Web Traffic Profiling and Characterization

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

Steve Hutchinson
ICF Jacob and Sundstrom, for
U.S. Army Research Laboratory
2800 Powder Mill Road
Adelphi, MD 20783
301-394-1237
Steve.E.Hutchinson.ctr@mail.mil

Joshua Edwards
ICF Jacob and Sundstrom, for
U.S. Army Research Laboratory
2800 Powder Mill Road
Adelphi, MD 20783
301-394-1578
Joshua.S.Edwards.ctr@mail.mil

ABSTRACT
This research discusses our approach to extract relevant semantic information from web traffic such that host behavior can be characterized. Once the semantic information is extracted, data mining techniques, particularly clustering and visualization, are applied to aid analysis of the data. Specific analysis goals include identification of: changes in behavior, known behavior indicative of malicious activity, identification of host or session function, etc. In aggregation, the capabilities will aid in the management of web based activity. Both LDA and CLUTO have been applied as relevant clustering algorithms. We provide examples of the results of these techniques and discuss the implications of the research.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *clustering, information filtering*; K.6.5 [**Management of Computing and Information Systems**]: Security and Protection – *invasive software.*

General Terms
Algorithms, Management, Experimentation, Security.

Keywords
Web traffic, profiling, characterization, clustering, information retrieval.

1. INTRODUCTION
Managing and securing computer networks has become a challenging enterprise, and for numerous reasons. However, three critical components are the chaotic and complex nature of the underlying data, the variability in attack vectors, and the overlapping use of protocols and function. Obviously, a critical analysis mechanism is narrowing the scope of the analysis. For instance, by focusing specifically on malware, keyloggers, or P2P traffic detection, intrusion detection systems can be made far more effective. This, in effect, results in improved security and management of the network.

Our focus in this research has been on the implications of web traffic profiling and characterization. The focus is to analyze web traffic for individual hosts with the goal of:

- Identifying the function of the individual hosts based solely on their web access patterns. Consider that sets of functions may be specifically malicious.
- Identify changes in behavior, i.e., anomalies that are indicative of changes in function. Ultimately, this may be one indicator of a compromised host.
- Identify *specific* behavior indicative of known malicious activity, i.e., botnets communicating over http ports.
- Identify patterns deemed to be intentionally focused on providing obfuscation.
- Incorporate learned function categories into alert rules for more intelligent event analysis.

The focus on web-traffic is a result of its applicability to malicious traffic. Web traffic easily obfuscates malicious traffic due to its high volume, apparent randomness, and wide acceptance. More sophisticated malware can operate entirely within web traffic protocols, so as to resist protocol and firewall blocking. Many of the control servers appear inter-mingled with normal, benign web-server service points.

Profiling and characterizing activity is a well-known approach to anomaly detection and has been applied to a broad range of domains within cybersecurity. In fact, the first well known intrusion detection paper focused on behavior, Denning [4]. Our focus on web accesses as well as our attempt to characterize function are both novel approaches within this domain.

The critical challenge resolved within this research is the identification of the actual human-associated activity indicative of functional behavior and retrieval of the semantic meaning of that activity. In essence, this required:

- Collecting the semantic context associated with the web-based activity. The semantics can be derived from the URL, meta-tags, and page contents themselves.
- Segregating user initiated from machine initiated actions. This classification in particular aids identification of malicious activity.
- Eliminating redundant and repetitive activity.
- Eliminating unusable and obfuscating non-semantic context.
- Clustering related elements for identification of associations and function classification.

2. Related Work
Profiling has been used extensively in the monitoring of behavior to identify unexpected changes in behavior [5]. Xu [10] applied data mining, e.g., clustering, to communication patterns from flow data in order to identify behavior patterns. This work identified unwanted traffic and anomalies for potential blocking. Sekar [9] provided event monitoring and comparison against known

unacceptable events, allowing identification of known unacceptable behavior. In relation to the monitoring of web traffic, Abdullah [2] had the most relevant work. However, the focus of Abdullah's research was on resource modeling from a performance perspective to assist network management. The goal was to analyze proxy data to characterize user sessions such that network resources can be better allocated for user needs.

3. Semantic Context Extraction

Our approach uses clues from the content, semantics, as well as the timing sequences to identify and assemble the human-caused traffic. As one can imagine, the characteristics of such traffic, the timing, and volumes are quite dependent on the expected use of web technology in the job function performed by a particular host.

The process at large is exemplified in Figure 1. The process contains two primary components. First is the pre-processing phase which prepares the data for the clustering. Second is the analysis stage which takes the clustered documents to identify the function, behavior, relationship, and goals of the document set. The current state of the research has focused most heavily on the pre-processing and clustering phases. We have acquired preliminary results related to the analysis phase and continue to explore and apply the results.

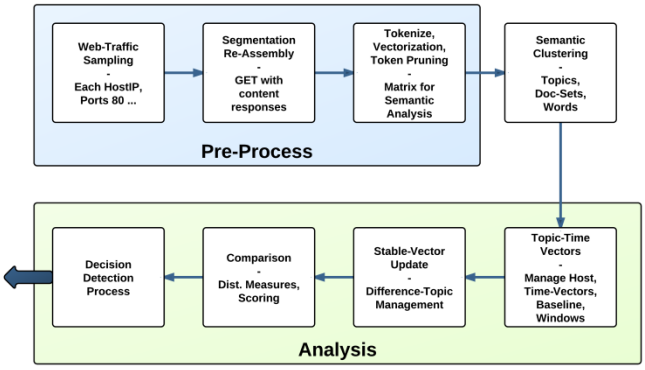

Figure 1: Process diagram exemplifying the pre-processing and analysis process associated with the analysis of host-based web activity.

3.1 Pre-Processing Goals

The nature and use of the web has evolved and changed dramatically, even in the recent half-decade. Web pages used to be predominantly tagged textual content with occasional images. Today, relevant content tokens are rare, and a page typically consists of far greater portions of code, such as javascript, formatting and tagging (style sheets), images and glyphs, iFrames, and multiple layers of content embedding. Animations and periodic updating content lists are commonplace. The token literals comprising this non-topic content often are identifiable, but only in the negative, by observing that they are not an expected, English word/stem.

Unfortunately, proper but unexpected tokens will affect the topic semantic, such as words from an advertisement or popup list box. Even though these tokens are valid they should not be construed as related to the prevailing topic requested by the user.

Our current approach attempts to mitigate the adverse effects of these non-content tokens, and to group the request content with corresponding responses.

3.2 Pre-Processing Algorithm

For this research, the network specifically captures every packet designated to or from port 80. Information collected includes: source and destination IP address, port numbers, protocol, timestamp, IP version, and the string representation of the data itself. Once the data is collected the processing begins:

Step 1: aggregate the data into complete transactions. This requires keeping track of two lists of transactions, i.e., the set of complete and the set of incomplete transactions. New transactions are identified when a "GET" request is observed, completing previous transactions for the associated IP-IP-port triple.

Step 2: only accept files likely to contain readable or useful contents; eliminating file types such as images or css files. This requires checking file extensions against a whitelist, including: html, htm, shtml, php, asp, jsp, and xml.

Step 3: data beyond the GET line in the packet is ignored.

Step 4: packets that do not include a specific GET request are aggregated with existing IP-IP-port triples to complete existing transactions.

Step 5: file meta-data is examined to identify additional files that that will not likely contain readable or useful content, such as javascript.

Step 6: eliminate calls to external CSS files, i.e., the appearance of "@charset".

Step 7: clean out HTTP headers (e.g. User-Agent, Location, etc.): check against the list of known headers and have them removed. The metadata they hold are not needed.

Step 8: sort completed transactions based on the timestamp of their initial GET request.

Step 9: strip embedded Javascript and CSS from file contents. This is primarily text between <script> and <style> tags.

Step 10: all HTML tags are removed. A regular expression ("<[^<]*?>") is used to find and replace all of them with a blank space.

Step 11: eliminate stop words [1]. This removes words that will not aid in identifying the semantics of the activity.

Step 12: apply a stemmer [1]. The stemmer ensures we are always providing analysis on the same ground truth.

Step 13: compare each words against a dictionary of valid word stems. Word stems not found in the dictionary are purged. This ensures that the words we are dealing with will be of assistance in providing semantic context.

Step 14: aggregate machine requested content. A significant source of machine generated content is ads being incorporated into the web page for revenue generation. In essence, if the secondary request associated with the IP-IP-port triple, e.g., the request for the ad image, occurred below a specified threshold time-period, then the request is assumed to have been machine generated. This machine requested content is associated with the most recently occurring transaction. Currently the threshold is set arbitrarily at two seconds. Future research will identify threshold values more scientifically.

The results of this process are exemplified in Table 1. This example clearly shows the removal of enormous amounts of extraneous information, leaving only semantically relevant data. This data can now be used to characterize the function of the host's actions and profile its behavior.

Table 1: A subset of raw data and the resulting cleaned data representing 169 lines of raw data, resulting in 30 lines of comprehensible text. This comprehensible text provides the semantic context associated with the user-directed web activity.

Raw Data	Cleaned Data
 <script type="text/javascript">awo2partner="BALTIMORESUN";</script> <a target="_blank" href="http://www.traffic.com/Baltimore-Traffic/RT-50-WESTBOUND-17_WB_50_35.html?AWOPARTNER=BALTIMORESUN"> 50 RT-50 From N. Philadelphia Av (Ocean City) To RT-90/Ocean City Exwy <script type="text/javascript">awo2partner="BALTIMORESUN";</script> <a target=" i4_blank" href="http://www.traffic.com/Baltimore-Traffic/US-29_Columbia_Pk-SOUTHBOUND-17_29_SB.html?AWOPARTNER=BAL i5TIMORESUN"> 29 US-29 Columbia Pk From RT-99/Old Frederick Rd To I-495 Capital Beltway <script type="text/javascript">awo2partner="BALTIMORESUN";</script> <a target="_blank" href="http://www.traffic.com/Baltimore-Traffic/US-40_Baltimore_National_Pk-WESTBOUND-	HTTP/1.1 200 OK 295 RT-295 Baltimore Washington Pkwy From Russell St To RT-32/Savage Rd 40 US-40 Baltimore National Pk I-70 To I-695 Beltway 50 RT-50 From I-495 Capital Beltway (#7) To RT-295 Baltimore Washington Pkwy 50 RT-50 From N. Philadelphia Av (Ocean City) To RT-90/Ocean City Exwy 29 US-29 Columbia Pk From RT-99/Old Frederick Rd To I-495 Capital Beltway 40 US-40 Baltimore National Pk From I-695 Beltway To I-70 395 I-395 I-95 To Conway St 695 I-695 Beltway (Northside) Moderate From I-95 i5(#33) To I-83 Harrisburg Exwy (#24) 195 I-195 From RT-170 Aviation Blvd (#1) To RT-166 Metropolitan Blvd 295 RT-295 Baltimore Washington Pkwy From RT-32/Savage Rd To Maryland/District of Columbia Line

3.3 Challenges

The organization and design of some websites leads to enormous amounts of irrelevant data. This includes heavily image oriented or Javascript oriented sites. This is further complicated by sites that do not properly label the site's files. Future work will look at employing forensics techniques to aid in identification of the types of data actually in use [8]. This will provide further analysis as to the potential for malicious code to be in use.

The result of cleaning such sites can result in empty documents. These empty files could be considered as false positives for the identification of malicious activity as botnet activity, once cleaned, will likely have a similar appearance.

3.4 Clustering Algorithms

The raw data files are used to generate formatted input for clustering techniques. Each unique word in the file is assigned a unique number. Each document is then given a line with a count of each instance of each word within it.

It is this clustering that goes to the heart of our research goals. The clustering essentially groups documents based on semantic relevance. This will, in essence, identify function and, through outliers, will identify documents not satisfying normal behavior e.g., malicious behavior; this may necessitate incorporating empty documents.

To this end, we employed document clustering techniques based on CLUTO [11] and LDA [6]. Document clustering was chosen due to the textual nature of the data ultimately extracted from the web documents and the desire to extract semantic meaning.

Clearly, even with just individual words it's obvious in the LDA clustering that one of the most common document types, topic 2, is associated with access violation or disallowed web sites. Many of the rest of the topics have a wider array of terms, most likely indicative of news sites.

The standard output for the CLUTO clustering provides fewer words and thus less context but it can still be seen that cluster 4 deals with issues in the local Washington DC area. Additionally, cluster 18 clearly deals with documents associated with malicious code.

3.5 Visualization

The goal of the research is exemplified in Figure 2 using the graphware visualization technique, based on GUESS [3]. This visual representation of web traffic data clearly shows that most of the web traffic is tightly associated and linked together. However, there are a number of disassociated web traffic connections. Of particular interest are the isolated nodes with larger numbers of

Table 2: Example of clustering using the LDA and CLUTO algorithms. Both algorithms were applied to cleaned data. This is a different data set than is shown in Table 1, exemplifying more realistic data. Only a subset of the results is shown to exemplify the results of the algorithms.

LDA Clustering of Cleaned Data	CLUTO Clustering of Cleaned Data
Topic 1 softwar (6) data (51) size (299) name (330) frequenc (333) int (335) time (337) var (339) passback (341) length (343) freq (344) cooki (346) pop (347) copyright (351) imp (352) media (353) tag (354) flash (355) salt (358) arrai (360) ad (367) ban (369) portion (371) emb (381) Topic 2 polici (12) access (33) request (329) forbidden (331) wish (332) result (334) system (336) coat (338) viewabl (340) filter (342) categor (345) assist (348) deni (349) network (350) suspici (356) refer (357) tribun (359) team (361) support (362) pleas (364) disput (365)contact (366) blue (368) question (370) modifi (1181) Topic 3 click (37) nation (97) found (102) domain (202) path (203) talk (204) expir (205) net (206) gmt (207) move (225) biz (239) sci (241) job (242) food (243) sport (244) fin (245) tech (246) temporarili (248) polit (249) fam (252) thu (253) health (309) item (1118) po (1329) Topic 6 visibl (280) window (363) fals (614) function (653) bodi (769) els (949) wed (1328) retarget (1330) substr (1332) practiv (1333)	Cluster 0, Size: 1, ISim: 1.000, ESim: 0.003 100.00% sync Cluster 1, Size: 1, ISim: 1.000, ESim: 0.008 100.00% size emb ad Cluster 2, Size: 1, ISim: 1.000, ESim: 0.008 100.00% atyp cad gen content Cluster 3, Size: 2, ISim: 1.000, ESim: 0.009 100.00% client content Cluster 4, Size: 1, ISim: 1.000, ESim: 0.014 100.00% beltwai pkwy washington ocean columbia Cluster 5, Size: 1, ISim: 1.000, ESim: 0.028 100.00% aq weekdai opinion event baltimor Cluster 6, Size: 1, ISim: 1.000, ESim: 0.037 100.00% detect except potenti danger request Cluster 7, Size: 3, ISim: 0.989, ESim: 0.035 100.00% cid beacon feb wed expir Cluster 18, Size: 2, ISim: 0.588, ESim: 0.016 50.00% trojan gen threat 0.00% trojan gen ransom threat 50.00% trojan zeu gen ransom 0.00% trojan zeu gen ransom threat

connections. What is the cause of these isolated nodes? What types of documents are they accessing? Who is accessing them?

One of the next steps will be to correlate the clustering research with the visualization technique such that an analyst will be able to examine such visualizations and immediately identify the nature of the accesses. This will aid rapid network management in the identification of potential malicious code.

4. Future Work

A significant research question arises from our unique application of document clustering: What modifications to natural language processing (NLP) and computational linguistics must be made to properly deal with web-traffic content to allow revelation of topic semantics?

For example, it will aid effectiveness to associate the web page URL with response content; i.e., the URL can provide as much meaning in some cases as the actual words in the document. The structure and nature of the URL will typically not survive the identified pre-processing. Of additional interest is the role of punctuation in web-traffic which can act as a token concatenation operator rather than a delimiter as would be expected in typical English document analysis.

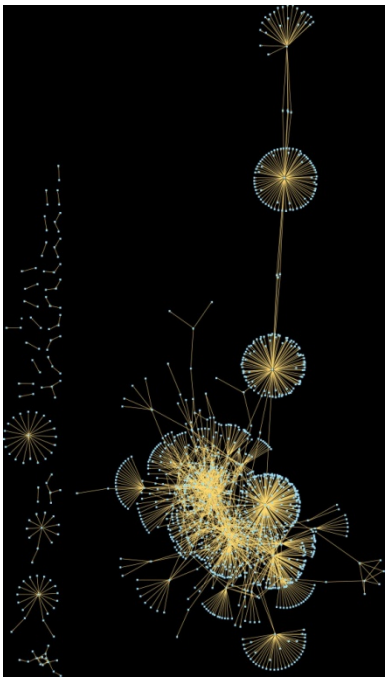

Figure 2: Graphware visualization of web traffic. The primary activity is exemplified in the connected structures on the right of the display. Of interest are the unconnected structures at the left of the display, especially the isolated highly active nodes.

5. Conclusions

The current results of the clustering, while appearing chaotic in nature, already provide sufficient semantic context to identify the basic focus of the documents. This is exemplified by topic 2 from the lda clustering and cluster 18 from the CLUTO clustering. Clearly, however, the semantic meaning of many of the other clusters and topics can be difficult if not impossible to interpret. Is this context sufficient to identify clusters of interest to network management and computer security?

Lastly, we need to perform a study of supervised web-traffic generation so that we can associate the actual user intent from the content that survives the pre-processing, so as to determine the effectiveness of the current process.

6. References

[1] Eija Airio. 2006, "Word normalization and decompounding in mono- and bilingual IR," *Information Retrieval*, Vol. 9, No. 3 (June 2006), pp. 249-271.

[2] Ghaleb Abdulla, *Analysis and Modeling of World Wide Web Traffic*, PhD. Dissertation, Virginia Polytechnic Institute and State University, Computer Science Department, May 1998.

[3] Adar, E.; Miryung Kim, "SoftGUESS: Visualization and Exploration of Code Clones in Context," *Software Engineering, 2007. ICSE 2007. 29th International Conference on* , vol., no., pp.762-766, 20-26 May 2007

[4] Denning, D. E., "An Intrusion-Detection Model," *IEEE Trans. on Software Eng.*, Vol. SE-13, No. 2, Feb. 1987, pp 222-232; also in Proc. of the 1986 Symp. on Security and Privacy, IEEE Computer Society, April 1986, pp 118-131.

[5] Tom Fawcett and Foster Provost. 1999, "Activity monitoring: noticing interesting changes in behavior," In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '99). ACM, New York, NY, USA, 53-62.

[6] R.A. Fisher, "The Statistical Utilization of Multiple Measurements," *Annals of Eugenics*, vol. 8, pp. 376-386, 1938.

[7] Anup K. Ghosh, Aaron Schwartzbard, and Michael Schatz. 1999, "Learning program behavior profiles for intrusion detection," In *Proceedings of the 1st conference on Workshop on Intrusion Detection and Network Monitoring - Volume 1* (ID'99), Vol. 1. USENIX Association, Berkeley, CA, USA, 6-6.

[8] Sarah J. Moody and Robert F. Erbacher, "SÁDI Statistical Analysis for Data type Identification," Proceedings of the *3rd IEEE International Workshop on Systematic Approaches to Digital Forensic Engineering*, Oakland, CA, May 2008, pp. 41-54.

[9] R. Sekar, T. Bowen, and M. Segal. 1999, "On preventing intrusions by process behavior monitoring," In *Proceedings of the 1st conference on Workshop on Intrusion Detection and Network Monitoring - Volume 1* (ID'99), Vol. 1. USENIX Association, Berkeley, CA, USA, 4-4.

[10] Kuai Xu, Zhi-Li Zhang, and Supratik Bhattacharyya. 2008, "Internet traffic behavior profiling for network security monitoring," *IEEE/ACM Transactions on Networking*, Vol 16, No. 6 (December 2008), pp. 1241-1252.

[11] Ying Zhao and George Karypis. 2002, "Evaluation of hierarchical clustering algorithms for document datasets," In *Proceedings of the eleventh international conference on Information and knowledge management* (CIKM '02). ACM, New York, NY, USA, 515-524.