD ( /sbin/rmmod -as)
Jan 9 12:52:33 karp.cs.albany.edu in.telnetd[27137]: connect from 169.226.14.70
Jan 9 13:00:12 visualizer-s.cs.albany.edu xinetd[899]: START: pop3 pid=28109 from=169.226.2.54
Jan 9 13:00:12 visualizer-s.cs.albany.edu xinetd[28109]: USERID: pop3 WIN32 : Analyst
Jan 9 13:02:29 visualizer-s.cs.albany.edu CROND[28111]: (root) CMD ( /sbin/rmmod -as)
Jan 9 13:03:29 visualizer-s.cs.albany.edu CROND[28113]: (root) CMD (run-parts /etc/cron.hourly)
Jan 9 13:08:30 cs.albany.edu in.telnetd[16702]: connect from cm-24-29-78-15.nycap.rr.com
Jan 9 13:11:43 karp.cs.albany.edu in.telnetd[27175]: connect from grande.cs.albany.edu
Jan 9 13:12:29 visualizer-s.cs.albany.edu CROND[28115]: (root) CMD ( /sbin/rmmod -as)
Jan 9 13:14:55 cs.albany.edu named[25266]: dangling CNAME pointer (mdl.doubleclick.net)

```

Figure 1: Example of a log file over the course of an hour in a lightly loaded environment.

systems, and with temporal relationships. Interactions on different systems may be occurring simultaneously or at different times. It is necessary, therefore, to visually represent parameters, relationships, as well as temporal information contained within the data set. The log files contain all of this critical information. The goal is to derive that information and effectively present it to the analyst.

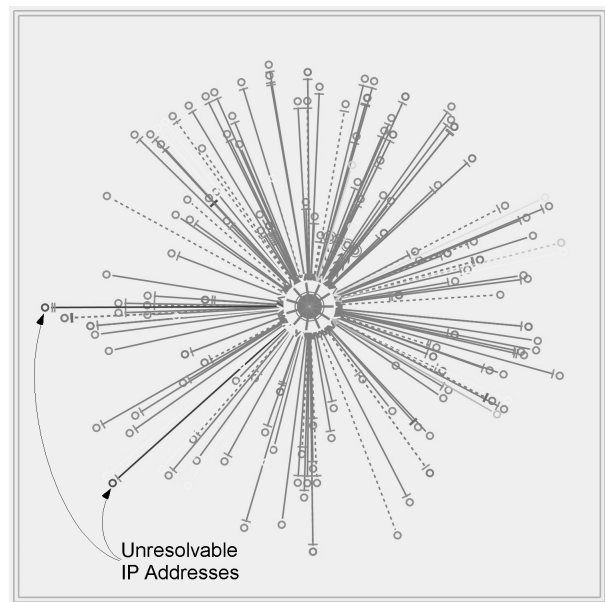
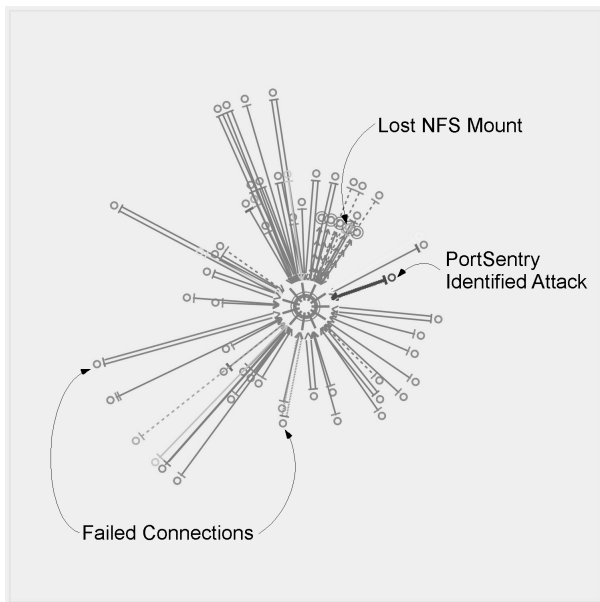
The total volume of data collected through log files varies significantly from one system to another, depending on usage, varying from 30,000 records a week for a lightly used workstation to 200,000 records a week for critical servers. We collected approximately 500,000 records over the course of a week from the university's principal server and a dozen other workstations. If you imagine the difficulty of making sense of this many log records you will get a sense of the task placed on the shoulders of system administrators.

The examples in Figure 2 use abstract visual representations. Systems accessed over NFS mounts are represented by lines with several arrows along their length. A lost NFS mount is highlighted yellow on screen. Initial connections, before the user has passed the authentication challenge, are represented with two parallel lines. Failed connections have their lines highlighted in red. Telnet connections are represented with solid lines, privileged ftp connections by long dashes, and anonymous ftp connections by short dashes. All lines are directed. The intensity of the node shows the duration since the node's last access. The node under examination shows the number of users through the protruding lines and system

load through the thickness of the inner circle. Portsentry identified attacks are shown in bright red with thick lines.

Figure 2a shows a snapshot from early in the morning so there are many nodes attempting initial connections. In order to assist with determination of the time of day the full display includes a border around the screen. The intensity (gray level) of this border is white at noon and black at midnight. An additional yellow border is displayed in association with PM. This assists in determining if the intensity is increasing or decreasing at any particular point in time. The example in Fig 2a. has a very light border without a yellow border so it is clearly approaching noon at the point this image is taken. The larger circles are principal systems being monitored. The smaller circles are remote connections, including both local and remote connections as well as UNIX and non-UNIX based systems. Fig. 2a also shows two connection attempts which failed to successfully pass the authentication challenge before timing out, an NFS mount which does not seem to be responding to queries, and a portsentry identified attack. Note that the portsentry identified attack appears to be a local system and thus should be investigated.

The nodes are positioned on the screen in five rings. The ring for a node is chosen based on the difference between its IP address and that of the monitored system. If only the right most number differs then the node is on the local subnet and is placed in the first ring, and so on. This is representative of the user's locality. Users that are on the same subnet as the monitored system can be easily identified. This assists in identifying the activity of local users as they will be clearly different from non-local users



(a) (b)
Figure 2: Basic visual representation of network and system activity. Information represented using color on the screen is annotated here.

and also shows multiple and indirect accesses that otherwise would not be visible. If the IP address for the node could not be resolved, i.e., the system is not known by the domain name server for some reason, it is placed in the fifth ring and is colored red. This occurs frequently when attackers or other users spoof their IP address to avoid detection or identification. The position of a node is recorded and node positions are not reused. This ensures that nodes or hosts always appear in the same position. This is critical for identifying temporal relationships or anomalies related to a single host's activity on the network.

Fig. 2b shows an example in the PM, actually very close to midnight. There are many more connections. This example also shows a single system with many individuals connected to the monitored system, fading nodes, and two systems whose IP addresses cannot be resolved. The unresolved hosts appear red on the display but are annotated here due to the lack of color. The fact that it is PM can be identified by the fact that there are two borders surrounding the image and the intensity of the outer border is representative of the time of day.

3.1. ATTACK ANALYSIS

Analyzing the collected information and determining if an attack or misuse is occurring requires that the intent or behavior of the individual be analyzed. Currently, when suspicious behavior is noted the individual's activities are examined, most often after the fact. With visualization it is possible to examine an individual's activity as it is occurring and determine immediately, before substantial harm has been done, that the individual's activities are unacceptable. Even suspicious activity can be difficult to detect with standard log file based approaches that require the analyst to peruse textual information.

Network traffic data can be incorporated into the display, allowing the user to quickly examine the data for particular types of traffic, such as illegal systems on the network, improper application usage, connections from unknown systems or users, etc. With this integration, additional cases of misuses and intrusions can be detected very quickly. For example, the case of the personnel at the CIA running an illegal chat room could have been detected through the analysis of network traffic information that would have identified the characteristic IRC packets on the network and would have been a clear indication of misuse. Since the information is not being read textually but rather interpreted visually through a graphical display, the gigabytes of information related to network traffic and user space applications can quickly be analyzed at intervals.

Single actions by an individual do not provide much context or basis for their motivation in their activity.

Certain activities clearly indicate illegal system usage, however, these actions are most often identified in users who are inexperienced in subverting a system. These types of novice users are easily identified with conventional techniques. Our concern must be with the more experienced users who will attempt to hide their tracks or camouflage their actions. In these situations, even though the user may be attempting to cover their tracks or hiding their true intent, the user's overall actions, when taken together, will clearly indicate the motive of the individual. In this fashion, we are providing analysts with tools that allow them to visually examine the activity on the computer systems under the analysts control as well as network usage in a merged environment.

3.2. BEHAVIOR IDENTIFICATION

In everyday life we must ascertain the intent and motivation of individuals. In a computing environment, the same level of information we use socially is not available. We must collect the information that is available and provide it in a form such that the activities of the user can be examined. The behavior of an individual can be derived from the activities the user performs, when these activities are performed, the order they are performed in, and how the presence of others affects their activities. At issue is the need to collect information that system administrators currently either do not collect, or collect but do not analyze due to the clutter it adds to log files.

For example if we consider the example in Fig. 2b we can see that there are numerous users accessing the system. Most of these accesses are static, carryovers from the daytime resulting from individuals who did not log off. There is one node of interest, a user's connection that is colored red since a reverse hostname lookup failed for that system. Taking individual actions alone aren't enough to comprehend the meaning of this activity. If, however, we consider that the user is performing a telnet in the middle of the night from a hostname that we cannot lookup then the situation begins to appear objectionable. Thus, it is all the characteristics taken together that tell a story about the user and that user's activity. This can then be used to derive the meaning of said activity and determine if it should be considered objectionable or not and what level of action needs to be taken. It is this ability to take multiple characteristics, through multi-parametric visualization techniques, and integrate them to find a greater understanding, that is the key to analyzing network and computer usage for intrusions.

A second example of interest is shown in Fig. 2b. The six nodes at the top are connections that were made in rapid succession from different IP addresses. Notice that these nodes are not within the university's local network itself. Is this an indication of an attack? Had they been

local to the university's network they would have been deemed to be students logging on immediately after class. Once removed from the university directly, seeing such sequences in rapid succession should raise a level of concern. At most organizations such sequences will be seen when classes or meetings end or typical starting times for employees arrive. Other types of activity in conjunction could have made the scenario more or less objectionable, particularly if they had been port scans. These types of activities are warning signs as to possible intrusions. The administrator's knowledge of local behavior is critical to make sense of the data and understanding its meaning.

It is important to note that these behavioral issues are observable as the system executes and the changes in state are animated. The behavior of the individuals observed in these animations can be interpreted and characteristics of the individuals determined. Static images and text will not exhibit these qualities as clearly.

4. OTHER INDICATORS

Watching the animation of the display over time aids in the perception of temporal as well as spatial relationships indicative of the overall activity of individuals or systems. The identification of an attack or misuse is dependent on the identification of related activities which are unexpected or unusual. This can be as simple as a user logging remotely into a machine that is not assigned such activity or complex interactions over time whereby a port scan is observed from a remote machine from which a failed connection is later identified; note the importance of the temporal relationship. The goal in the latter example is to identify the attack *before* a successful connection occurs. Other indicators may include:

- Unsuccessful connections to many different machines from a single host, particularly if different protocols are also attempted.
- A remote individual who successfully connects to a local machine and then immediately attempts to connect to other systems, either local or remote, with most of the attempts failing. This shows an example of an indirect relationship.
- An excessive number of connections between two hosts. Can be all local, an indication of internal misuse, or a combination of local and remote hosts.
- And many more

The identification of indicators and their interpretation is critical to the identification of an attack. A future goal is to develop a full set of classifications associating a set of known activities with the known activity indicative of the activity. In this way, system administrators will be provided a strong foundation on

which to base their monitoring activities. However, with the ever adaptive nature of hackers and the continuously changing and developing forms of attack this will only act as a foundation and the system administrator will need to be perceptive of *unexpected* activity that may be indicative of some new form of attack or misuse.

5. CONCLUSIONS

Computer and network security are becoming critical issues. The capabilities are not yet in place to allow analysts to efficiently detect and counteract intrusions of the systems and networks under the analysts control. Only through the innovation of new technologies can we hope to be able to counteract the growing threat from hackers. Of the techniques available, visualization appears well suited to take on the brunt of this task. Perusal of textual log files is totally inadequate.

By providing sufficient attribute mappings within the visualization we can represent substantial characteristics as to the overall behavior of users within the environment. By analyzing user behavior as a whole we can gain insight into the user's intent and ultimate goals. It is only through the combination of attributes when taken together that the whole of the meaning of the user's activity can be discerned. By focusing analysis on the user's behavior we are reducing the number of false alarms and increasing the reliability of the system analyst's analysis. Ultimately, the incorporation of visualization tools should prove to greatly improve the detection of intrusions before damage is incurred to the system.

The visualization tools will also aid in reducing false alarms and identifying potential problems that would otherwise go undetected. These types of situations can be a great drain on an analyst's time. Ultimately, this will become much more than just an early warning system for analysts, rather it will become a filtering device allowing the analyst to filter out unwanted details and identify real activity of concern.

6. REFERENCES

- [1] Joel Scambray, Stuart McClure, George Kurtz, *Hacking Exposed*, 2nd Edition, Osborne/McGraw-Hill, 2001.
- [2] Georges Grinstein, iWorkshop on Information Exploration Shootout Project and Benchmark Data Sets: Evaluating How Visualization does in Analyzing Real-World Data Analysis Problems, *Proceedings of the IEEE Visualization '97 Conference*, IEEE Computer Society Press, Phoenix, AZ, pp. 511-513, 1997.

- [3] Markus Gross, *Visual Computing, The Integration of Computer Graphics, Visual Perception and Imaging*, Springer-Verlag, 1994.
- [4] Richard Becker, Stephen Eick, and Allan Wilks. *Graphical methods to analyze network data*, In *IEEE International Conference on Communications ICC 93 Proceedings*, Geneva, Switzerland, pp. 946-95, May 1993.
- [5] Taosong He and Stephen G. Eick, *Constructing Interactive Visual Network Interfaces*, *Bell Labs Technical Journal*, Vol. 3, No. 2, pp. 47-57, April-June 1998.
- [6] Kenneth Cox, Stephen Eick, and Taosong He, *3D geographic network displays*, *ACM Sigmod Record*, Vol. 25, No. 4, pp. 50, December 1996.
- [7] Eleftherios E. Koutsofios, Stephen C. North, Russel Truscott, and Daniel A. Keim, *Visualizing Large-Scale Telecommunication Networks and Services*, *Proceedings of the IEEE Visualization 97 Conference*, IEEE Computer Society Press, San Francisco, CA, pp. 457-461, 1999.
- [8] Deborah Estrin, Mark Handley, John Heidemann, Steven McCanne, Ya Xu, and Haobo Yu, *Network Visualization with Nam, the VINT Network Animator*, *IEEE Computer*, Vol. 33, No. 11, pp. 63-68, November 2000.
- [9] Stephen G. Eick and Graham J. Wills, *Navigating Large Networks with Hierarchies*, In *Visualization 93 Conference Proceedings*, San Jose, California, pp. 204-210, October 1993.
- [10] Richard Becker, Stephen Eick, and Allan Wilks, *Visualizing Network Data*, *Readings in Information Visualization: Using Vision To Think*, Stuard Card, Jock D. Mackinlay, and Ben Shneiderman, editors, Morgan Kaufman Publishers, pp. 215-227, 1999.
- [11] Tim Bray, *Measuring the Web*, *Readings in Information Visualization: Using Vision To Think*, Stuard Card, Jock D. Mackinlay, and Ben Shneiderman, editors, Morgan Kaufman Publishers, pp. 469-492, 1999.
- [12] Allan Heydon, Mark W. Maimone, J. D. Tygar, Jeannette M. Wing, and Amy Moormann Zaremski, *Miro: Visual Specification of Security*, *IEEE Transactions on Software Engineering*, Vol. 16, No. 10, pp. 1185-1197, October 1990.