A Novel Data Reduction Technique

Robert F. Erbacher
U.S. Army Research Laboratory
2800 Powder Mill Road
Adelphi, MD 20783
301-394-1674
Robert.F.Erbacher.civ@mail.mil

Anupama Biswas
Dept. of Computer Science
UMC 4205
Utah State University
Logan, UT 84322
Anupama.Biswas@aggiemail.usu.edu

Trent Cameron
Dept. of Computer Science
UMC 4205
Utah State University
Logan, UT 84322
Trenton.Cameron@gmail.com

ABSTRACT
Large-scale networks generate enormous numbers of events that network analysts must parse through in order to determine which are malicious attacks and which are not. Additionally, network analysts must prioritize the events such that the most severe attacks are resolved first in order to limit the potential for damage to the network as much as possible. While there exist many data reduction and event correlation techniques for reducing the amount of data needing analysis, these techniques do not provide prioritization capabilities.

This paper discusses our novel impact assessment technique geared towards the prioritization of events. This will aid network analysts and managers in identifying and resolving the most critical events first. Our techniques will work with the already existing data reduction techniques. The impact assessment technique identifies the potential impact of an. This impact assessment as an automated prioritization scheme will greatly improve the efficiency of the analysis process and reduce the amount of data needing to be transmitted over the network.

Categories and Subject Descriptors
K.6.5 [**Management of Computing and Information Systems**]: Security and Protection – *invasive software.*

General Terms
Algorithms, Management, Experimentation, Security.

Keywords
Data reduction, intrusion detection, impact assessment.

1. INTRODUCTION
For this research, we developed an automated impact assessment algorithm designed to work with existing capabilities to assist analysts in prioritizing network events based on the potential severity of the identified event sequences. The impact assessment scores identify the potential (expected) impact of an attack on the network and associated resources. It identifies the extent to which resources will be degraded by the attack. Thus, our capabilities allow for the analyst to more efficiently and effectively target the events with the greatest potential impact on the network. In turn, this will greatly reduce any potential impact on the legitimate system users. Additionally, we propose the application of vulnerability assessment techniques to be used in conjunction with the impact assessment algorithm to determine the potential impact of an event or event sequence more accurately. A highly vulnerable system is more likely to be susceptible to a given attack and thus the priority for an attack against such a system must be higher. The network analyst would thus give priority to high impact attacks being targeted at highly vulnerable systems, Figure 1. A future situational awareness visualization [1][3][4][6][8] of the impact and vulnerability assessment data essentially makes the network state far more approachable and comprehensible to the network analyst.

2. Impact assessment

The primary component of this research was the development of techniques for the identification and representation of the impact of an event. Our concept with respect to the presentation of impact is to perform an analysis of a wide range of events and associate these events with associated military impacts. For instance, the most applicable association is with readiness, i.e., cyber readiness. In other words, how ready is the network to deploy missions or countermeasures? Specifically, a denial of service attack would decrease available bandwidth associated with a specific computer system. This would be associated with an impact on the readiness of the effected computer system and a reduced ability for said system to deploy operations. In highly load-balanced environments, the associated impact would be similarly reduced. Corresponding impacts must be identified, associated with the full range of analyzed attacks, and visually represented. In essence, when computing an impact assessment score, we are attempting to compute the amount of degradation of available resources. More importantly, however, is identifying the levels at which the degradation in available resources negatively affects our ability to use those resources.

It is this concept of impact that will bring together, link, and correlate groups of events in a meaningful way for the network analyst and network manager. Rather than simply identifying the existence of an event, we will be attempting to identify what this event and other events together mean to the decision maker.

In effect, the goal of these mappings is to identify the operational readiness of the cyber infrastructure as a whole, as well as individual components. The mappings are designed in a generic format such that they can easily be adjusted based on considerations of each local environment. Of particular importance will be the need to be able to change the associated value with each impact as this can vary greatly.

Ultimately, the impact assessment will provide a prioritization score to the network analyst. In essence, the network analyst must deal with enormous numbers of events on a daily basis, especially for larger organizations. These events must be prioritized in order to identify a sequence in which they will be resolved. Currently, this prioritization is adhoc and based solely on the network analyst's expectations of the event significance. The goal here is

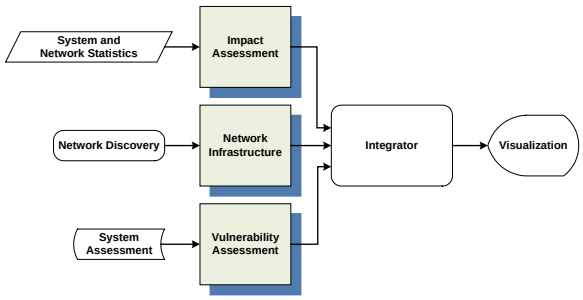

Figure 1: Diagram of impact and vulnerability assessment and integration for visualization.

to create an impact assessment that identifies for the network analyst the current and predicted impact of each event according to the network analyst's specified metrics.

2.1 Impact assessment computation

Our impact assessment computation is derived from the work by Hariri et al. [7].

In essence, Hariri et al. provide two mechanisms by which impact can be computed. These are dependent on how the network analyst wishes to specify acceptable impact. First, the network analyst can specify the minimum acceptable resources that must always be available. Alternatively, the network analyst can specify the maximum acceptable usage of available resources. The equations specified by Hariri would then be:

- CIF(Router, FSk)=|Bfault-Bnorm|/|Bmax-Bnorm|
- CIF(Client, FSk)=|TRnorm-TRfault|/|TRnorm-TRmin|

Where CIF is the computed impact factor. Here, the assessment is per resource and the individual values for each resource would have to be summed to generate total system impact. For our purposes, we also include priority. Different systems will have different levels of importance to the network analyst, i.e., having a server heavily impacted is of greater concern to the network analyst than a common desktop. Thus, our inclusion of priority is designed to account for the variation in systems importance. Therefore:

$CIF' = CIF * Priority_{system}$
CIF=0..1
0 is normal
1 is abnormal
Priority=0…1

For this research, we initially computed impact factors manually as follows:

- Identify a set of attacks covering a wide array of attack types.
- Simulate these known attacks to quantitatively and accurately identify their impact on resources. This formed our initial attack corpus [2].
- Map the attacks identified in the sample attack data to these known attacks. In essence, we are attempting to map the characteristics of new simulated attacks against attacks in our known attack corpus to determine if the sequence of events should be considered an attack and to derive its future potential impact. This can result in an attack in the sample attack data being partially mapped to multiple known attacks, resulting in a list of percentages.
- The resultant impact factors for the sample attack data are calculated.
- These values are stored in our database for use by the network analyst, likely through situational awareness techniques.

We have performed simulations of attacks to identify the impact those attacks have on network bandwidth, CPU utilization, memory utilization, and disk utilization. The second step in the process was to manually map the identified impacts onto the attacks identified in sample attack data. These impacts could be further mapped to available services, systems, and missions. The full set of results is beyond the scope of this paper but appear fully in [2]. While we currently map the simulated (new) attack to the attacks in the attack corpus manually, this could be done in an automated manner in the future through attack graph analysis or related techniques.

Caveat: The fact that we store the potential impact rather than simply calculating it allows for the analyst to directly modify some of these values. This is critical for attacks in which the rate of resource consumption, pour main measure of impact specification, is inappropriate. A common example for which this is the case is botnets. For botnets we will want to relate a much greater impact than the simple rate of consumption of resources.

2.2 Formula derivation for mapping sample attack data to simulated attacks

This section provides more specifics as to how the impacts factors of the sample attack data were generated. For the purpose of this analysis, we considered the following performance parameters:

- CPU Usage
- Memory Usage
- Disk Usage
- Network Bandwidth Usage

Threshold values of the attacked system and network: These are the values of the victim network and system before the attack:

Network Bandwidth Capacity: 94.13 Mbps
CPU: 3 GHz
Memory: 1 GB
Disk: 2.5 GB

For our initial example of the impact assessment computation, we will consider two attacks: ICMP Flood Attack and ICMP ping NMAP. All attacks were implemented to perform the simulation, amounting to ~1100 lines of code. The ICMP Flood Attack was previously computed and its impact metrics stored in our attack corpus. The goal was to deploy an actual simulation of the ICMP ping NMAP Attack, correlate it with the ICMP Flood Attack, and determine how the ICMP ping NMAP Attack's resource usage compares to that of the ICMP Flood Attack's resource usage. These two attacks are compared due to their similarity. In a deployed capability, identifying this relationship would allow the network analyst to predict the future potential impact of a new attack, an attack not currently in the attack corpus.

For this evaluation, we are considering only a single system under attack. The simulations used three systems, running through a single switch, namely: the target systems, the attacker system, and the detection system running the Wireshark protocol analyzer. This was a controlled environment to measure only the impact of the attack so there were no other users on the system. The values of all parameters were measured before the attack was initiated to acquire the ground truth data. The parameters were again

measured after the attack was initiated. The difference in values identified the impact of the attack.

All results presented are associated with the single target system associated with the simulation. If multiple systems were being affected then the CPU usage, memory usage, and disk usage would include the impact of the attack on all affected systems; thus scaling them appropriately. Differences in system configuration may prevent a simple linear scaling of the impact.

The severity of the attack's impact is measured by the rate at which the resources are being used. Hence, the measurements should be in such a format that represents the amount of consumption of the resources as well being comprehensible to the network analyst. Hence, the network bandwidth usage is measured in Mbps. Memory usage and disk usage are measured in GB/min, which indicates the amount of storage-based resources being used if the attack continues. For CPU resources, a typical usage based on CPU load is acquired, which has a value of *zero* for no load and a value of *one* for the full-time usage of one equivalent CPU. A difficulty arises, however, since CPU load is not directly related to other CPUs. Thus, to generalize the same here, there is a need to map the CPU load to an indicator of that CPU's capability. Here, it is mapped to the speed of the CPU, GHz, essentially indicating how much of the available processing power would be consumed by the attack. Another option would be to map to the MFLOPS rating of the processor, but this is less accessible and can be even more misleading since the attacks are not going to be optimized for efficiency of instruction usage.

ICMP Flood Attack: An ICMP Flood Attack is a denial of service attack in which one or more attacking systems send a sufficient number of ICMP echo requests to overload the target system's resources. This prevents the target system from being able to respond to any future requests, even valid ones. The values were recorded during simulation of the attack for a period of 20 minutes.

Available Network Bandwidth during Attack: 39.0 Mbps

NW Bandwidth Usage: (94.13 - 39.0) Mbps*60/1024/8= .4038 GB/min *300 = 121.14 GB/min

CPU Usage: 1.26 GHz/20 = 0.063GHz/min

Memory Usage: 0.53GB /20 = 0.0265 GB/min

Disk Usage: 2.5 GB/20 = 0.125 GB/min

Total Resource Usage: 121.14 GB/min + 0.063GHz/min + 0.0265 GB/min + 0.125 GB /min = *121.3545*

Thus, Total Resource Usage is a representation of the total resources used by the attack. We convert all values to similar scales of units, namely giga* per minute. A side effect of this is that the network bandwidth becomes an extremely low value in comparison with the other values; this is seen below where we compute maximum availability values. To compensate we multiple the network bandwidth by a scale factor, namely 300. This could also be construed as a priority value. For instance, should network bandwidth be considered the most important resource then it could be multiplied by a greater amount.

Network Bandwidth Capacity: 94.13 Mbps*60/1024/8= .6894 GB/min *300 = 206.82

CPU: 3 GHz*60 = 180 GHz/min

Memory: 1 GB*60 = 60 GB/min

Disk: 2.5 GB*60 = 150 GB/min

Total Resource Availability: 206.82+180+60+150 =596.82

Percentage of Total Resources Used by Attack = 121.3545/596.82 *100%= 20.3335%

One thing that becomes clear very quickly is that most of the values are extremely small, except for network bandwidth usage. The percentage of available network bandwidth being used is very high but the other values are extremely small. This is representative of ICMP Flood Attack being an older denial of service attack using many network packets that current network services handle efficiently. An alternative computational strategy would be to compute the individual usage percentages for each resource and then average them together.

This is an example of an attack in our attack corpus. The interpretation of the specific impact of this attack would be up to the individual network analyst. What is more important is when a new attack is identified and we can associate that attack with an attack already in our attack corpus. Then this would allow the network analyst to predict the impact of the new attack. For instance, we can simulate an ICMP Ping NMAP attack, which is similar to and would be mapped to the ICMP Flood Attack.

ICMP Ping NMAP: This is a network scan attack. In essence, it is an indication that nmap was used to generate a sequence of pings of the target network. This is often seen as a precursor to more direct and malevolent attacks. Since there is no direct way to measure the usage of the mentioned parameters, we have assumed certain values given below. It is also assumed that these values were calculated over a period of 20 minutes.

Available NW Bandwidth during Attack: 50.5 Mbps

NW Bandwidth Usage: (94.13-50.05) *60/1024/8 = .3229 GB/min *300 = 96.8555

CPU Usage: 0.6 GHz/20 = 0.03 GHz/min

Memory Usage: 0.5 GB /20 = 0.025 GB/min

Disk Usage: 2.5 GB/20 = 0.125 GB /min

Total Resource Usage: 96.8555 + 0.03 + 0.025 + 0.125 = 97.0355

Ratio of resources used compared to ICMP Flood Attack = 97.0355/121.3545*100% = 79.9603%

Thus, the impact of an ICMP Ping NMAP attack is 79.96% of the impact of ICMP Flood Attack. In other words, the ICMP Ping NMAP uses approximately 79% of the resources used by ICMP Flood Attack. This will essentially allow a network analyst to predict the future impact of an attack.

2.3 Generalization of the formula

The threshold values of the attacked system and network are the values before the attack and are represented as follows:

Let the availability of the resources of the attacked system be defined as:

AV_{CPU}: Amount of CPU available for usage in GHz

AV_{MEM}: Amount of memory available for usage in GB AV_{DSK}: Amount of disk available for usage in GB

AV_{NW}: Amount of network available for usage in Mbps.

Let the values of the above parameters for the simulated attack for time period T1 minutes be defined as:

S_{CPU}: Percentage CPU usage by simulated attack

S_{MEM}: Percentage memory usage by simulated attack

S_{DSK}: Percentage disk usage by simulated attack

S_{NW}: Network bandwidth available by simulated attack (in Mbps)

Let the Priority Factor be defined as :

PF: Priority factor which is assigned a value 1.

Let the Priority value for each of the parameters be defined as follows:

$CPU_{PRIORITY\_VALUE}$ = 1 min/GHz

$MEMORY_{PRIORITY\_VALUE}$ = 1 min/GB

$DISK_{PRIORITY\_VALUE}$ = 1 min/GB

$NWBANDWIDTH_{PRIORITY\_VALUE}$ = 1 min/GB

Let the usages of the above parameters for the simulated attack in T1 minutes be defined as:

$US_{CPU} = (AV_{CPU} * S_{CPU} * PF * CPU_{PRIORITY\_VALUE}) / T1$

$US_{MEM}=(AV_{MEM} * S_{MEM} * PF * MEMORY_{PRIORITY\_VALUE}) / T1$

$US_{DSK} = (AV_{DSK} * S_{DSK} * PF * DISK_{PRIORITY\_VALUE}) / T1$

$US_{NW}=((AV_{NW}-S_{NW})*PF*NWBANDWIDTH_{PRIORITY\_VALUE})/T1$

Total resources used by the simulated attack (AS_{TOTAL}) in T1 minutes is as follows:

$$\mathbf{US_{TOTAL} = US_{CPU} + US_{MEM} + US_{DSK} + US_{NW}} \quad (1)$$

It is important to note that the specific terms for the different components will be factored out in equation 3. The goal with adding the different terms is to relate the total impact of the attack while treating these terms as equally important. This differs from Hariri et al. [7] in which each component is individually checked to determine if it is anomalous and the percent of anomalous components is determined. We felt this was insufficient since it allows attacks to go unrecognized should they limit the number of impacted components, even if those components are significantly impacted. By adding the individual components, we ensure that all impacts of an attack are represented and presented to the network analyst. The individual priority terms for each component can be used should one of the component values be too small relative to the other components.

Let the values of the identified parameters for a new attack for a time period T2 minutes be defined as:

N_{CPU}: Percentage CPU usage by new attack

N_{MEM}: Percentage Memory usage by new attack

N_{DSK}: Percentage Disk usage by new attack

N_{NW}: Network bandwidth usage by new attack (in Mbps)

Let the usages of the above parameters for the new attack in T2 minutes be defined as:

$UN_{CPU}=(AV_{CPU} *N_{CPU} * PF * CPU_{PRIORITY\_VALUE}) / T2$

$UN_{MEM}=(AV_{MEM}*N_{MEM}* PF * MEMORY_{PRIORITY\_VALUE}) / T2$

$UN_{DSK}=(AV_{DSK}* N_{DSK} * PF * DISK_{PRIORITY\_VALUE}) / T2$

$UN_{NW}=((AV_{NW}-N_{NW})*PF*NWBANDWIDTH_{PRIORITY\_VALUE})/T2$

Total resource usage by the new attack in T2 minutes is as follows:

$$\mathbf{UN_{TOTAL} = UN_{CPU} + UN_{MEM} + UN_{DSK} + UN_{NW}} \quad (2)$$

Thus, the impact severity percentage calculation is as follows:

$$\mathbf{(UN_{TOTAL} / US_{TOTAL})* 100} \quad (3)$$

Equation 3 is particularly critical as it essentially eliminates the remaining terms leaving a value without GHz, Mb, GB, etc. Additionally, this converts the total impact into a percentage. This helps to neutralize the deviation is scale of terms from equations 1 and 2.

3. Related Work

Ultimately, the goal with our research is to provide a more effective and usable data reduction technique for network managers. This is a necessity due to the scale of the current problem and the insufficiency of current techniques such as data reduction, automated classification, and intrusion detection systems. The data reduction domain focuses on three primary techniques, namely: event correlation, clustering, filtering, and dimensional reduction [5].

Our technique provides data reduction while simultaneously providing prioritization information for the network analyst. Existing capabilities may provide data reduction but do not provide the prioritization. The goal in this section is to provide examples of other relevant techniques for data reduction. We also identify how they relate to our technique. Many of these techniques can be used in conjunction with our proposed technique.

4. Conclusions

This research presents significant advances in the automatic generation of impact assessment data. This will have numerous direct benefits to network managers and network analysts. First, the focus on impact assessment essentially amounts to a distributed data reduction technique. This approach puts much lower strain on any database system attempting to store the network data. The proposed data reduction technique distributes the computation and reduces the volume of data, reducing the network and database impact and increasing the amount of data feasibly analyzable by network analysts. Finally, the ability to prioritize attacks more effectively and target the attacks of greater potential threat allows network analysts to provide a more secure environment while continuing to support the connections needed for large-scale connections.

5. References

1. Adam, E. C. (1993), "Fighter cockpits of the future," *Proceedings of 12th IEEE/AIAA Digital Avionics Systems Conference (DASC)*, pp. 318-323.
2. Biswas, A., "Impact and Analysis of System and Network Attacks," Master's Thesis, Department of Computer Science, Utah State University, Defended 11/2008.
3. Endsley, M. R. (1995b), "Toward a theory of situation awareness in dynamic systems," *Human Factors*, 37(1), pp. 32-64.
4. Erbacher, R. F., Walker, K. L., and Frincke, D. A., "Intrusion and Misuse Detection in Large-Scale Systems," *Computer Graphics and Applications*, Vol. 22, No. 1, January/February 2002, pp. 38-48.
5. Fodor I. K., "A survey of dimension reduction techniques," *LLNL technical report*, June 2002, UCRL-ID-148494.
6. Foresti, S., Agutter, J., Livnat, Y., Erbacher, R., and Moon, S., "Visual Correlation of Network Alerts," *Computer Graphics and Applications*, Vol. 26, No. 2, March/April 2006, pp. 48-59.
7. Hariri, S., Qu, G., Dharmagaddam T., Ramkishore, M., Raghavendra, C., "Impact Analysis of Faults and Attacks in Large-Scale Networks," *IEEE Security and Privacy*, September/October 2003, pp. 49-54.
8. Lewis, L., Jakobson, G., Buford, J., "Enabling Cyber Situation Awareness, Impact Assessment, and Situation Projection," *Proceedings of IEEE SIMA/MILCOM*, 2008.