Visual Network Forensic Techniques and Processes

Robert F. Erbacher, *Member IEEE,* Kim Christiansen, Amanda Sundberg
Department of Computer Science, Utah State University, Logan, UT 84322

Abstract—Network forensics is the critical next step in the analysis of network attacks, intrusions, and misuses. It is the forensic process that will aid identification of what occurred and how. With the explosion in numbers and types of attacks it is critical that new techniques be developed to aid in the analysis of said attacks. For instance, with the recent widespread deployment of botnets, it is forensic techniques that will allow these botnets to be dissected to determine their extent, their capability, their control mechanisms, etc. In this paper we discuss visualization techniques designed around the analysis of network traffic data and tailored to the scalability issues intrinsic to such data. In conjunction with these techniques we discuss how these techniques fit into an analysts repertoire, how we foresee them being used, their advantages to the forensic process, and the process through which they will be effectively applied.

Keywords—Network Forensics, Legal Issues, Visualization, Intrusion Detection, Interactive Feedback Loop

1. INTRODUCTION

Cyber security has become a major point of concern given the criticality of the network infrastructure in today's society. Additionally, many techniques have been explored to identify potential attacks on a network, including: pattern matches [30], data mining [12], and visualization [6]. Meanwhile, after an attack or an intrusion is identified, the analyst is left with little recourse accept to manually and tediously examine available system log files, network traffic data, and recorded system statistics, in order to determine what did in fact occur in sufficient detail as to resolve the incident. In essence, the goal of network forensics is the examination of data collectable from networks of computer systems in order to examine some form of criminal activity. Using the aforementioned data sources, analyst must be able to identify and examine any form of compromise of a network, whether it is an intrusion from an outside source, an inside job, unauthorized modification of websites, etc. The goal at this stage is for the analyst to answer typical forensic questions.

- What happened?
- How did it happen?
- Why did it happen?
- When did it occur?

More importantly, these questions must be related to the networked infrastructure:

- Who broke into the system/network? More specifically, from where did the attack initiate? The validity of any identification is limited due to IP Spoofing and the use of intermediate compromised hosts. However, it can aid identification of compromised systems being used to launch attacks and aid correlation with other attacks and further identification of compromised systems.
- What did they compromise/damage? We must identify what the focus of the attack was such that we can determine what actions need to be taken in order to recover the systems within the local network from the attack, identify any data that may be compromised, identify local legal liability, and identify any appropriate legal recourse that must be initiated.
- What did they gain access to? Was it sensitive?
- How did the intruder break in? This is critical since if we can not identify how an attack was deployed we will not be able to defend against it in the future.
- What systems need to be repaired?
- How can future such attacks be protected against?
- Identification of when the attack occurred. The duration of time during which the system was compromised can have an enormous impact on any of the above considerations.

In essence, the analyst must determine what the attacker or intruder did in as much detail as possible such that they can be prevented in the future and the appropriate civil and criminal procedures followed (if needed), based on the associated damage and risk assessments [21]. Should a system be missed by the analyst then a compromised system, and thus vulnerabilities, will continue to exist within the organization. Such compromised systems can be used for future attacks, to sniff the network for data, or be used as future dissemination vehicles for viruses, worms, or denial of service attacks. Consider, for example, the threat of a compromised system at a bank, e-commerce site, or credit card company. The information available is highly sensitive. Should the exact mechanism the attacker used not be identified then the organization will be subjected to a future, perhaps more organized attack. Should the details of the compromised data not be identified then the extent of damage to the companies' Intellectual Property will not be completely determined and the extent of the needed criminal or civil action will not match the damage to the company's bottom line. For example, it is critical that it be identified should credit card information be stolen such that customers can be warned and the credit card numbers deactivated.

With the above explanation of the criticality of the analysis process following an intrusion, analysts are still left with principally reading log files for the relevant data. Simple searches [19], pattern matching [30], and statistical analysis [13] can help but by no means reduce the tediousness of the effort. With the volume of such attacks on the rise, capabilities for improving the task are sorely needed.

2. NETWORK FORENSIC PROCESS

Forensics is a well established process through which typical criminal activity is investigated. The process for such typical forensics is well established and accepted by the courts. With network forensics we must provide new analysis capabilities to deal with the volume and type of data needing analysis. Additionally, a process is needed to ensure the acceptability and validity of the underlying data as well as the analysis process and resulting data themselves. Our proposed process for network forensics is exhibited in figure 1.

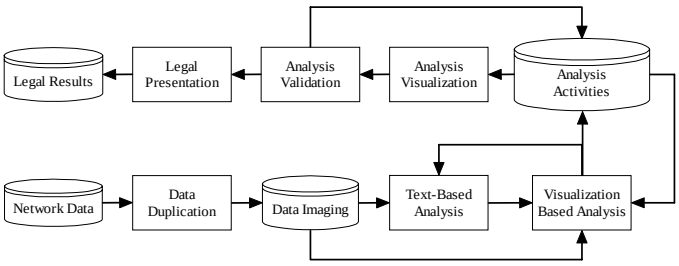

Figure 1: Abstract network forensic process diagram. Of key importance is the interactive feedback loop affecting the visualizations. Also of importance is the incorporation of legal needs into the process design. In this way we ensire that all aspects of the analysis process are recorded for later examination.

2.1. Data Validation

Clearly, we will be starting from some form of raw data associated with the identified intrusion or compromise. This original data source must be protected in order to ensure validity and verification. The idea here is to be able to prove that the data has not been modified since it was collected. This can be done by encrypting the data with a dated certificate or making a complete copy of the hard drive and storing the entire hard drive, following complete chain of custody and evidence storage procedures. The key is to not operate on the original data but protect the original data in some way such that it can be shown that any evidence identified exists in this original data source and this original data source is uncompromised.

2.2. Text-Based Analysis

The fundamental capabilities for analyzing network traffic data hinges on the analysis of textual representation of the underlying network traffic data and binary-based pattern matching to a limited extent.

For instance, when analyzing network traffic data, the analyst will have identified at least one system that has been compromised. This compromised system provides a focal point for initial analysis and the first step is generally to perform an initial filter to leave only events related to the target system. Later in the analysis process, after more details of the attack and attacking system are well known then the analyst can return to the full set of raw data in order to identify other systems that may have been targeted by this attack.

Given the volume of data intrinsic to network forensics, the filtering of data to narrow analysis focus is crucial. Additional analysis capabilities, however, are also necessary, including:

- **Pattern matching** – In order to identify the attack sequence which led to the compromise it is critical to be able search for patterns of activity. The patterns may be fixed or regular expressions. Typical activities searched for would include sequences of bytes used in known buffer overflow attacks, or sequences of packets that don't match typical connection sequences, etc.
- **Stream identification** – With the number of packets collected, the ability to cluster packets can greatly reduce the amount of effort involved. The most fundamental clustering parameter is based on event streams. Associating packets with event streams allows the analyst to rapidly eliminate large numbers of packets if the event stream can be identified as being acceptable, i.e., not part of the compromise.
- **Data browsing/examination** – At a final level, analysts will need to examine the raw packet data in a binary, textual, or hybrid format. Ultimately, there is no way to do searches for all types of attacks. This requires that the analyst examine this raw data in order to perform the final steps in identifying what packets were involved in the attack and how the compromise was instigated.
- **Domain knowledge** – Given the level of analysis at all levels required in order to forensically analyze network data, the analyst must have extensive experience and understanding of network data in order to determine if packets are innocuous or malicious. Additionally, knowledge of the local network will aid in improving of the analysis process as often identification of whether a packet or packet sequence is innocuous is dependent on local policy.

These capabilities will need to be integrated into the visualization capabilities. The work by Lakkaraju et al. [16] provides a starting point for some of this integration but much more work is needed to provide analysts with all of the capabilities they have become familiar with.

2.3. Visualization Based Analysis

Clearly, much of the analysis process is inhibited by the volume of data and the need to filter or cluster data. The extensive amount of data examination required makes network data analysis a very slow process. Consequently, visualization techniques work with the analyst to improve the process. These visualization techniques must be designed to work with the typical forensic analysis capabilities. Through visualization and graphical user interfaces the typical commands are still available but are made more accessible.

While the visualization techniques will rely extensively on the raw data, they will also rely on the results of the more traditional techniques. This creates an interactive feedback loop in which the textual analysis capabilities provide more meaning and context as the results are incorporated into the visualization. The visualization then aids better comprehension of these results and provides better access to the traditional commands such that the next sequence of commands can be initiated in a more effective fashion.

Thus, the goal of the visualization is to to make the analysis process more effective and efficient. We discuss our specific

capabilities for visual forensic analysis in section 4.

2.4. Analysis Results

Typically, the key aspect of an analysis is simply the results of the analysis. With our proposed model the analysis process itself is of critical importance. The analysis results provides a second interactive feedback loop to aid future analysis as well as aiding in proving the legal admissibility of the results.

In terms of the analysis process, the goal is to keep track of what the analyst has examined, why (through annotation), when this was done, and how. The sum of the identified activities of the analyst provides a picture of the analysis process.

This database of activities provides benefits for future analysis, both by this analyst and by other analysts. For instance, we envision the development of machine learning techniques to guide analysts by identifying typical activities performed during previous analysis sessions that bore out successfully; especially if a previous session can be deemed to be particularly similar to the current session. Additionally, these machine learning techniques could automatically create macros for typically performed activities, shortening aspects of the analysis process.

In terms of legal validity, the database will provide background as to what types of activiuties are typically done by analysts. This will show that the current analysis follows convention. Given the wide variety of analysis scenarios this paradigm isn't perfect as it will require a very large database in order to validate an analysis sequence. However, it is an effective and necessary first step.

2.5. Analysis Visualization and Validation

As the analysis process data is collected it will be necessary to analyze this data itself. More importantly, it will be necessary to present this data in court proceedings. This essentially will require the development of further visualization techniques. The goal of these techniques will to a small part be for analysis, for instance:

- What analysis techniques or processes appear to be most effective?
- What techniques are missing or not used effectively?
- How effective are individual analysts and their associated techniques?

In essence, these techniques are similar to the visualization-based analysis techniques previously introduced but applied to the collected analysis data and with the change in focus of validating and presenting the data, with less focus on actual analysis capabilities.

The more important portion of these visualization techniques will be to present the results in court proceedings and otherwise validate the analysis process under examination. Thus, the visualization techniques must be designed to present the data, rather than providing the ability to analyze the data. In terms of presentation, the visualizations must show the difference between the analysis under discussion and typical analysis processes. Given that no two data sets will be exactly the same, there will obviously be differences but we must be able to show that these differences are not significant and that the overall process is identical. This will amount to a validation process with respect to the forensic analysis process.

The results of this validation will further feed back into the analysis results database such that analysts using processes or steps deemed invalid can be warned of such and focused on more appropriate techniques. Additionally, the machine learning techniques would be able to provide indications as to the likelihood of evidence admissibility given the processes followed by the analyst and the relationship of the identified evidence to the analysis task.

3. Issues of Scale

We have already indicated that the critical issue in the analysis process is the volume of data needing analysis. As we are primarily concerned here with network forensics, network traffic data will be our primary concern. However, within any network analysis scenario, there may be any amount of other data sources available, such as:

- System log files, especially for critical servers
- Firewall logs
- Web logs
- Router logs
- System statistics

In terms of specifics, typical network log files easily achieve multiple gigabytes in size with many organizations acquiring multiple terabytes of data. Current analysis capabilities are completely lacking in their ability to analyze such large volumes of data.

Clearly, the network traffic data will create the bulk of the data both in terms of sheer volume as well as in terms of number of individual events. However, the other sources of data will create enormous numbers of additional events needing analysis.

Further, we must be concerned not only with the sheer volume of data but also with the number of data dimensions needing analysis. In the simplest sense, network traffic data has dozens of dimensions. Of course, with a change in perspective, network data can have tens of thousands of dimensions. For instance, often ports are a critical parameter that provides indication as to the validity of activity. If we consider ports as a primary dimension, then there will be 65K dimensions as well as numerous parameters, including: number of packets, originating host, destination host, packet time, packet validity, etc. Thus, the developed capabilities must handle the enormous number of dimensions with whatever view of the data the analyst may require. Other issues of scale include:

- Number of data elements in both the network flow data and system log data
- Number of ports
- Number of local machines to be examined
- Number of remote machines exhibiting activity on the local network
- Complexity of relationships. Events may be temporally and spatially distributed resulting in enormous numbers of possible combinations of events leading to relationships.
- Number of packets involved in a single event stream

The developed capabilities must aid without hindering the

analyst. There may be enormous volumes of data with enormous numbers of parameters. Given the rapidly changing nature of attacks it is not possible to ascertain how the analyst will need to look at the data and the developed capabilities must handle all worst case scenarios.

4. Visualization for Forensic Analysis

In terms of the visualization techniques themselves our goal was to develop the techniques with the analyst's needs in mind. In terms of the visual analysis, the visualization techniques actually play second fiddle to the interaction techniques as it is the interaction techniques that will make or break the forensic analysis process. The essence of the visualization design process is shown in figure 2.

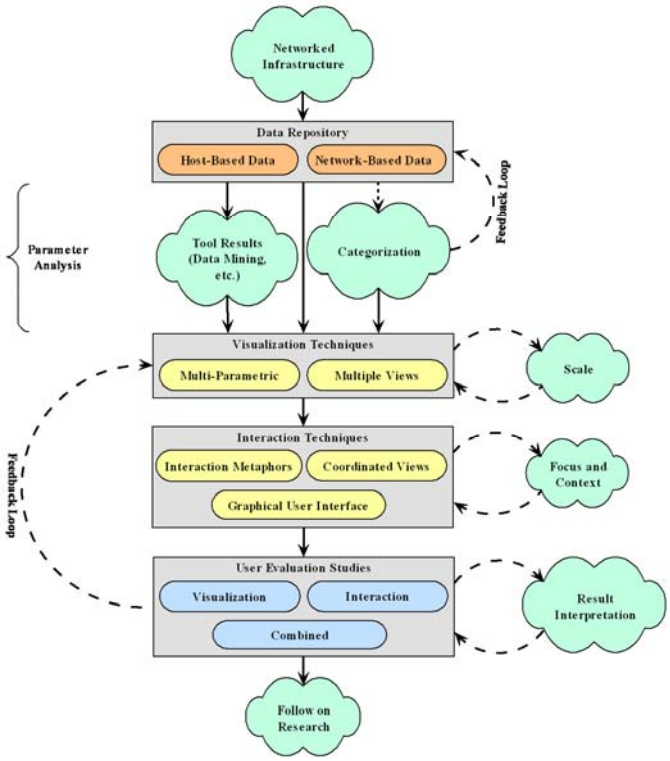

Figure 2: Issues in the development of visualization techniques and associated processes.

As exemplified in the diagram the entire process is iterative and all of the components are interdependent. The techniques we design are focused specifically on the network forensic analysis task. Additionally, both the interaction and the visualization techniques are designed based on analysis of the specific data needs.

We also do not forego the value of other analysis techniques within the analysis process. For instance, data mining algorithms can potentially be quite effective at aiding identification of anomalous activity or packets that may aid identification of the original attack or compromised event stream. However, current data mining techniques have enormous false positive and false negative rates with respect to network traffic data, due to its complexity, the complexity of attack streams, and the low signal to noise ratio with respect to attack data. Thus, while the data mining techniques can not be used effectively in isolation they can aid in providing focus for the analyst. Combining multiple techniques in cooperation with visualization should prove quite effective.

4.1. Advantages of Visualization and Interaction Techniques

Given the volume of data needing analysis, there is a need for capabilities to allow for the rapid examination of this data. Typical, textual processing or simple searches still leaves an enormous amount of manual and time consuming processing by the analyst. Visualization techniques designed carefully in conjunction with interaction techniques can greatly aid analysts and allow for rapid examination of even very large-scale datasets. Developing such visualization and interaction techniques is the goal of our research.

4.2. Related Techniques and Research

While other visualization techniques have been applied to the forensic problem they at the very least did not integrate interaction into the design of the visualization, making the interaction more difficult and clumsy. In general, however, these visualization techniques do not even integrate substantial interaction techniques. Given the similarity of the problems between intrusion detection and network forensics we discuss examples of both techniques.

The work by Teoh et al [23], [24] focuses on Internet routing data and thus is limited in its applicability in intrusion detection and will have no applicability to forensics. The work by Foresti et al [9] is designed for situational awareness and thus is designed for decision makers. This technique does not provide the low level detail, can't handle the volume of data needed, and does not provide the interaction techniques needed for network forensics. The work by Krasser et al. [15] is limited with respect to what attributes of the data are presented and provides only limited interaction techniques, reducing the clarity and exploratory ability to the environment. The work by Eick et al. [5] strictly deals with e-mail and subsequently resolves many fewer nodes and attributes than is needed for intrusion detection

These environments typically deal with small numbers of processors that are working on a single task and thus have a common grounding and have not been applied to intrusion detection. Many environments are geared towards naïve monitoring of port activity, Teoh et al [18], [24], but such work can not handle the scale of real data and infrastructures, differentiate sophisticated (e.g., low and slow) attacks, or handle the diversity of data and attack types that are truly exhibited in today's environments.

Finally, the VizSec community has begun publishing a body of work related to intrusion detection through a series of workshops [2], [17]. Again, the focus of this body of work is on intrusion detection and the body of work does not deal with the unique issues intrinsic to network forensics. For instance, few of the techniques provide the level of detail or the level of interaction necessary and certainly not combined. This body of work can be used as an indicator of what visualization

strategies will and will not work. Thus, it can be used to avoid reinventing the wheel and while our tasks and challenges differ there are characteristics of the visualizations that will remain similar.

4.3. Visualization Techniques

Given that analysts must be directly involved in the process of data analysis, the visualization techniques are the key to effective forensic analysis. However, one key concept with visualization is that no single visualization can solve all problems or are appropriate for all tasks. Thus, for given forensic problems, different visualization techniques will likely be more appropriate than others. While we currently focus our efforts on a single visualization technique, the architecture for our tool provides support for the incorporation of many visualization techniques as well as support for coordinated views.

4.3.1. Coordinated Views

The idea with coordinated views is to allow multiple visual *views* of the data; the term view originating from database concepts. These views may be the same or different visualization technique, the same or different subsets of the data, or the same or different parameter mappings, etc. By coordinated we are referring to the fact that when the user interacts with one of the visualizations the interactions are transferred to the other visualizations. Coordinated views will greatly improve the effectiveness of the exploration and analysis process by allowing relationships and data elements identified in one display to be associated with the corresponding visual elements in the other displays. The other displays can then carry on the exploration process through additional visual and interaction metaphors.

4.3.2. Multiparametric Visualziation Techniques

The visualization techniques we focus on are geared around multiparametric techniques. This is fairly obvious given the number of parameters we are dealing with and may vary greatly depending on the focus of the visualization. For example, when representing network connectivity we will often need to represent: port number, source IP, destination IP, fragmentation, packet length, etc. When viewing port accesses as a principal dimension we will have the following parameters: port number, source IP, destination IP, volume of packets, mean time between packets, packet length, etc. Which parameters should be used, particularly as principal attributes? What visual forms will cognitively, effectively, and in a form comprehensible to the target analysts, represent the available data?

The visual representation will ultimately take the form of glyphs presented on the screen. Each glyph will represent one or more matching events, parameters, or systems. Attributes of the glyph will be representative of an associated parameter. The attribute can be representative of the presence of a parameter, i.e., an on/off mapping. In the more complex case the attribute will be representative of the value of a parameter. In this case, we will map the actual value of the parameter to the grayscale intensity of the attribute, i.e., black represents the lowest value and white represents the highest value; other color scales can also be explored.

Additional parameters can easily be mapped to the visualization, i.e., alerts identified by other tools. These parameters may be mapped to the connector intensity, connector style, connector thickness, etc.

4.3.3. Example Visualization Techniques

The basic visualization technique is exemplified by figure 3. The idea behind this technique is to represent network activity as efficiently and concisely as possible. This is to handle as much activity as possible and begin to resolve some of the many scalability issues inherent to network data and the associated network forensic task.

The developed visualization technique in its default form begins by representing the local IP (Internet Protocol) address of a connection around the radius of the internal circles. The technique then represents remote IP addresses along the top and bottom edges of the window. This redundancy aids in reducing clutter and line crossings. The top edge is used if the local IP appears in the top semicircle and the bottom edge is used if the local IP appears in the bottom semicircle. Similarly, port numbers are represented on the left and right edges of the window. If the local IP appears in the right semi-circle then the right edge of the screen is used and if the local IP appears in the left semi-circle then the left edge of the screen is used. Essentially, we are using the inherent rectilinear nature of the display to provide segregation between the port numbers and remote IP addresses. The outer rings represent the most current data while each inner ring represents data m^n time units older, where m is the ring number and n is a user configurable parameter; in essence the rate at which n increases is user controlled.

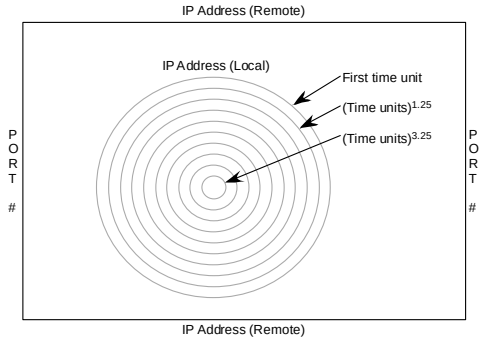

Figure 3: Basic layout of the visualization techniques. The goal of this basic visualization technique is to handle basic scalability issue as well as provide representation of many of the key parameters associated with network traffic data.

Each of these mappings is user configurable. Thus, the user interface allows the user to quickly and dynamically change the data parameters mapped to each of the visual attributes. The behavior of the visualization itself remains intrinsically the same.

By default, ten rings are presented. The number of rings is also user configurable. Thus, the multiple rings provide for the persistence of the display, allowing extensive periods of time to be represented simultaneously. The use of multiple rings

allows for the differentiation of activity within and between temporal periods. In other words we can differentiate: the most recent activity, long term activity, and old activity. This allows for more extensive interpretation and analysis of the identified activity than if we had only a single ring.

An example of this visualization technique in action is shown in figure 4. This image shows an example applying actual network traffic data. As can be seen from this display the rings associated with older data have lower intensity to reduce their impact while maintaining the information for the analyst. Line crossings other than crossing ring boundaries are significantly limited, not only in number but also in visual impact. A future goal will be to reduce these even further.

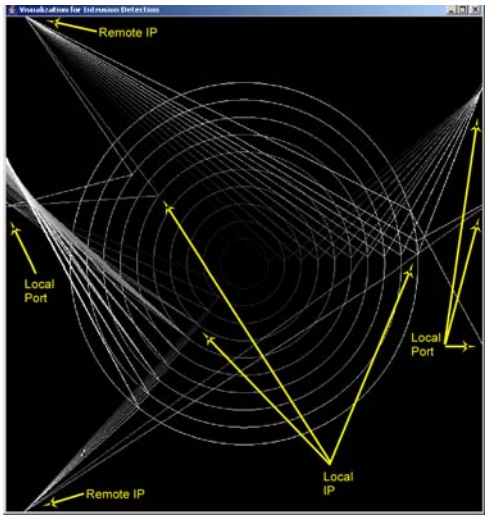

Figure 4: Annotated visualization technique showing actual network traffic data.

Analysis of this display shows that there is consistent activity from several remote hosts. This quickly shows the relationship or lack of relationship between activities within the display.

4.4. Interaction Techniques

Our primary goal with the interaction techniques was to develop the capabilities typically applied through command line based metaphors but through easier to use metaphors. Thus, not only are the capabilities provided but they are provided in a far more accessible paradigm. Additionally, the goal is to create analysis capabilities within the visualization. The visualization itself is designed to show the basic characteristics of the data as well as relationships. However, at some level, forensic analysis will always require extensive capabilities to examine the meaning behind the data and thus behind the visualization. These capabilities go from the fundamental, such as the ability to select and probe visual entities to garner feedback as to the underlying raw data, figures 5 and 6, to extensive filtering capabilities to remove elements from the display already analyzed and deemed not relevant for further analysis, figures 7 and 8. It is the filtering capabilities that are of particular interest as they greatly aid the ability to deal with the scale of the data to which the forensic analysis is being applied.

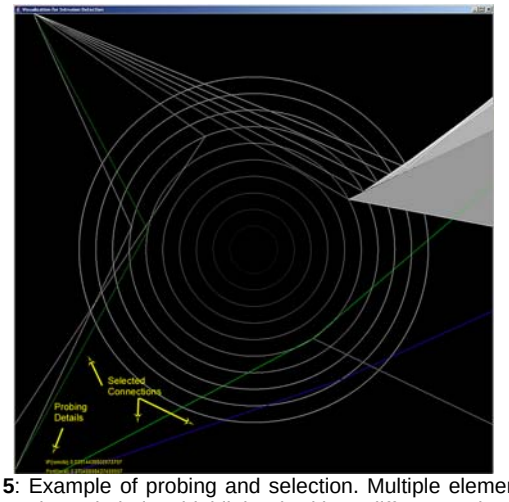

Figure 5: Example of probing and selection. Multiple elements can be selected, each being highlighted with a different color. Select details of the most recently selected element are displayed within the visualization (probing details). Selected elements remain highlighted for tracking and analysis. Through coordination, selected elements will also be selected within any other visualization.

4.4.1. Selection and Probing

Figure 5 shows the most fundamental technique to garner feedback from the display. While the visualization is critical for analyzing the data at a basic level, ultimately, the analyst must have effective access to the underlying details in order to fully analyze and assess the data. In essence, the visualization allows multiple elements to be selected simultaneous for focal point maintanence and tracking. Colors are selected pseudo randomly, allowing a large number of elements to be selected, each with a different color. Obviously, the more selections made simultaneously then the more difficult it will begin to become to differentiate between the selected elements. Additionally, the last selected element provides fundamental information as to the selected packet within the visualization display itself. This includes details such as local and remote IP and port numbers.

The environment also supports coordination with respect to such selections. This means that when an element is selected in one visual display, this selection is passed onto all other visual displays. This is done whether the visual display incorporates the same visualization or a different visualization. This allows for more rapid and complete analysis. Since no single visualization can effectively show all characteristics of the underlying data, it is critical that the analyst be able to examine the data using additional visualization techniques. The coordination allows the analyst to rapidly refocus on a different visualziation technique and identify the same exact elements.

Figure 6, shows the extensive detail view within the primary control panel. The details of the entire packet are displayed here, including association with the visual display by showing the color with which the packet has been highlighted. Using the tabbing the analyst can rapidly switch between multiple selections. This allows for the analyst to select and track multiple elements while simultaneously maintaining the ability to examine any number of packet details related to these selected elements.

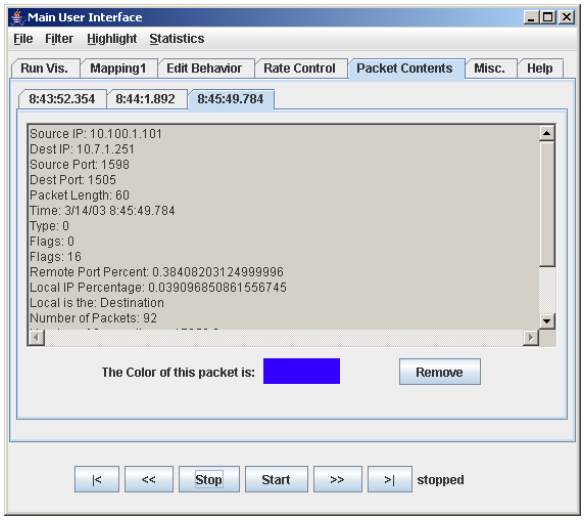

Figure 6: The details of selected visual elements, i.e., the raw data, are displayed for each selected visual element within a tab within the main control panel.

4.4.2. *Selection and Filtering*

In addition to the fundamental ability to rapidly acquire feedback and details from the visual display, we have developed extensive filtering capabilities. The filtering capabilities are particularly critical given the volume of data we are dealing with.

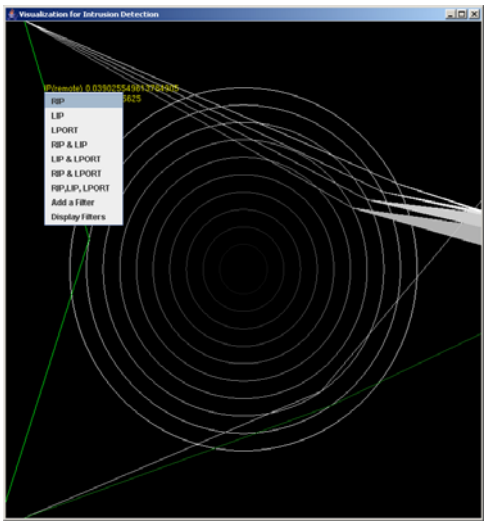

Figure 7: Example of probing and selection. Multiple elements can be selected, each being highlighted with a different color. Select details of the most recently selected element are displayed within the visualization (probing details). Selected elements remain highlighted for tracking and analysis. Through coordination, selected elements will also be selected within any other visualization.

We have developed two fundamental techniques for providing for the filtering of data elements. The first, shown in figure 7, is associated with the packet selection capability. By right clicking a dialog is displayed that allows rapid filtering of all packets based on this packet's: Remote IP, Local IP, Local Port, or a combination of these elements. Thus, if the analyst deems that a particular local port could not have been the source of the attack then all access to that port can be filtered from the display, greatly reducing the volume of data presented to the analyst and easing the remainder of the analysis process.

The second technique for filtering we have incorporated is essentially a separate control panel for specifying what information should be removed from the display, figure 8. In this display, the user can enter sequences of ports and IPs to be filtered. Ports or IPs entered on the same line are interpreted as being OR'd together. When both ports and IPs are entered simultaneously the elements between seperate lines are AND'd together. After a filter is entered and "saved", a new dialog is displayed to allow the user to enter additional sets of filters. This is necessary given the AND'd nature of the rows. This allows for the analyst to create quite extensive sets of filters.

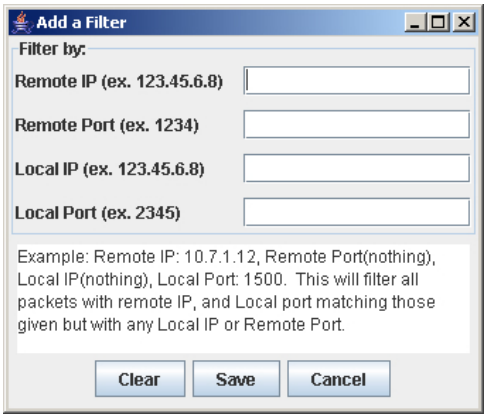

Figure 8: Dialog allowing the user to enter detailed filter sets.

A final display, figure 9, allows for the analyst to examine currently set filters. From this display the analyst may examine or remove any filters currently set.

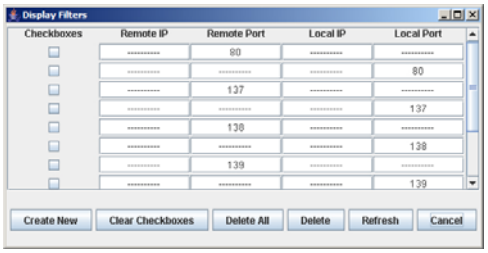

Figure 9: Display showing the summary of currently enabled filters with the ability for the analyst to remove filters.

4.5. *Applying the Techniques*

In terms of the forensic analysis, the probing in conjunction with the visualization techniques provides the main functionality needed. The visualization itself will aid identification of anomalies. The probing allows analysis of these identified anomalies and determiniation as to whether they are innocuous or malicious. Innocuous events or resolved events can then be filtered from the display to allow additional events to be identified and examined in detail. The unique design of the visualization allows for event identification, event analysis,

and selection of both individual events as well as groups of events.

When a malicious event *is* identified, the unique design of the visualization shows the path of the visualization, including the originating host and the target host. The target host can then be restored or otherwise corrected. Additionally, by selecting all other activity originating from the attacking host, the analyst can quickly identify any other systems that may have been attacked. These selected events can further be analyzed to determine the extent of the threat to these additional hosts. The group select just requires dragging a selection rectangle over all activity originating from the already identified attacker. Future work will automate portions of this task.

We envision the following process being followed in the application of the detailed visualization and interaction capabiliites to the forensic analysis:

1. Identification of the details and knowns of the compromise existing before analysis begins. These knowns will be used initially to guide the analysis process.
2. Examination of a subset of the entire database. This will be a temporal subset, i.e., the first *N* period of time.
3. Given the knowns from step 1, the analyst must identify regions within the display and determine groups or clusters of activity deemed to be related.
4. Elements in an identified grouping are slected to identify the low level details. Again, by selecting a visual element, the details are displayed in the control window.
5. The analyst must identify if the selected elements are malicious, innocuous, or not yet know.
   a. Innocuous elements are removed from the display through filtering. This reduces the clutter and aids in resolving issues of scalability.
   b. Unknown events are left as is with the goal of incorporating additional events into the display until all events are resolvable.
   c. Malicious events are examined to determine if they are part of the compromise, whether they are protected against, and whether they require a response.
   d. Should the events be deemed to be part of the compromise then they will be left in the display until all elements of the compromise are identified.
6. The analyst will repeat from step 3 until only compromise and unknown events remain.
7. The analyst will return to step 2 to add additional events to the display.
8. When all events related to the compromise are identified, the analyst will proceed with identifying the appropriate steps for recovery and further protection of the network.

5. Relation to Previous Work

With the rapidly growing concern for computer security an enormous interest is developing with respect to visual intrusion detection. This is due to the volume of data available and the lack of availability of comprehensive and effective tools. Fundamentally, much of the current work is limited in terms of the visual effectiveness, the applicability to visual cognition, lack of scalability, and lack of effectiveness towards sophisticated attackers.

Aside from our ongoing work in this area, little prior work has been applied to the use of visual analysis as an aid to network forensics. Much interest has been garnered for visual intrusion detection but network forensics has yet to garner the same level of interest. There is a difference due to the analytical nature of forensics in contrast to the pure identification requirements of intrusion detection. While in section 4.2 we covered some of the visualization techniques with direct applicability to network forensics, here we provide a summary of the more foundational work related to network forensics.

In terms of previous work we are principally concerned with intrusion detection due to the lack of work on forensics as a whole and their close relationship. In fact, many intrusion detection systems provide the foundation for network forensics as they deal with the same data (online vs. post-mortem). For example, eTrust Network Forensics (formerly SilentRunner) [25] shows the confusion on the issue as their environment primarily incorporate tools for security monitoring and intrusion detection.

A summary of forensic tools can be found in [3]. Clearly, a majority of the available tools incorporate only simple visualization techniques (i.e., graphs). Others, such as eTrust Network Forensics [25] incorporate slightly more sophisticated visualizations (essentially 3D graphs) but nothing on a scale that will aid in resolving the true network forensics issues.

5.1. Intrusion Detection Systems

As shown previously, the distinction between network forensics and intrusion detection is narrow. Snort [30] and Bro [20] integrate rule systems, for example, which could be used to identify patterns in network traffic data for either purpose, particularly with snort's ability to load any libpcap generated file. Consequently, we must explore the extent of capabilities of such environments.

While many intrusion detection tools have begun to incorporate basic graphical user interfaces (BlackICE [26], RealSecure [27], Cisco Secure IDS [28], eSecure [29]) they fall short of providing effective visualization displays to aid in interpreting the generated information. For example, most of the tools will provide an indication when it received an unexpected packet. But was this an attack, a misdirected packet, a casual attack, or a real attempt to break into the system? These systems do not adequately provide the detail and event interrelationships needed to analyze the activity in the detail needed forensically.

5.2. Visualization systems

In contrast to intrusion detection, quite a bit of visualization research has been applied to network accesses. 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) [10]. In this project, researchers were given access to a data set consisting of network intrusions. The goal was to identify which researcher's techniques were effective at identifying the intrusions.

The previous work involving visualization related to net-

works emphasized network performance and bandwidth usage [1], [4], [11], [14], [22], even down to the router [4], individual packets [8], and individual e-mail messages [5]. The techniques developed for these purposes do not provide sufficient detail or handle sufficient numbers of nodes and attributes in combination for our needs.. Other work has been geared towards visualizing systems for program analysis and program development [7].

6. CONCLUSIONS

We have developed techniques and processes for the forensic analysis of network traffic data. The process takes into account the complexity of real network forensics, the need for new technology both for fundamental analysis as well as for training, and the need to plan for legal validation of the resultant analysis.

In terms of the visualization itself we have developed and followed a process to ensure needs are met throughout the research process. This visualization process shows the critical components for the development of successful techniques.

Through the application of this process we have developed a fundamental infrastructure for meeting the identified needs of the visualization and interaction techniques. In particular, through realization that the interaction techniques are as important if not more important than the visualization techniques for forensic analysis we have ensured the architecture is scalable for both. Additionally, we have developed a simple visualization technique in conjunction with extensive analysis capabilities for an initial capability at forensically analyzing network traffic data. Of particular concern in the development of these techniques is the need to deal with issues of scale. Consequently, we have incorporated techniques for dealing with these scale issues, particularly the filtering capabilities applied in conjunction with the data selection and probing capabilities.

REFERENCES

[1] 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.

[2] *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security*, Editors: Carla Brodley, Philip Chan, Richard Lippman, Bill Yurcik, ACM Press, 2004.

[3] Eoghan Casey, “Forensic Network Analysis Tools,” *Digital Forensics Research Workshop*, 2003, http://www.dfrws.org/dfrws2003/presentations/Brief-Casey.

[4] Kenneth Cox, Stephen Eick, and Taosong He, “3D geographic network displays,” *ACM Sigmod Record*, Vol. 25, No. 4, pp. 50, December 1996.

[5] Stephen G. Eick and Graham J. Wills, “Navigating Large Networks with Heirarchies,” In *Visualization ‘93 Conference Proceedings*, San Jose, California, pp. 204-210, October 1993.

[6] Robert F. Erbacher, Kenneth L. Walker, and Deborah A. Frincke, “Intrusion and Misuse Detection in Large-Scale Systems,” *Computer Graphics and Applications*, Vol. 22, No. 1, January/February 2002, pp. 38-48.

[7] Robert F. Erbacher, “Visual Assistance for Concurrent Processing,” *University of Massachusetts at Lowell Doctoral Dissertation* (1998 CS-3), Lowell, MA 01854.

[8] Deborah Estrin, Mark Handley, John Heidermann, 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] Stefano Foresti, James Agutter, Yarden Livnat, Robert Erbacher, Shaun Moon, “VisAlert: Visual Correlation of Network Alerts,” *Computer Graphics and Applications*, Vol. 26, No. 2, 2006, pp. 48-59.

[10] Georges Grinstein, “Workshop 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.

[11] 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.

[12] Klaus Julisch, “Data Mining for Intrusion Detection: A Critical Review”, in D. Barbará and S. Jajodia, editors, *Applications of Data Mining in Computer Security*, Kluwer Academic Publisher, Boston, 2002.

[13] Akira Kanaoka and Eiji Okamoto, “Multivariate Statistical Analysis of Network Traffic for Intrusion Detection,” *Proceedings of the 14th International Workshop on Database and Expert Systems Applications* (DEXA'03), September 2003, pp. 472-476.

[14] 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.

[15] S. Krasser, G. Conti, J. Grizzard, J. Gribschaw, H. Owen, “Real-time and forensic network data analysis using animated and coordinated visualization,” *Proceedings of the 6th annual Information Assurance Workshop*, IEEE Computer Society Press, West Point, NY, pp. 42-49, 2005.

[16] Kiran Lakkaraju, Ratna Bearavolu, A. Slagell, W. Yurcik, “Closing-the-loop: discovery and search in security visualizations,” *Proceedings of the 6th annual Information Assurance Workshop*, IEEE Computer Society Press, West Point, NY, pp. 42-49, 2005.

[17] *Proceedings of the 2005 IEEE workshop on Visualization for computer security*, Editors: Kwan-Liu Ma, Stephen North, Bill Yurcik, IEEE Press, 2005.

[18] Jonathan McPherson, Kwan-Liu Ma, Paul Krystosek, Tony Bartoletti, Marvin Christensen, “PortVis: A Tool for Port-Based Detection of Security Events,” *Proceedings of CCS Workshop on Visualization and Data Mining for Computer Security*, ACM Conference on Computer and Communications Security, October 29, 2004.

[19] Gonzalo Navarro, Mathieu Raffinot, *Flexible Pattern Matching in Strings*, Cambridge University Press; 1st Edition, 2002.

[20] Vern Paxson, “Bro: A System for Detecting Network Intruders in Real-Time,” *Computer Networks*, 31(23-24), pp. 2435-2463, Dec. 14 1999.

[21] Thomas R. Peltier, *Information Security Risk Analysis*, Auerbach Pub; 1st Edition, 2001.

[22] Robert Spence, *Information Visualization*, Addison-Wesley, 2001.

[23] S.T. Teoh, K.L. Ma, and S. F. Wu, “Visual exploration process for the analysis of internet routing data,” In *Proceedingsof the IEEE Conference on Visualization* 2003, 2003, pp. 523-530.

[24] S.T. Teoh, K.L. Ma, S. F. Wu, and X. Zhao, “Case study: Interactive visualization for internet security,” In *Proceedingsof the IEEE Conference on Visualization* 2002, 2002, pp. 505-508.

[25] http://www3.ca.com/Files/DataSheets/etrust_networkforensics_pd.pdf

[26] http://www.networkice.com/

[27] http://www.iss.net/securing_e-business/security_products/intrusion_detection/index.ph

[28] http://www.cisco.com/univercd/cc/td/doc/pcat/nerg.htm

[29] http://www.esecurityinc.com/

[30] http://www.snort.org