Online Behavioral Analysis and Modeling Methodology (OBAMM)[1]

David J. Robinson[†], Vincent H. Berk[†], and George V. Cybenko[†]

[†]Firstname.M.Lastname@Dartmouth.EDU, Thayer School of Engineering at Dartmouth College, Hanover, New Hampshire

Abstract This paper introduces a novel method of tracking user computer behavior to create highly granular profiles of usage patterns. These profiles, then, are used to detect deviations in a users' online behavior, detecting intrusions, malicious insiders, misallocation of resources, and out-of-band business processes. Successful detection of these behaviors significantly reduces the risk of leaking sensitive data, or inadvertently exposing critical assets.

1 Introduction

The World Wide Web (WWW) has become the single largest repository of information in the world, with people utilizing it to address all aspects of their home and work lives. As our dependence on the web increases, so does the amount of information that can be gathered about an individual using the Internet. E-commerce and marketing firms have taken advantage of this fact for years by accumulating information on individuals for purposes ranging from tailoring ad campaigns to personalizing a shopping experience. Although of obvious importance to marketing firms, this data offers potential benefits in other areas as well. So far, in computer security and intellectual property protection, little attention has been paid to the area of user behavioral profiling. While profiling is generally frowned upon when dealing with personal privacy issues, the ability to accurately profile and depict user activity could eliminate many forms of computer related criminal activity. Using profiling information in the area of computer security can aid significantly in the detection of rogue users, malicious activity, policy violations and unauthorized data exfiltration, and in many cases will even prevent them from happening.

[1] This work was supported under DHS contract number 2006-CS-001-000001. The views expressed in this work are those of the authors alone, and do not necessarily represent the official position of the US DHS.

In this paper, we suggest an Online Behavioral Analysis and Modeling Methodology (OBAMM) to accurately and efficiently categorize users based on their individual web browsing activities and propose how this information may be used in the realm of computer security.

Section 2 of the paper provides background on other approaches that take advantage of profiling information. Section 3 briefly describes OBAMM followed by initial results in Section 4. Section 5 proposes how this technology can be used in computer security with Sections 6 summarizing and proposing future work.

2 Background

E-commerce and marketing firms have taken advantage of profiling for years by collecting volumes of information on individuals. Such profiling is accomplished by aggregating information on individuals purchase history (online and offline), finance records, magazine sales, supermarket savings cards, surveys, and sweepstakes entries, just to name a few. This information is then cleaned, organized, and analyzed using a number of statistical and data mining techniques to create a "shopping" profile of that individual. These profiles can then be used to target ad campaigns, personalize a shopping experience, or make recommendations on additional products a user may find they "can't do without". Amazon (see http://www.amazon.com) is a perfect example of the effective conduct and implementation of this type of profiling with its tailored "*recommendation"* and "*customers who bought"* features in addition to their personalized e-mail campaigns that let users know when items you may be interested in or have looked at in the past go on sale.

Network traffic profiling, on the other hand, is an emerging area quickly filling up with commercial products that record and plot network usage by selected characteristics. These profiles, however, operate only at the level of network sessions and flows, while application level information, which is crucial in user behavioral profiling, is not used. Products like Qradar (see http://www.q1labs.com) and Mazu Profiler (see http://www.mazunetworks.com) consider TCP/IP traffic patterns as memory-less indicators of behavior. While interesting, we seek a more comprehensive profile, targeted towards user behavior, instead of straightforward host traffic graphing.

A final area that has seen great success in the field of behavioral modeling is business process modeling. Although primarily used as a tool to model work flows in an organization for the purposes of optimization of productivity and efficiency, much research has been done on how to model an individual user's actions and key characteristics. Data collection is traditionally done through the use of questionnaires, interviews, and direct observation, and the field provides a great deal of information and valuable lessons learned on what data is needed and how it can be best utilized to create an accurate profile of an individual. Our automated data collection techniques have the potential to advance the science of business process modeling by providing an easy and efficient way to test hypotheses and gather observations.

3 Approach

OBAMM is a new technique that uses information about a user's web browsing activities only, to accurately categorize what the primary interest areas are of that individual. We passively sniff network traffic to obtain browsing information, not instrumentation of the machine's browser application. Based on what we term "*reverse category lookup"*, OBAMM provides for very accurate categorization models to be built from fairly minimal user data. While fundamentally different in its implementation, the approach being proposed follows the same general methodology used in data mining (see [9]); collect data, process data, discover patterns, analyze patterns. The remainder of this section will briefly describe how each of these is accomplished using OBAMM.

3.1 User Data

With online users taking advantage of the Internet for everything from research, to hobbies, to online shopping, it would seem that all the information needed to describe a user is ready and available. While research in specific areas has been done to take advantage of users browsing activities in areas such as improving searches in peer-to-peer networks [10] and recommender systems for publications and retail [11, 12], little has been done to use this information to create a more complete user behavioral profile. The critical piece of information that provides the details needed for these types of applications comes in the form of the Uniform Resource Identifier (URI). The URI represents the global address of documents and resources that are present on the World Wide Web. The most common form of the URI is a web page address. When a user operates their web browser and requests a URI, or clicks a hyperlink on an existing page, a HTTP/GET request is generated. This GET request can be sniffed and captured, pulled from log files, or captured by agent devices installed on the user's machine. Timestamps can be added by recording the time that the GET request was captured, while the URI, the destination host, and assorted browser information are available directly from the request. By collecting the URIs that a user has visited, it is possible to have access to a large portion of the information that a given user has viewed in a period of time. By taking into account sites visited, frequency of visits, and duration of visit, it is straightforward to abstract information that provides a sketch of who an individual is.

3.2 Categorization Data

The key to OBAMM is the existence of a reverse lookup category data repository. This data comes in various forms, but for the purposes of this research, we are only interested in URIs that have been categorized in some hierarchical manner. An example of this type of hierarchical data can be seen in the Open Directory Project (ODP). ODP is an open content directory of World Wide Web links that has been constructed and categorized by humans. Web URIs based on similar

content are grouped at a high level category while lower level sub-categories define varying levels of specificity for each site. For example, category information for http://www.dartmouth.edu would return the following category hierarchy:

Reference: Education: Colleges and Universities: North America: United States: New Hampshire: Dartmouth College

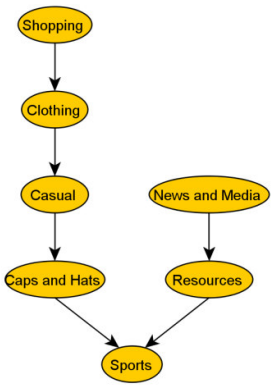
This information can be transformed into a directed graph G=(V,A) where V is the set vertices represented by the category description and E is the set of edges connecting each category. Representing the above category in this manner would yield the directed graph in Figure 1.

Figure 1: Graphical Representation of Category Information

As reverse categorization is done on additional URIs, the respective graph information can be added creating a graph representation of the users browsing activity. Edge weights are added to represent frequency while general category interest is represented by node weight. Building the graph in this manner provides an easy way to visualize a user profile as well as see data correlations that may have otherwise gone unnoticed. For example if dealing with the URIs http://espn.go.com and http://www.vivacheap.com it may not be intuitively obvious of any relation between these two sites. Storing the category information as a directed graph allows this correlation to be determined both visually (Figure 2) and mathematically using matrix algebra. Constructing the graph that represents this information illuminates the fact that both URIs share a common sports node. This may signify the beginning of a trend that can be used to begin to describe the individual. In addition, clustering techniques can be employed in order to allow the data to be viewed at varying levels of abstraction while pruning algorithms may be used to filter potential “noise”.

A number of resources in addition to ODP provide commercial and open source options for gaining this type of URI categorized information (YellowPages.com, http://kc.forticare.com/). An important point that must be considered when utilizing multiple sources for this type of information is that they all must be normalized to the same format and category structure before being used for reverse category lookup.

Figure 2: Visual Correlation of Seemingly Disparate Data



3.3 Pattern Discovery/Analysis

The actual format of a user profile is at the core of our technology. While a number of commercial tools exist today that claim to monitor and track user profiles and behaviors, most are only working at the TCP level and are doing little to identify or describe an individual based on what they are actually doing. When we consider behavioral profiles, three distinct orders of models can be described:

1. 0^{th} order models: binary event recorded only, for instance: the list of websites that were visited.
2. 1^{st} order models: frequencies, probability distributions, for instance: a Bernoulli style model indicating the likelihood that a site will be visited, based on a frequency count of previous visits.
3. 2^{nd} order models: causality relations, time-dependencies, for instance: a hidden-Markov style model, with transitional probabilities between site accesses.

Most behavioral profiles fall in the 1^{st} order model category, meaning they do not model important time-ordered sequences of events. For instance it is more likely that someone will visit their preferred shopping website first before going to others. Likewise, most people visit news and email websites, on a daily basis, in exactly the same order. These are important signatures of user behavior. In our work, we consider a user profile as a record of user browsing activity and can be described by the following key characteristics; destination, frequency, duration, and order. Destination represents where the user is actually browsing to. This information provides details on likes, dislikes, hobbies, interests, and other details relating to the individual. Frequency represents the number of individual times a destination is visited. This information combined with duration alludes to the importance of that data to the given individual. Order relates to the sequencing of the browsing activities from which it is possible to derive patterns of behavior about a given user. Although based solely on a users browsing activities, the aggregation of this information provides detailed attributes of the individual that can be used for the purposes of classification. By basing our profile on a combination of 0^{th}, 1^{st}, and 2^{nd} order models, we have the ability to not only graph what the user is doing, but how they are doing it. Figure 3 is a graphical depiction of the catego-

ries of information describing a given user and the temporal interaction between them.

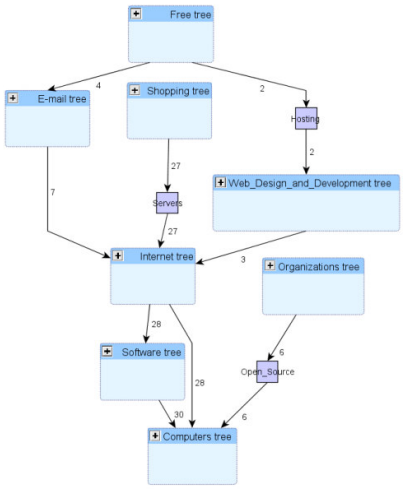
Figure 3: Temporal Aspects of Categorical User Data

4 Experimental Results

To test OBAMM, we chose to use ODP and a portion of the blacklist archive from URL Blacklist.com as our categorization dataset. At the time of the experiment, the DMOZ data set contained 4,830,584 URIs broken into sixteen main categories while the blacklist data contained 2,555,265 URIs split into 69 categories (all of which represented either primary or sub-categories of the DMOZ data). Normalization was done to the blacklist data in order to match the format and overall category structure of the DMOZ data set. The combined category archive was then hashed by URI. While it is understood that more optimal methods exist for the storage and retrieval of this type of information, for the purposes of this experiment, a has proved sufficient in regards to speed and accuracy.

Once all categorization data was collected and stored, approximately one week's worth of sample data was obtained from our local research lab, monitoring consenting individuals only. Data was preprocessed to include just the HTTP GET requests. The data was then broken out by source IP (where it is assumed that each unique source IP represents a unique user) and cleaned to remove adware and other automatically generated GET requests. The final data consisted of files each representing an individual user containing URIs of all the sites visited by that user. Data from each file was then run against the categorization hash in order to extract the category information required to construct a directed graph with category names representing nodes and link weights representing the number of times the user visited a site relating to that category. Graph information was stored in GraphML (see http://graphml.graphdrawing.org) file format so that it could be easily viewed, analyzed, and manipulated using graphical tools such as yED (see http://www.yworks.com). Figure 4 shows a subset of a user graph, representing the users browsing habits as they relate to the category *Computers*.

Figure 4: Subset of a directed graph representing a user's browsing habits as they relate to computers



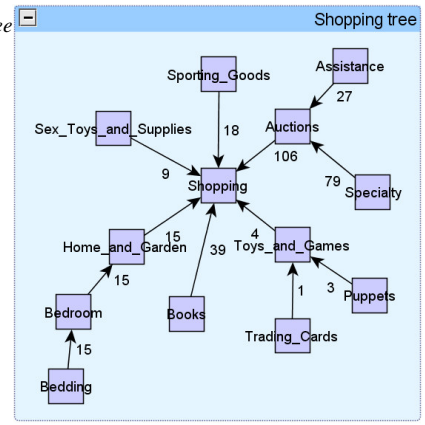
As can be seen in Figure 4, it is immediately apparent at a high level what type of activities the user was taking part in as well as the frequency of those activities. One of the keys to OBAMM, however, is the ability to view the users activity at varying levels of abstraction. Each of the larger boxes in the figure actually represents a grouping of two or more nodes in undirected subtrees of the principle node. Figure 5 is an example of how the *Shopping tree* node can be further expanded to display more detailed specifics about the user. From here, specific details can be obtained as to specifics about the individual, in this case in the form of shopping preferences.

Five individuals were broken out of the data set and categorized in this manner with approximately one week's worth of data for each user. Overall, 63% of the URIs were accurately categorized using only ODP and one category from the blacklist data set. While not a seemingly significant number at first glance, a number of factors are worth noting. First is that only one true categorization source was utilized (only one category from the black list data was used). Second is that because a hash was used, only exact matches were categorized. After closer examination of the user data, it is believed that we could increase the categorization percentage by 15-20% by utilizing a data structure that allows for fuzzy matching.

5 Security Applications

Using this methodology, we were able to quickly generate profiles summarizing the topics of interest to each user, how much time is spent on each topic, and how many times a topic or URI is visited.

Figure 5: Expansion of *Shopping tree* category node



These profiles are very unique indicators of a persons online behavior, provide insight in to a number of key interest areas:

1. Malicious compromise. Hosts showing increased levels of activity at times of day, or volumes that are atypical for the profile on record. This could indicate a compromised machine exfiltrating sensitive information to an outside attacker.
2. Activity outside of normal work hours can be an indication of a malicious insider using the trusted environment to obtain and distribute sensitive documents.
3. Business process modeling. Unusually high levels non-work related browsing activity is an indication of less than optimal work efficiency by employees.
4. Out-of-band business processes. Two or more users in a network spend unusual amounts of time communicating with the same group of websites might indicate an ad-hoc business process that was created by users to circumvent an inefficiency in the organization. Not only can additional efficiency be gained by detecting and optimizing this step, it is also a security risk when an outside third-party is used for exchange of sensitive, mission critical information.

Although these security and efficiency implications are obvious, the potential impact of this intelligence is not. Many of today's successful organizations are held together by tightly guarded intellectual property that is at constant risk of compromise, either by insiders, outside attackers, or negligent behavior. Off-the-shelf network security solutions focus on packet or file monitoring for known-bad signature detection. So far little effort has been expended to detect users' behavior, which is the ultimate objective of our research. A direct correlation to this can be seen in the areas of business processes and trust models as they relate to computer security. While tools and techniques are in place to detect direct and malicious attacks against an information system, less obvious and often times unintentional violations of business processes and trust models can often be more harmful. One of the issues with detecting these types of violations is that business processes tend to address the more general aspects of security rather than specific

standards and protocols (i.e. not storing corporate data on public resources) and because of this, make detection difficult to impossible.

6 Summary and Future Work

In this paper we have presented a new methodology to conduct behavior modeling using pre-categorized data and emphasized how this data can be used in the field of computer security. While not a "silver bullet" in and of itself, our initial testing has demonstrated the huge potential for using this type of information. A network security solution that monitors email for IP leakage is easily circumvented by using a printer, or a USB memory stick. The detectable event, however, is therefore not the actual exfiltration, instead it is the process of the user obtaining the data, and the intent to illegitimately redistribute. Since this is a "behavior", not an "event", some heuristic or stochastic estimation needs to be applied to deduce intent. We strongly believe that creating models of user behavior is the first step in this process, and that future research will decide if behavior profiling is a usable, and morally justifiable way of limiting an organizations exposure to IP related risks.

A number of key factors still require additional research for this methodology to be even more effective. While we are currently storing temporal information as it relates to users and categories, little is being done with it at this time. Research into how temporal aspects of this information can be used to effectively describe an individual is still needed. In addition, clustering and pruning algorithms need to be implemented more effectively to ensure to information of value is not inadvertently lost or mis-categorized. Lastly, research into how to handle both search engine queries and URIs that cannot be categorized using this technique is required for the system to be complete.

References

1. Brown B, Aaron M (2001) The politics of nature. In: Smith J (ed) The rise of modern genomics, 3rd edn. Wiley, New York.
2. Olmez A E (2006), Exploring the Culture-Performance Link in Heterogeneous Coalition Oganizations. PhD Thesis, George Mason University, Fairfax, VA.
3. Smith J, Jones M Jr, Houghton L et al(1999) Future of health insurance. N Engl J Med 965:325-329
4. South J, Blass B (2001) The future of modern genomics. Blackwell, London
5. Tantipathananandh C, Berger-Wolf T Y, Kempe D (2007) A framework for community identification in dynamic social networks. In: proceedings of the 13th international conference on knowledge discovery and data mining 717-726
6. V. Berk an G. Cybenko, "Process Query Systems", IEEE Computer, January 2007, p 62-71
7. Qradar, http://www.q1labs.com
8. Mazu Profiler, http://www.mazunetworks.com
9. Ian H. Witten and Eibe Frank, "Data Mining, Practical Machine Learning Tools and Techniques", 2nd edition, MK publishers 2005
10. Lu J, Callan J (2006) User Modeling for Full-Text Federated Search in Peer-to-Peer Networks. In: Annual ACM Conference on Research and Development in Information Retriev-

al Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval 332-339
11. Quatse, Jesse and Najmi, Amir (2007) "Empirical Bayesian Targeting," Proceedings, WORLDCOMP'07, World Congress in Computer Science, Computer Engineering, and Applied Computing
12. Parsons, J., Ralph, P., & Gallagher K. (2004) Using viewing time to infer user preference in recommender systems. AAAI Workshop in Semantic Web Personalization, San Jose, California, July.