2011 Fifth IEEE International Conference on Semantic Computing

Multiresolution Semantic Visualization of Network Traffic

Alefiya Hussain
USC/Information Sciences Institute
Email: hussain@isi.edu

Arun Viswanathan
USC/Information Sciences Institute
Email: aviswana@isi.edu

***Abstract*—Multiresolution semantic analysis of data involves inferring increasing levels of meaning from data. Due to the large volume and complexity of network data, multiresolution visualizations allow the user to rapidly focus on meaningful and relevant information. Current tools and techniques for visual analysis of network data are limited in their ability to operate at semantically relevant resolutions and impose great cognitive burden on users to manually infer semantics from low-level details. Extending our previous work that allows users to define semantics as abstract models, we apply these models to construct multiresolution visualizations of network traffic data. Our methodology for visual exploration allows the user to rapidly analyze and understand network traces, by providing *intuitive* and *interactive* representations of the network. We demonstrate the effectiveness of our approach by applying it to analyzing network trace data from a cyber security incident involving DNS cache poisoning.**

I. Introduction

Network and cyber security researchers routinely face many difficult challenges when attempting to analyze large volumes of dynamic and complex message exchanges generated during an Internet incident. In this demanding environment, we need sophisticated analysis mechanisms to convert raw network traffic traces into *meaningful* visual representations. These representations will then allow the researchers to rapidly discern temporal and spatial interaction within the network traces.

There are two typical approaches for semantic visualization of network traffic. In the first approach, a specialized visualization tool is created for a particular purpose. The semantics of the network protocol operation, packet headers, and payload are encoded within the tool. For example, DNSViz, the DNS visualization tool allows understanding and troubleshooting deployment of DNS Security Extensions (DNSSEC)[3]. Similarly, tcptrace, enables detailed analysis of TCP-based flows [7]. In the second approach common network analysis tools, such as wireshark [9], FloVis [6] and zenmap [10] include visualization plugins and modules that allow static drilling down of the network data from a high-level summary to packet level details. Typically, these visualizations do not associate any semantic information with the network data and mainly enable the user to zoom in or out and visually identify patterns and groupings within the data. Although, wireshark can be extended with specialized plugins for visual analysis, the semantics required for the visualizations are defined within the plug-in source code.

Generating meaningful visualizations from network data is extremely hard as it requires identifying interesting relationships in large volumes of multivariate, multi-type, timestamped network data. These interesting relationships capture useful semantic associations, such as, ordering, causality, dependence, and concurrency allowing intuitive exploration of the network data. Multiresolution control allows quantization of the multivariate space progressively into smaller spatio-temporal regions to gradually and dynamically analyze the data at several distinct levels of detail or resolution. A fully detailed visualization can be distracting, and may not allow identifying interesting and evolving relationships within the data. Our methodology provides an intuitive and interactive interface to control the level of semantically relevant detail to enable answering the analysis questions at hand and promote visual thinking.

In this paper, we focus on the problem of generating *semantically relevant multiresolution* visualizations of network traffic. We first introduce an example cyber security incident, a DNS cache poisoning attack, commonly seen on the Internet in Section II and then in Section III, we discuss the various types and levels of semantic relationships using the example. In Section IV we discuss how we extend our previous work to create temporal and spatial visualizations of semantically relevant relationships identified by the semantic analysis framework [8].

The fundamental contribution of this paper is a multiresolution semantic visualization methodology for confirmatory and exploratory analysis of multivariate, multi-type, time-ordered network data. The methodology leverages the semantic analysis framework to extract interesting behaviors from network data at multiple resolutions specified intuitively and interactively by the researcher.

II. An Example: DNS Cache Poisoning Experiment

In this section, we discuss a well known cyber security attack that is used throughout the rest of the paper as an example to illustrate the multiresolution semantic visualization methodology.

In Dan Kaminsky's popular DNS attack [4], the attacker's objective is to poison the DNS cache of a victim nameserver in order to redirect all domain name resolution requests for a domain, for example eby.com, to a fake name server.

978-0-7695-4492-2/11 $26.00 © 2011 IEEE
DOI 10.1109/ICSC.2011.80

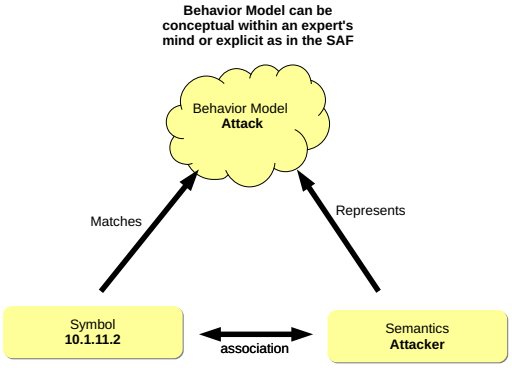


Fig. 1. Associating semantics to network traffic

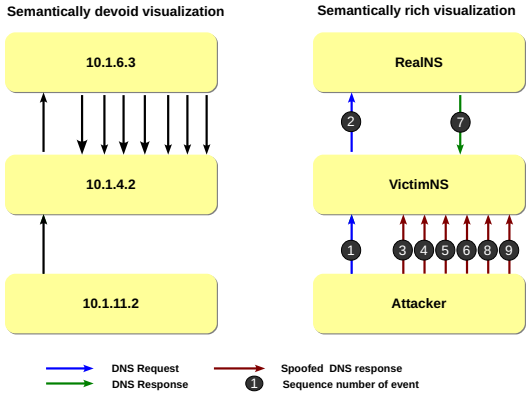


Fig. 2. Semantic visualization of the DNS cache poisoning attack

The attack begins when *attacker* sends a DNS query to *victimns* for some non-existent domain name under eby.com. The non-existent domain is created by appending a randomly generated string to the domain name, for example, azxs83.eby.com, instead of www.eby.com. Since this domain name is unique, it does not have an entry in the DNS cache, hence the *victimns* forwards the DNS query to an upstream name server *realns*. The *attacker*, meanwhile, starts sending forged DNS responses in an attempt to poison the cache of *victimns* with the address of the fake name server. There are two conditions required for the DNS Kaminsky attack to work; (a) Attacker must be able to guess the correct DNS transaction identifier and source port used between the *victimns* and the *realns*, (b) The forged responses from the attacker to the victim nameserver should arrive before the real responses from the real nameserver. The attack exploits a race condition to poison the DNS cache at the *victimns*. We recreated the attack on the DETER testbed [2] for the analyses discussed in the next section.

III. Semantic Analysis

Semantic analysis provides the ability to associate meaning or significance to the addresses, individual messages or groups of messages observed in network traffic. For example, in Figure 1, the behavior of the host 10.1.11.2 within the network trace is consistent with the behavior of an attacker as defined in the attack model. The host is hence associated with a symbolic name of an attacker. If the analysis is done manually, the behavior model of the attacker is conceptual and within the user's mind. If the analysis is automated, the behavior model of the attacker is encoded within the source code of the analysis tool or defined explicitly and read by the analysis tool. In this section, we discuss the semantic analysis framework (SAF), that enables automated analysis of multivariate, multi-type, timestamped network data [8].

The main elements of SAF include (a) a specialized formal language for specifying behavior models and (b) an analysis engine for extracting instances of user-specified behavior models from data. At any time, the condition of a networked system or process can be captured as a sequence of *states*, where a state is a collection of attributes and their values. A *behavior* (b) is a sequence of one or more related states and a *behavior model* (ϕ) is a formula that makes an assertion about the overall behavior of the system. For example, consider a simplified IP flow in networking, where a flow is a communication between two hosts identified by their IP addresses. For simplicity we assume an IP flow to be broken into two states: `ip_s2d` denotes a packet from some source to destination host and `ip_d2s` denotes a packet from a destination to source. Then, a valid IP flow behavior, `IPFLOW`, is one where `ip_s2d` and `ip_d2s` are related by their source and destination attributes with the additional criteria that `ip_d2s` always occurs after `ip_s2d`. The behavior model (ϕ_{ipflow}) is an assertion that `IPFLOW` is valid.

The language proposed in SAF combines operators from Allen's interval-temporal logic [1], Lamport's Temporal Logic of Actions [5] and boolean logic. Temporal logic allows expressing the ordering of events in time without explicitly introducing time. Interval-temporal logic allows expressing relationships like concurrency, overlap and ordering between behaviors as relationships between their time-intervals. Additionally, complex behaviors are easily composed from simpler ones using boolean operators. These operators help encode a large range of interesting relationships in network traffic. Also, the model constructs within SAF enable specifying dependency relationships between event attributes while leaving the values to be dynamically populated at runtime.

Thus, behavior models encode interesting relationships between two or more network events, or collection of events. These interesting relationships are primarily of two types (i) factual or true relationships that are present due to well defined or standard network protocol interactions; and (ii) useful or valuable relationships that the researcher wants to identify

through the behavior model. The first type of relationships are typically ground truths and capture the relationships that do not change with time. The second type of relationships are typically analyses dependent and ephemeral and used to highlight the relationships the researcher want to explore.

We analyzed the network traffic traces generated from the DNS cache poisoning experiment on the DETER testbed [2]. The success of such attacks rely on subtle packet content and ordering information, and hence it is extremely challenging to determine the experiment behavior manually. Figure 2 shows how semantically rich visualizations can be created leveraging our methodology. The behavior models allow identifying three types of semantic relationships: (i) nodes are associated with semantically relevant names; (ii) network events can be ordered, visually coded, and dependencies and causal relationships can be identified; (iii) forged DNS response originating at the *attacker* are distinguished from response originating at the *realns*. These semantic cues are completely absent in the visualization on the left.

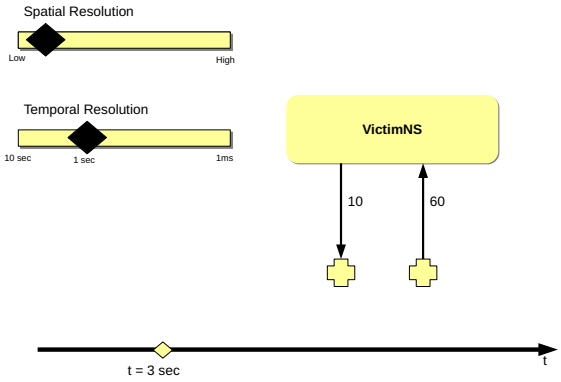
Such semantic analyses and visualizations provide an intuitive way to analyze large volumes of network data with complex relationships. In the next section, we discuss how multiresolution analysis can be used to generate interactive visualizations.

IV. Multiresolution Visualization

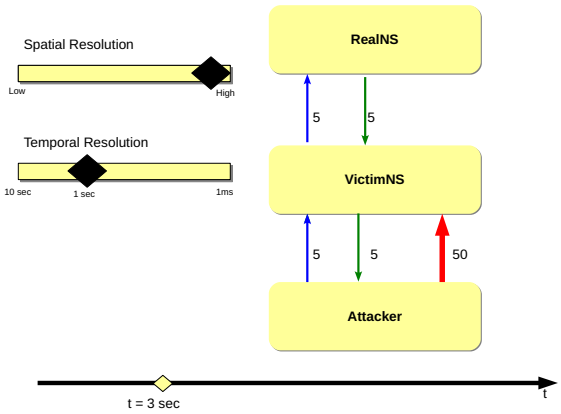
Visual exploration is an indispensable technique in the analysis of spatial and temporal network data as the visualizations allow the researcher to examine data, conceptualize, and search the data creatively to reveal interesting relationships. In this section we discuss an interactive visual exploration technique to analyze spatio-temporal changes in network traffic through animated layouts.

Network data is highly voluminous. Our strategy for exploring such large volumes of temporal and spatial network data, is to create multiple levels of abstraction or resolutions by decreasing the detail in the visualization using temporal and spatial summarizations. Hence the multiresolution visualizations as shown in Figure 3 are a conscientious effort to present the data at several levels of detail or resolution in the hope of revealing interesting patterns and relationships in the data. The interactive multiresolution control allows the user to systematically explore the data by quantizing the multivariate space into smaller and smaller spatial and temporal regions. Although the example below explores DNS request–response pairs, several other types of information can be simultaneously explored.

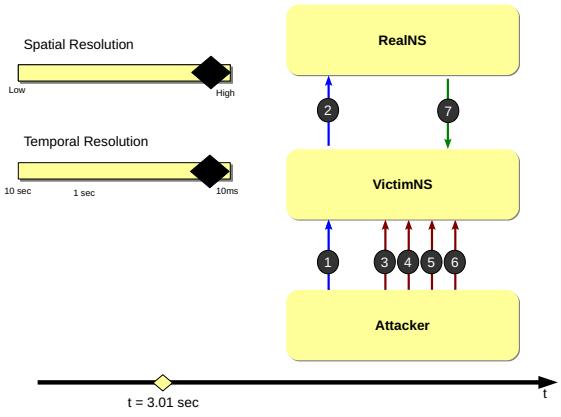
Formally, given a finite sequential trace T of events and a user-defined behavior model ϕ, the SAF as discussed in the previous section, finds all behavior instances $(B^1_\phi, B^2_\phi, \ldots)$ from T that satisfy the behavior model. The goal of the multiresolution visualization, is to then create *temporal* summarizations and *spatial* summarizations of the behaviors B_ϕ to create a unified visualization of all the semantically relevant relationships in the trace T.



(a) Low spatial resolution, medium temporal resolution



(b) High spatial resolution, medium temporal resolution



(c) High spatial resolution, high temporal resolution

Fig. 3. Interactive multiresolution control allows the attribute space to be quantized into smaller temporal and spatial regions.

Spatial Summarization. Spatial summarizations are typically useful in large networks or in highly connected networks where there is a significant amount of spatial redundancy in semantically relevant relationships. For example, a single node can be used to represent a group of end hosts in an enterprise network topology if they have similar communication patterns. The first step towards constructing a low resolution spatial summarization is to aggregate all the behaviors seen at a central node. The centrality of a node is currently guided by the behavior models but in the future, we will explore other measures of centrality. All the flows ingressing and egressing a node are summarized at the central node. In the subsequent levels, the flows are further categorized based on flow destination and additional attributes. Referring to Figure 3(a), at a low spatial resolution, we can observe a single node, the *victimns* with flows originating and terminating at the node. The annotations next to the arrows indicate the number of flows observed during the current time period set by the temporal resolution control. At the highest spatial resolution, the complete network of nodes is expanded allowing visualization of more detail, such as, distinguish between the direction and type of the flows during each time period. As shown in Figure 3(b), keeping the temporal resolution constant, the spatial resolution is increased to show additional nodes within the network layout. Thus the increased spatial resolution enables the researcher to observe detailed source and destination flow information. Due to the simplicity of this example, we consider only two levels of resolution, but if the network is large and the interactions are more complex, there can be several levels of spatial resolution.

Temporal Summarization. Temporal summarizations are useful to observe trends or change patterns in network data by aggregating groups of events based on time. For example, at low temporal resolution the aggregation time period is large and several individual events are represented by one high-level event, such as, a TCP connection setup, data transfer, and TCP connection teardown can be represented by a single TCP flow event. Similarly, at higher temporal resolutions, the time period for aggregation of events is much smaller, and at the highest resolution, each event is individually represented to create a detailed per event visualization. The first step towards constructing a low resolution visualization, consists of reducing event stream at each node to categorize all the events based on common attributes, such as the same source and destination pairs. Each subsequent resolution level, can then successively introduce additional attributes to subdivide each categorization further. Referring to Figure 3(b), at a 1 second resolution, we can observe all the types of flows that are active during each time period. The annotations next to the arrows indicate number of flows active during the time period, that is, during the last one second time period (2nd second to the 3rd second in the experiment), there where five DNS requests that were sent by the *attacker*, five DNS requests that were sent by the *victimns*, five DNS responses sent from the realNS and *victimns* respectively. Additionally, the *attacker* also sent 50 spoofed attack flows. At a higher level of temporal resolution, we can observe detailed packet and ordering information and identify flow sequences and lifetimes during each time period. At this level of visualization, all the attributes are used to uniquely identify each event. Hence during the last 10ms time period, lasting from the 3.0 seconds to 3.01 seconds one DNS request packet was send from the *attacker* to the *victimns*, the *victimns* forwarded the packet to the *realns*. Additionally, there were four attack packets sent from the *attacker* to the *victimns*. The *realns* also sent an authentic DNS response back to the *attacker*. For this example, the temporal aggregation of events of each resolutions was done manually however, in the future we will explore using wavelet-based techniques for temporal summarization.

V. Conclusion and Future Work

In this paper, we presented a methodology for multiresolution semantic visualization of network traffic. These visualizations can be created using the semantic analysis framework with user-defined behavior models which allow identifying patterns visually by: (i) associating semantically relevant names to hosts in the network traffic; (ii) visually coding dependencies and causal relationships in the network traffic; and (ii) distinguishing between forged and real network traffic. The multiresolution control interface allows the user to intuitively and interactively control the level of detail using temporal and spatial summarizations. This approach allows the user to systematically explore the attribute–value space by progressively partitioning it into smaller and smaller spatio–temporal regions.

Generating effective visualizations for network data is extremely challenging for two reasons. First, in large scale systems, efficient and effective visualizations is extremely resource intensive due to the sheer volume of traces and logs. Second, the definition of "interesting" varies widely in different situations and can greatly influence the level of detail shown in a visualization. As future work, we will explore various semantic analysis and summarization algorithms for cyber security and networking scenarios.

Acknowledgements: This work is funded by the Department of Homeland Security under Contract No: N66001-10-C-2018

References

[1] J.F. Allen. Maintaining Knowledge about Temporal Intervals. *Communications of the ACM*, 26(11):832–843, November 1983.

[2] Terry Benzel, Robert Braden, Dongho Kim, Clifford Neuman, A. Joseph, K. Sklower, R. Ostrenga, and S. Schwab. Experience with DETER: A Testbed for Security Research. In *2nd Intl. Conf. on Testbeds and Research Infrastructures for the Devel. of Networks and Communities - TRIDENTCOM*, page 10, 2006.

[3] DNSViz:a dns visualization tool. http://www.dnsviz.net/.

[4] D. Kaminsky. Multiple DNS Implementations Vulnerable to Cache Poisoning. http://www.kb.cert.org/vuls/id/800113, 2008.

[5] Leslie Lamport. The Temporal Logic of Actions. *ACM Trans. Program. Lang. Syst.*, 16(3):872–923, 1994.

[6] The FloVis System for Network Data Analysis. http://www.flovis.net.

[7] Tcptrace Website. http://www.tcptrace.org/.

[8] Arun Viswanathan, Alefiya Hussain, Jelena Mirkovic, Stephen Schwab and John Wroclawski. A Semantic Framework for Data Analysis in Networked Systems. *Proc. of the USENIX Symposium on Networked Systems Design and Implementation*, 2011.

[9] Wireshark Website. http://www.wireshark.org/.

[10] Zenmap Website. http://nmap.org/zenmap/.