Web Adoption: An Attempt Toward Classifying Risky Internet Web Browsing Behavior

Alexander D. Kent
Los Alamos National Laboratory

Lorie M. Liebrock
New Mexico Institute of Mining and Technology

Joshua Neil
Los Alamos National Laboratory

Abstract

- **Background.** This paper explores associations of computer compromise events in relationship to web browsing activity over a population of computers.
- **Aim.** Our hypothesis was that computers are more likely to be compromised in comparison to other computers when the computer regularly browses to web sites prior to other computers visiting the same site (early adopters) or browses to unique web sites that no other computer visited (unique adopters) in a given time period.
- **Method.** Web proxy data and associated computer-specific compromise events covering 24,000+ computers in a contiguous 6 month time period were used to group computers in various adopter categories and compare potential compromise events between the groups.
- **Results.** We found distinction in web surfing behavior, in some cases differentiating the chance of compromise from 2.5-fold to over 418-fold between certain adopter categories. However, the study also showed no additional value in predicting compromise using these more complex adopter categories when compared to using simple unique web activity counts. As additional contributions, we have characterized several large, real-work cyber defense relevant data sets and introduced a method for simplifying web URLS (client web requests) that reduces unwanted uniqueness from dynamic content while preserving key characteristics.
- **Conclusions.** We found that a count of unique web visits over time has the same level of predictive power for potential compromise as does the more complicated web adopter model. Both models have better than chance levels of prediction but also reinforce the idea that many factors beyond elements of web browsing activity are associated with computer compromise events. Nonetheless, our adopter model may still have value in objective computer risk determination based on web browsing behavior.

1 Introduction

Using a web browser application on a computer to query and fetch information from the Internet is a primary and regular activity for many computers within an organization. Unfortunately, web browsing is also a primary vector for a computer to become compromised by malicious entities. According to Symantec's 2011 Internet Threat Report, web-based attacks have increased by 36% compared to 2010 with over 4,500 new attacks every day; additionally, 39% of all email-based attacks used a web link within the email as the malicious vector [26]. Given this concerning situation, methods for differentiating web browsing behavior across an organization that is benign versus that which increases the risk of computer compromise is of particular relevance and use.

The focus of the work presented in this paper is on improving the security of all computers as a comprehensive system within an organization. We assume that compromises, originating from the Internet through a variety of mechanisms, occur at a low but continuous rate for all significantly-sized organizations. In addition, we assume that the contemporary organization's goal for cyber security is to rapidly detect and then reduce the frequency and damage caused by these Internet-originating compromise events. Note that this assumption is different than the traditional philosophy of the cyber Maginot Line with maximization of the perimeter and assumption that any breach is failure. The work we present in this paper is about learning from ongoing compromise-oriented activity across a large, coherent population of computers to manage and minimize future malicious activity to the aggregate population. It is not about how to improve the protection of individual computers in isolation.

Our work demonstrates that risk of compromise is not

uniform across a set of computers (representing users). We find that, indeed, there are patterns of risk when computers exhibit certain web browsing behavior over time. We also find that a simple quantity of web activity (the number of unique web locations visited) may have an equivalent association and provide the same predictive capability as the more complicated adopter models we explore below.

In this paper we begin with an overview of existing related and relevant research. We then present a model for describing web access behavior across a population of computers. Next we define a variety of data sources used for validating this model including some useful characteristics in large-scale web browsing traffic and relevant compromise indicating data sets. We also propose data reduction and normalization processes to help allow better comparison and analysis. These data sets are then used to explore the model. We conclude the paper with a discussion of applications for our models, shortcoming of our approach, and future directions for using the data and outcomes of this study.

2 Related Work

A variety of work has been done to characterize web browsing behavior and activity. Many focus on various means of content and search classification. Kumar et al. provide a large-scale study of web browsing content classification using a one week data sample collected from browser toolbar searches [18]. A variety of others look at content classification, general browsing patterns and content valuation [1, 2, 19]. Another area of research has been on re-visitation of web content and understanding how often and perhaps why users are returning to the same or similar content [14, 5, 27]. All of this research focuses on single users and not necessarily information spread between users. Most of the research, to varying levels, found significant re-visitation of web locations and content by users. No research was found that explicitly examines a single large-scale organization over many months in terms of web browsing behavior quantification. We note that our behavior modeling is focused on the activity external to the computer and not traditional on-computer behavior modeling seen in existing research [11].

Additionally, a variety of research exists in web-based compromises, methods, and understanding. Provos et al. provide a well cited, comprehensive overview of recent web attack methods and the significant volumes of malware seen on the web [24]. Their results on over 4.5 million URLs showed 10% as malicious. A slightly older study by Mushchuck et al. show similar results over a smaller set of URLs; 13.4% of web downloaded executable content being malicious and 5.9% of dynamic (script) web content being malicious [22]. Another study by Provos et al. shows that 1.3% of Google searches returned at least one URL result that was malicious [23]. Note that this malicious content is not hosted by Google's site but is instead on the sites presented by Google in the research results.

The work by Moore et al. provides some interesting and relevant time frames for how long web sites serving malicious content exist before take-down events occur [21]. Their study shows that phishing-oriented web sites existed for an average of 58 hours before take-down but had long-lived sites as well, producing a lognormal distribution and a median of just 20 hours. They also imply the difficulties of just blocking and blacklisting web locations as a solution to defending against malicious web content given the variable nature and ease at which web locations are changed.

Ma et al. propose a machine learning approach to determining malicious web content solely through the URLs and a variety of associated, non-content attributes (WHOIS and similar data) [20]. Using a sample of approximately 30,000 known benign and malicious websites (URLs) from several sources, they showed a 95-99% accuracy in determining malicious content. Invernizzi et al. demonstrate an efficient approach for determining and finding malicious web content throughout the Internet using attributes from existing malicious web locations to help narrow the search [15].

Hein et al. provide an overview of contemporary attacks against web browsers and mitigation strategies [13]. Their paper describes several mechanisms of how exploits are injected into web browser clients without direct user knowledge and both infrastructure and browser improvement mitigation techniques. They end the paper by proposing a crowd-sourced trust model that allows web browsers to determine potential harm of content based on others' prior experience. A variety of novel approaches to the traditional detection of malware delivered through the web continue to be developed [4, 17]. Grier et al. show an interesting ability to determine the root (actual) source of malicious content that attempted to compromise a web browser as a function of their proposed browser [12]. Davis et. al demonstrate the use of time series data to present web access volumes before and after publicized cyber incidents to determine the effects of incidents on activity by the public to the web site [7].

Few validated methods for objectively determining risk with a computer based on activity or behavior exist. In contrast, existing research focuses on the behavior of computer attackers themselves and not on the recipient's perspective [6]. Unfortunately, quantifiable methods of validation are lacking in much of the related, existing research [28]. An objective, data-driven approach to de-

termining risky behavior is of particular importance to large-scale cyber defense [10]. We believe this approach is a key aspect of the model presented and discussed below.

3 Approach

In this section, we introduce our model for describing web access behavior across a population of computers. It assumes a source of historical web access logs representing the population of computers. Internet-accessing web proxy logs are a common source of such data for large organizations.

Our web adopter model (WAM) hypothesis is inspired by Rogers' sociological model of technology early adopters and the repetitive mechanisms through which new technologies expand gradually to a larger group of adopters [25]. This technology adoption model can be succinctly described as follows: When a new technology becomes available, a subset of the populace (risk-taking and technology-centric individuals) quickly adopt the technology. With increasing propagation, a larger set of individuals follow in adoption, but with an assumed lower risk (since any problems were worked out by the earlier adopters). This process and the role individuals play within the process is generally static from one related technology to the next in terms of propagation distribution, speed, and path.

When applied to Internet web browsing behavior, we find that indeed there are well-defined patterns when computers (representing users) adopt specific web locations over time. We find this behavior, in combination with the count of unique web locations accessed by the computer, does have statistical power for predicting the risk of compromise. In particular, when we look at specific subclasses of compromises, there is a strong association between web adopter behavior and probability of compromise.

WAM specifically distinguishes three classes of web adopter behavior that we show associate well to risky and non-risky behavior. The first adopter type we refer to as unique adopter (UA) behavior. The UA behavior applies to computer access events to Internet web locations that are unique accesses within the computer population and time frame considered. The second type we call early adopter (EA) behavior, which are those computers accessing web locations in a well-defined time period before other computers within the population also visit the web location. The final type we define as mainstream adopter (MA) behavior. MA behavior occurs when a computer accesses a web location that is common in the computer population and it would be impossible to distinguish EA behavior; for example, much of the computer population visits `http://google.com`. These behavior classifications are applied to each unique web access that a computer makes over a time period and drives overall labeling of the computer's browsing behavior.

We now describe WAM more formally.

Given a set of all unique web locations W (Section 4.1 defines web locations), over a time period T, each unique web location χ is associated with an unevenly spaced time series (USTS) [9] of access events. The USTS has N elements over time period T beginning at T_{start} and ending at T_{end}.This USTS is a sequence of value pairs of the first access time by a computer and the identifier (name) of the computer (t,C):[1]

$$W_\chi = (t_1,C_1),...,(t_N,C_N)$$
$$\text{s.t. } T_{start} \le t_1 \le t_2 \le ... \le t_N \le T_{end}.$$

Note that any subsequent accesses by an individual computer C to the same web location χ are not within the USTS W_χ; only the first access by that computer to the web location are included.

We define a single element set containing computer C as an UA set, $\{(C_1,...,C_n)\}$, to web location χ when C was the only computer to access the location in the time period T:

$$UA_\chi = \{C : W_\chi = \{(t,C)\}\}.$$

Similarly, we define a set of computers $\{C_1,...,C_n\}$ as an EA set to web location χ when set members are the first to access χ in the time period T; at least three computers accessed χ; and the EA set accessed χ at least 24 hours prior to at least one additional computer accessing χ (not in the EA set):

$$EA_\chi = \{C_i : (t_i,C_i) \in W_\chi,$$
$$|W_\chi| \ge 3,$$
$$\exists (t_k,C_k) \in W_\chi \text{ s.t. } t_i + 24 \text{ hours } \le t_k\}.$$

The 24 hour time separation between early adopters and non-early adopters (visitors) is based on observation of the data set. Figure 1 shows the distribution of time between the early adopter(s) and subsequent adopters. Note the well-defined time steps between the set of early adopter(s) and the subsequent computers that access the web location. Specifically, we find that the time separation generally falls on well-defined time boundaries of 1 day (24 hour) increments. Our speculation is that this reflects human behavior and the propagation of information about the web location from the early adopter population to the subsequent adopters. It provides a useful boundary between early adopters and the rest of the adopting population of a web location. In addition, it

[1] Most cyber security relevant data sets are at one second resolution and events within the same second are randomly ordered within that second, otherwise the sequence is strictly ordered in time.

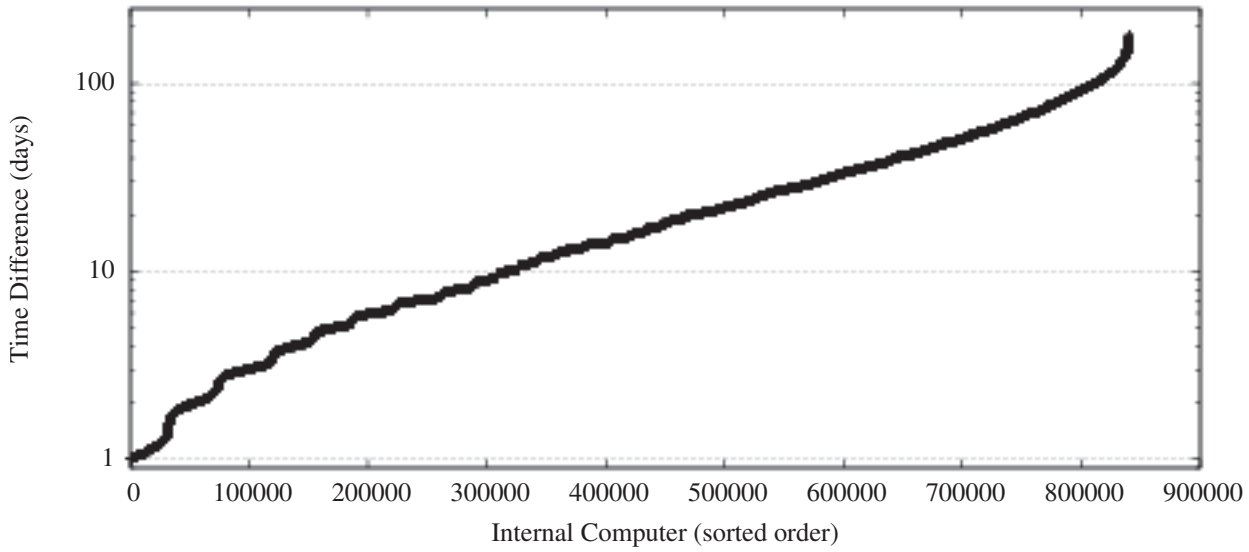

Figure 1: The empirical distribution of the time difference in days between the set of EA computers EA_χ and the additional adopters (computers who accessed it) of all web locations χ in our web data set. Note the step at well-defined boundaries of day intervals. Note the graph is log scale on the y-axis.

was important to allow for multiple early adopters, as can obviously occur. A good example of multiple early adopters would occur with a phishing email containing a web link that multiple receiving users then clicked soon after receiving it.

The top 1 percentile of web locations by number of unique accesses represents over 90% of all successful web access events, as detailed in Section 4.1. While the top percentile was chosen subjectively, it provides a useful differentiator for particularly common locations like `google.com` versus the rest of the web locations. We define ℓ, the set of lengths for all W_χ (number of unique accesses for all χ):

$$\mathscr{L} = \{|W_\chi| \forall \chi\}.$$

Let $p_{99}(\mathscr{L})$ be the 99th percentile of $\mathscr{L}$. We now define Λ as the set of web locations χ that are in the 99th percentile or above in terms of unique access length:

$$\Lambda = \{\chi : |W_\chi| \geq p_{99}(\mathscr{L})\}.$$

Finally, we define a set of computer $\{C_1, ..., C_n\}$ as a MA set to web location χ when the web location is in the set of most accessed web locations Λ:

$$MA_\chi = \{C : (t, C) \in W_\chi, \chi \in \Lambda\}$$

When applied to Internet web browsing behavior, we find that indeed there are well-defined patterns of web location adopter behavior for specific unique requests over time. In addition, we find web adopter behavior does have an association to riskier behavior relating to compromise in our data sets. In particular, when we look at specific subclasses of indicators of compromise (IOC),[2] there is a strong association between adopter behavior and probability of compromise. The results of the model in conjunction with IOC events are discussed in Section 5.

4 Data

Several sources of data were collected and analyzed for the purposes of validating and analyzing WAM. The data sets collected are from Los Alamos National Laboratory's (LANL's) organizational user networks over a period of 6 months during 2011 and uses activity data collected from 5 primary sources:

- *Web locations:* 6.4 billion outbound web proxy log entries representing Internet web requests from 24,292 computers.
- *Antivirus:* 306,135 antivirus log entries from approximately 18,000 Microsoft Windows-based computers.[3]

[2] An IOC is often also referred to as an intrusion detection or compromise *signature* in much of the related research.

[3] The count is approximate due to computers only reporting *events* and an exact inventory of computers with the antivirus reporting agent was not available. The additional 6242 or so computers seen in the web proxy logs are non-Windows computers or Windows computers with custom configurations that do not report antivirus data.

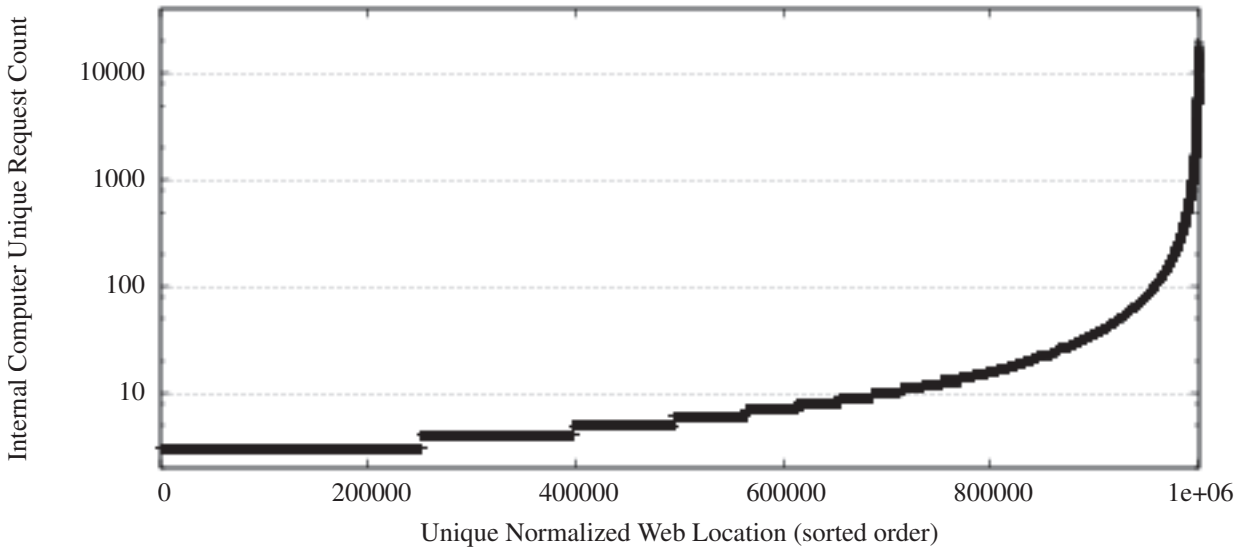

Figure 2: The empirical distribution by number of unique client computer accesses to unique normalized web locations over the 6 month time period. Only data for the 1,000,664 web locations that had three or more unique client visits is shown. The top 1% of web locations are still included in this representation. Note the graph is log scale on the y-axis.

- *IR:* 1256 Internet-intrusion incident response analyst tickets from LANL's incident response capability.
- *Phishing* 44,550 normalized web locations known for serving phishing-based malware derived from public Internet sources.
- *Proximity:* 19 normalized web locations derived from time proximity across two or more computers with existing and time-relevant IOC events.

More specific dates of the data analyzed are not disclosed to reduce the likelihood that an adversary could use the information presented in this paper for inappropriate purposes.

4.1 Web Location Access Events

For the first data set, the outbound HTTP proxy logs were analyzed. Using only successful GET and POST requests (and excluding CONNECT and others), this accounts for 6.4 billion individual web page request records over the 6 months from the 24,292 LANL computers. However, many of these requests are unnecessarily specific and redundant so we therefore normalize them.

A normalized HTTP web request has a somewhat complex definition as used in this paper. Most simply stated, it is a client computer's first successful request for a given file extension or file type from the base domain of the source server. This simplification allows us to condense the web request events that are overly unique due to load balancing servers or dynamically generated client-specific paths and file names; and to reduce the variety of file types possible (e.g. `x-pdf` and `pdf` resolve to the same type), but still retain some distinction of different file types coming from a web domain.

More specifically, the normalized web request uniform request locator (URL) is substantially shortened to include just the base 2-tuple of the server's domain (or 3-tuple in the case of two letter country code suffix or first two octets of the IP address if no domain). The URL's path is then replaced with just the file extension or MIME type if there is no file extension. Given these changes, the normalized web request becomes: `http://aaa.com/pdf` or `http://aaa.com.au/html` (we call these "web locations" and define them individually as χ). When reduced, this accounts for 3,942,541 normalized unique web locations (site and type pairs).

Borders et al. provide an intriguing, though more complex, method of dealing with the dynamic URL's generated by dynamic web content to enable site comparison and analysis [3]. Unfortunately, their method also requires elements of content beyond just the URL. Their intended use was also quite different than ours.

Even with this normalization, the breadth of uniqueness of the web sites is noteworthy: 2,389,586 (60.6%) of the normalized sites during 6 months are unique (only one client computer accesses it); 552,291 have two different client's requesting them; and 1,000,664 with three or more. We believe this high uniqueness is primarily the result of two factors: load balancing of server con-

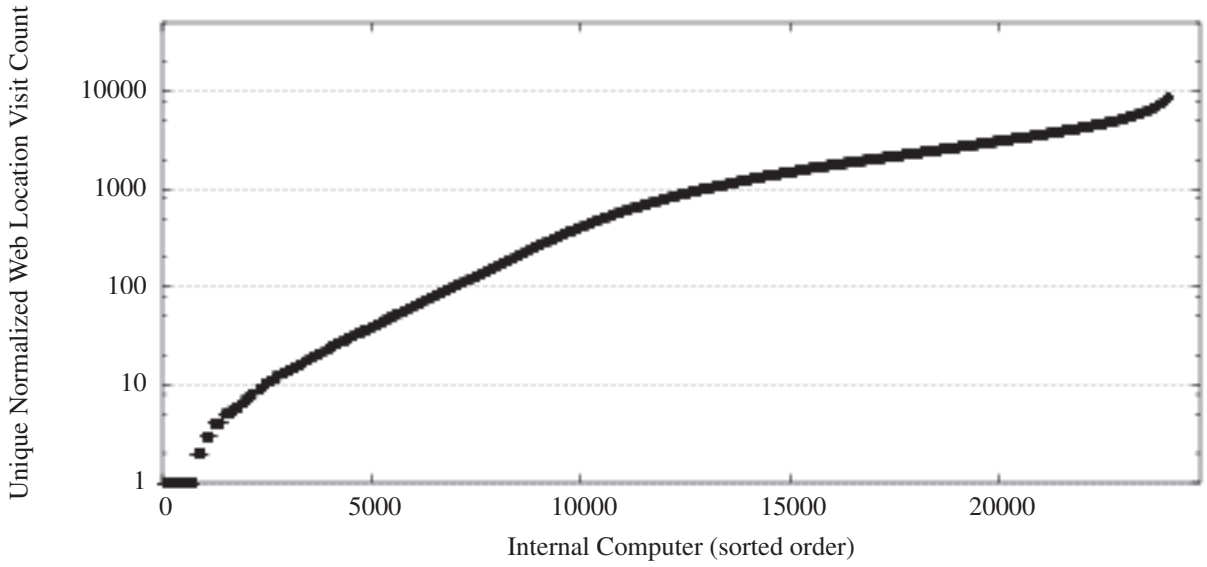

Figure 3: The empirical distribution of the 24,292 internal computers and the number of unique normalized web locations each has accessed over the 6 month time period. Note the graph is log scale on the y-axis.

tent and dynamic locations generated by dynamic content. Our normalization process attempts to combine location-related content, while still allowing some distinction through file type differences. Figure 2 shows the empirical distribution of clients visiting normalized web locations with 3 or more clients.

On the other end of the distribution, a vast majority of requests are made to a relative minority of unique locations on the Internet. For example, requesting the web location `http://google.com/html` occurs from 19,800 of the 24,292 total computers during the 6 months.[4] We find that the top 1 percent of web locations (39,424 web locations) in the month's traffic have 108 or more requests from distinct client computers. We make the assumption that this top 1% represents the most popular information on the Internet for LANL computers. In fact, we find that this set of web locations represents over 90% of the total web access traffic in the 6 months. While compromise is not impossible from these top locations, it is improbable and when it does occur, it is quickly and publicly mitigated. Thus for the purposes of this work, we exclude the data from these top 39,424 normalized web locations. Normalization and removal of these web locations reduces our data set to 3,903,117 unique web locations.

Figure 3 shows the empirical distribution of computers to the number of unique normalized web locations each accessed over the 6 months. Observe the extremely high number of unique web requests that exist for some web locations and the value in removing these few extremes for comparison purposes across the rest of the web locations. The average number of web locations that a computer accessed in the 6 months was 1565 with a standard deviation of 2043. The minimum was 1 web location and the maximum was 36,896. The median number of web locations accessed was 91.

4.2 Compromise Events

As previously stated, compromise data over the 6 month time period comes from two sources: the individual antivirus logging of approximately 18,000 Microsoft Windows-based computers and the incident response (IR) tickets for intrusions from LANL's incident response capability. During the 6 month time period, 848 computers, in 306,135 individual (and often repetitive) events, reported having the local antivirus engine detect malware. From analysts, 1256 unique IOC ticket events involving 401 computers were recorded during the time period.

In terms of compromise and intrusion detection, most methods of detection are an *existence* IOC that does not contain information regarding source or method of compromise. For example: an antivirus engine can detect that a piece of malware is on a computer as part of a nightly file check, but would not indicate how the malware came to be on the computer. Likewise, incident response tickets will indicate that a computer has been found to contain malware (perhaps it was unexpectedly

[4] In terms of all requests (not just the first one by a given computer), `http://google.com/html` represents 7,124,661 or 1.5% of all web traffic over the 6 months.

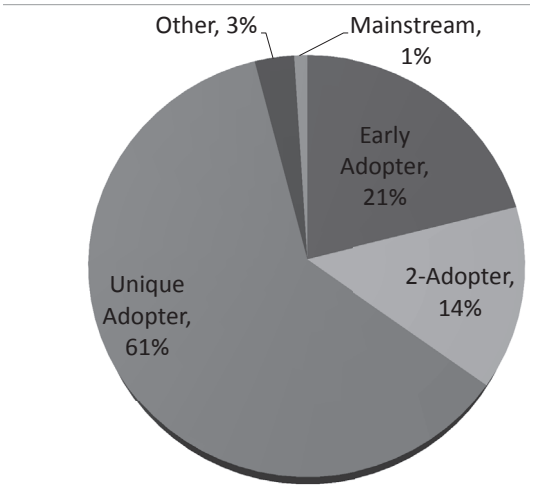

Figure 4: Chart comparing the unique normalized web locations in the 6 months (T) by type. Unique adopters are those web locations with exactly one client computer visiting during the 6 months where $|W_\chi| = 1$. 2-adopters are those locations with exactly two unique client computers, $|W_\chi| = 2$. Early adopters are web locations χ are those locations where $|EA_\chi| \geq 1$. Mainstream are web locations χ that account for the top 1% of unique visits by count where $|MA_\chi| \geq 1$. Finally, those sites that have 3 or more visitors but do not meet the conditions as an EA or MA containing location are labeled as Other.

beaconing to the Internet), but again does not indicate how the malware came to be resident on the computer. IOCs also contain an inherent level of error and more accurately they are indicators of *potential* compromise with often high and unquantified levels of false indication.

False positives likely do exist at some level within the two IOC data sets due to the inaccuracies of antivirus engines and IR investigations that did not associate to an actual intrusion. In addition, an unquantifiable number of malware events likely also exist within the 6 months that are not represented within these two data sets (false negatives). Even with these two limitations, we feel that this compromise data is particularly valuable and unique as a data source given its association with actual cyber intrusion events, as shown in the following results.

4.3 Internet-Published Phishing Websites

Publicly available malware-serving website lists for phishing attacks were used to generate another data set of potential IOC events for relational purposes. 12,863,857 and 1,447,310 unique, malware-serving URLs were retrieved from `malwaredomainlist.com` and `phishtank.com`, respectively on September 5, 2012. These URLs were then reduced to 44,550 unique normalized web locations using the same method described above for web proxy data. None of these phishing-based malware web locations matched the top 1% of web sites previously described (mainstream). The substantial reduction in URLs points to the significant reuse of base sites for serving malware. Obviously, such reduction does introduce the risk of false positive matching in this data set, though for the purposes of our study we believe it does not significantly impact the results. Of the 44,550 web locations, 4411 were seen within the 6 months of LANL's web request data. Of course, accessing a web location known to contain phishing-related malware does not mean guaranteed compromise but it is definitely risky behavior and we consider it to be a *potential* IOC.

4.4 Time Proximity to Existing IOC Events

The final IOC-related data set uses the web location USTS sets to determine IOC events of interest. This approach combines the time line of successful website accesses (or downloads) for each client computer with the time line of IOC events that also occurred for the computer. It then considers suspect any website accesses within a 24 hour time period before or after an IOC event. If two or more computers consider a website access suspect due to time proximity with an IOC, the website access is considered a *potential IOC*. In our data set, when the website access is seen by 3 or more computers in association with an IOC, we find no false positives–the website location has been found to always be a source of malicious activity. Additional details and the results of this IOC detection method can be found in [16].

5 Results

Using WAM, as described in Section 3 and the data sources described in Section 4, we now discuss results.

To provide a sense of how the adopter types are distributed, Figure 4 shows the ratios of each adopter type by web locations in the normalized web access data set over the 6 months. As expected, the largest volume of web locations are associated with UA behavior followed by web locations that have distinct EA computers. Note that the 2-adopters in the figure are those web locations χ where $|W_\chi| = 2$ and can have neither UA or EA associated computers, per the definitions. Similarly, "Other" applies to those web locations that do not have EA or MA sets, e.g. there are multiple computers accessing it but none initially more then 24 hours apart or enough to make it mainstream.

To apply WAM, we take the set of web locations χ that each computer C accessed from the normalized web

access data set (W) over the 6 months, defined as:

$$W_C = \{\chi : \chi \in W, C \in W_\chi\}.$$

For each web location χ in W_C, we apply an label based on computer C's membership in one of the various sets UA_χ, EA_χ, or MA_χ. For each computer C, all web locations χ in W accessed during T (the 6 months for our data set) are potentially labeled as UA, EA, or MA respectively.[5] The ratios of each of these labels to the total web location accesses ($|W_C|$) for each computer are then computed. Figure 5 shows the ratio for each of these labels across the population of computers within the data set.

Using the four different types of IOC data described in Section 4, we then label a given computer as being *compromised*[6] by one or more of those IOC types if the computer is associated with the IOC type during the 6 months. The four IOC labels, as previously discussed, are *Antivirus (AV)*, *IR*, *Phishing*, and *Proximity*. While this is a very coarse labeling of compromise over a potentially long period of time, we find that the model yields interesting results and suspect that more fine grained labeling would improve the model's fidelity and usefulness (with significantly higher data and computation costs).

When we consider the set of computers exhibiting any of these four types of IOC events, we find that the ratios of adopters are rare at the low and high endpoints, as seen in Figure 6. The lack of low ratios is easily explained by the notion that computers that do not access unique web locations or go to locations as the first set of visitors (when exploits may more likely exist) are much less likely to be compromised. While the result is more difficult explain, we see three explanations for why high adopter ratios would exist without associated IOC events:

- High MA ratio computers are very likely to have few to no compromise events since they only access very well known and often accessed locations on the Internet; sites we have previously asserted are not associated with compromise events.
- For a small set of computers with high UA and EA ratios, we find they are crawling many, diverse web locations in an intended automated fashion.
- Somewhat more speculatively, it may be that computers with very high levels of UA and EA activity are representing users with a higher level of knowledge in avoiding compromise. This idea requires further exploration to assert or reject.

[5] Again, note that some locations χ are not labeled since they do not meet the definitions for the 3 adopter types.

[6] We assume the logged IOC was correct and not a false positive for our statistical purposes.

When a computer's ratio of traffic for UA and/or EA exceeds 1%, we label that computer as being of generally type UA and/or EA. We choose the somewhat arbitrary 1% cutoff based on the significant increase that occurs approximately at this value, as seen in Figure 5. Similarly, we used a 99% cutoff for a computer to be labeled as a mainstream adopter; 99% or more of its accesses to unique web locations χ must be labeled as mainstream accesses. Using these computer-based labels and the association of computers to the various IOC events described in Section 4, Figure 7 shows the overall results of applying WAM using our data sets. The strong association to potential IOCs relating to computers visiting phishing labeled web locations is particularly noticeable. We assert that phishing and proximity data are most strongly associated adopter behavior due to their pure web-based association. In contrast, antivirus and IR events can occur through other compromise mechanisms that are not associated with web browsing activity. In addition, antivirus data is only collected from a subset of Windows-based computers compared with the larger set of web browsing computers (18,000 versus 24,000). Nonetheless, there is still significant association between all of the IOC types and computers labeled as UA and EA.

Using these results, we can now estimate probabilities of an IOC of various types occurring, given a computer C has one or more of these labels, We define the estimated probability for an adopter label L as:

$$\hat{P}(L) = \frac{|L(C)_\gamma|}{|L(C)|},$$

where $L(C)_\gamma$ is the set of computers $\{C\}$ with adopter label L associated with a potential IOC γ during time period T and $L(C)$ is the set of computers with adopter label L. These probabilities for both having and not having the associated labels can be seen in Figure 8. Note the particularly high probability of having a phishing IOC associated with a computer that is both an unique and early adopter. Also note the extremely low probability for computers label as mainstream adopters for any type of IOC–these computers are obviously not the source of most compromise from the Internet. The $UA \vee EA$ results are not shown due to insignificant differences in the values and brevity.

One particularly important question about how useful these adopter types are for predictive methods is whether these labels are leading or trailing indicators for the different compromise events. We used a coarse method of dividing the data sets in to two 3 month pieces and assigned adopter type labels and IOC labels to each computer independently within the 3 month data sets. Thus, if a computer had an adopter label in the first 3 months that drove to an IOC in the second 3 months, we assume

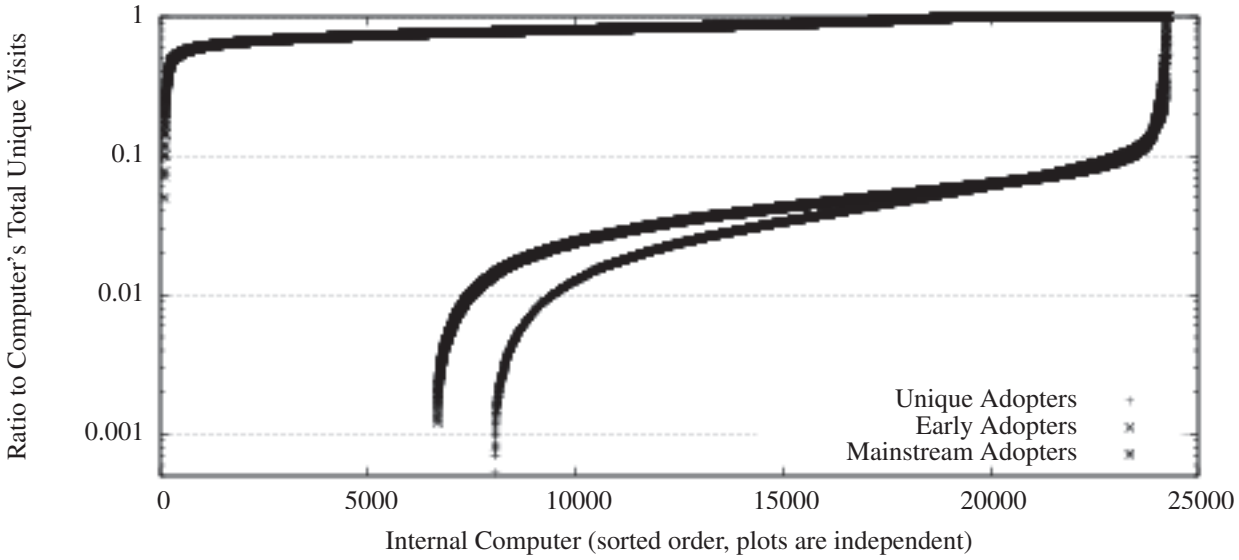

Figure 5: The empirical distribution of the 24,292 internal computers in terms of the ratio of their individual total unique web location visits over the 6 months as UA, EA, and MA. Note the graph is log scale on the y-axis.

the adopter label is a *leading indicator*. Inversely, if an IOC occurred in the first 3 months that led to an adopter label in the second 3 months, we assuming the adopter label was a *trailing indicator*. If the adopter label and IOC both occurred in the same three month period, we call that a (time) *local indicator*. These indicators are not mutually exclusive when applied to a computer's activity over the 6 months. For example, a computer could be a leading, trailing, and local indicator if the same adopter label and the same type of IOC event occurred in both 3 month time periods.

Figure 9 shows the volume of computers showing a leading, trailing, and/or local indicator for each adopter label and data set. As is shown, the adopter labels show association for the antivirus and phishing data sets. However, the labels for the IR data set are seen more strongly as a trailing indicator. The high volume of UA and EA labeled computers that show association within time local IOC events shows the strong association to UA and EA behavior and similar-in-time risky behavior. In contrast, the lower volume of time local IOC events to MA behavior reconfirms the relatively low risk that MA labeled computers show. This low volume is particularly apparent for phishing IOC's since, as previously discussed, the phishing IOC data set does not contain any mainstream web locations. The combined $UA \wedge EA$ and $UA \vee EA$ sets were not shown due to insignificant differences in the results and for brevity of presentation.

While this initial analysis on leading and trailing indicators lacks truly useful granularity, we believe our results provide some generalized differentiation.

5.1 Application

These differences in distinct risk groups that are represented in the data can be used to help prioritize and localize the placement of traditional intrusion detection sensors and the sensitivity of these sensors. Placing sensors at or near MA labeled computers has little value but placing more sensors near computers that are both labeled as unique and early adopters has much higher value than otherwise random placement. More specifically, non-MA labeled computers have a 10-fold increase in being associated with any of the IOCs and for the phishing IOC its a 417-fold increase.

5.2 Regression Model for Prediction

We also tested the significance of the overall model for statistical fit and predictive viability using logistic regression. We tested WAM for each computer within the data set at predicting IOC events against both an intercept (the null hypothesis that WAM has no effect on compromise) and using the total unique web location counts. For the adopter parameters we used the ratio of each adopter type for each computer to the total unique web location visits for the computer (we did not just use the simplified adopter labels discussed previously). The adopter parameters UA, EA, and MA fit the regression model while other parameters were redundant. Using these three parameters we find that WAM does much better then the intercept. However, when we use just total unique web location counters, we find equivalent predictive power indicating that this simpler counter has significant and equiv-

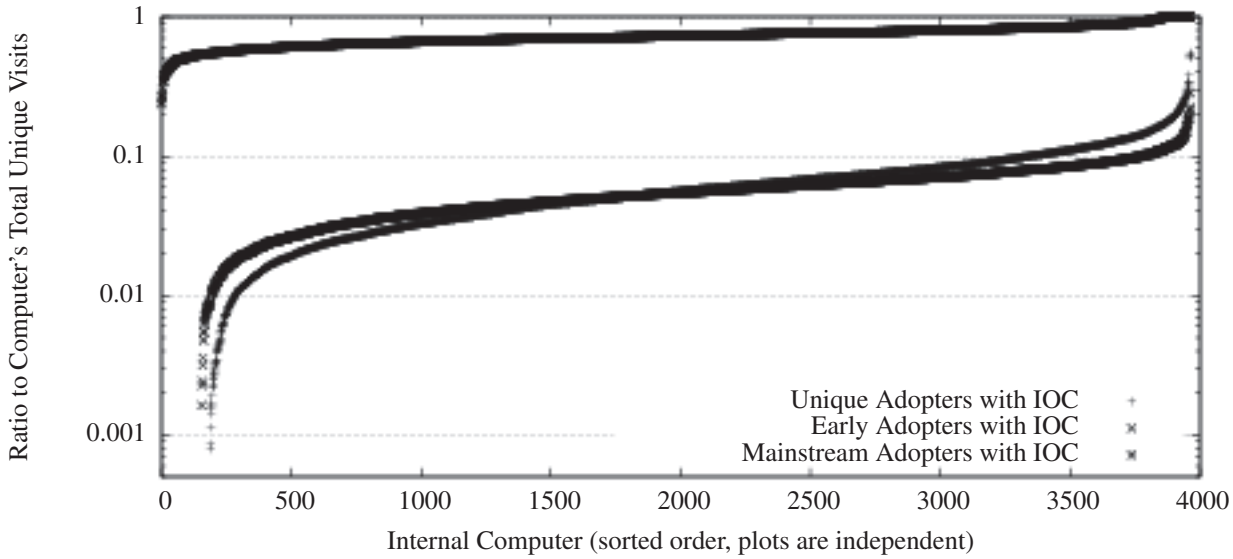

Figure 6: The empirical distribution of the 3964 internal computers that had any IOC during the 6 months in terms of the ratio of their individual total unique web location visits as UA, EA, and MA. In comparison to the total population in Figure 5, note the reduction in computers with no UA or EA ratios and the reduction computers with extremely high UA, EA, and MA ratios. Note the graph is log scale on the y-axis.

Type	**Population**	*UA*	*EA*	$UA \vee EA$	$UA \wedge EA$	*MA*
Pop.	24,292	14,825 (61.0%)	16,519 (68.0%)	17,453 (71.9%)	13,891 (57.2%)	5213 (21.5%)
AV	848 (3.49%)	685 (80.8%)	726 (85.6%)	747 (88.1%)	664 (78.3%)	63 (7.43%)
IR	401 (1.65%)	323 (80.6%)	337 (84.0%)	346 (86.3%)	314 (78.3%)	40 (9.98%)
Phish	3032 (12.5%)	2975 (98.1%)	2988 (98.6%)	3015 (99.4%)	2948 (97.2%)	2 (0.0662%)
Prox.	19 (0.0782%)	19 (100%)	19 (100%)	19 (100%)	19 (100%)	0 (0.0%)
Any	3964 (16.3%)	3681 (92.9%)	3746 (94.5%)	3801 (95.9%)	3626 (91.5%)	104 (2.62%)

Figure 7: Summary of each of the analyzed data sets and set memberships. For each of the IOC types, the number computers (and percentage of total IOC-tagged computers) is show as being labeled as UA, EA, either, both, or MA. To be labeled as UA or EA requires 1% or more of the computer's total unique web access traffic to be as an UA or EA (respectively) to web locations. To be labeled as MA, the computer needs 99% or more of its traffic to be to mainstream defined web locations. Population (Pop.) and Proximity (Prox.) are abbreviated for formatting purposes.

alent statistical predictive capability. In fact, we see that the adopter parameters, in combination, *represent* unique web location visits and individually delineated do not increase predictive capability. In other words, a computer's UA, EA, and MA parameters together present that computer's web behavior that when reduced can be equivalently stated as the unique location visit count; at least in terms of statistical prediction.

A comparison of predictive power between the two models is shown in the ROC curve in Figure 10. As seen in the figure, both the WAM model and the quantity of unique web locations visited provide better then chance but are still far from perfect predictors. The lines are nearly identical, indicating that they likely represent the same predictive capability. Given the fact that counting unique web location visits in a time period is much easier than calculation of the WAM parameters, the obvious conclusion is that it is simpler and more appropriate to use the unique visit count as the best predictor of future compromise behavior by a computer.

6 Lessons Learned

The most significant observation and theme from the research represented in this paper is the importance of simplicity in approach, whenever possible. While it is motivating to gravitate towards more complicated approaches as *good science*, we believe that given the relative immaturity of the cyber research domain, there is significant value and importance in the simplest approaches; at least

Type	$\hat{P}(UA)$	$\hat{P}(\neg UA)$	$\hat{P}(EA)$	$\hat{P}(\neg EA)$	$\hat{P}(UA \wedge EA)$	$\hat{P}(\neg(UA \wedge EA))$	$\hat{P}(MA)$	$\hat{P}(\neg MA)$
Antivirus	4.62%	1.72%	4.40%	1.57%	4.78%	1.77%	1.21%	4.11%
IR	2.18%	0.825%	2.04%	0.824%	2.26%	0.837%	0.767%	1.89%
Phishing	20.1%	0.606%	18.1%	0.570%	21.2%	0.810%	0.038%	15.9%
Proximity	0.128%	0.0%	0.115%	0.0%	0.137%	0.0%	0.0%	0.100%
Any	24.8%	3.00%	22.7%	2.81%	26.1%	3.25%	2.00%	20.2%

Figure 8: Estimated probabilities of various compromise types existing given a computer's membership in the various defined sets (or not). Using these estimated probabilities or frequencies as a basis for predicting future potential events occurring in the various populations (sets) assumes that set membership is a viable leading indicator of risky behavior.

Adp.	IOC	Pop.	Leading	Trailing	Local
UA	AV	681	77.8%	52.6%	96.9%
UA	IR	312	30.1%	69.2%	93.3%
UA	Phish	2809	52.1%	58.4%	99.4%
EA	AV	694	78.4%	55.0%	98.3%
EA	IR	323	31.0%	70.0%	94.7%
EA	Phish	2818	52.8%	59.2%	99.5%
MA	AV	88	59.1%	36.4%	67.1%
MA	IR	44	36.4%	43.2%	77.3%
MA	Phish	47	66.0%	34.0%	2.1%

Figure 9: Trending for each adopter label (Adp.) as to how often it has a leading, trailing or time local (similar in time) association to computers observed with each IOC type. The associations are not mutually exclusive. The population (Pop.) size represents the number of computers that have that adopter label in either time period and at least one associated IOC event of the given type over the two 3 month consecutive time periods.

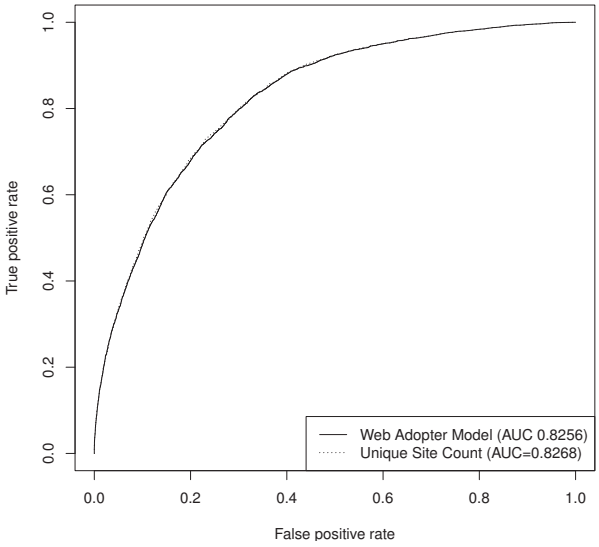

Figure 10: ROC curve showing both the fit (and likely predictive capability) of both early adopter logistic regression and simple unique web location logistic regression models. The two curves are nearly identical.

until these direct methods are *demonstrated* to be insufficient.

We assert that the regression-based prediction results we show are an excellent example of Occam's razor, where the simpler hypothesis with fewer assumptions produces equivalent results to the more complex approach [8]. We have shown that simple web counts provide equivalence to a much more complex and frankly better sounding approach.

Another key observation that was initially not expected in this research was the significant uniqueness seen within web URLs across the large population of hosts. This variation drove us to the web location approach described in Section 4.1 that allowed better URL grouping, while still enabling some diversity beyond just domain name. Without this approach, the data analysis was overwhelmed with one-event URL's that did not allow for multiple adopters except primarily in the Mainstream group. While we did try more complex methods to detect and normalize similar URLs (e.g., timestamp detection and random *ID* string detection within the URL string), we found that our simple approach provided an equivalent capability with significantly less processing and complexity. Our final approach is also much easier to describe and replicate.

Also worth noting, this paper shows the importance of using real world data in significant volume for effective hypothesis testing within the cyber domain. Using contrived data, WAM could have easily shown predictive capability stronger then the real data since such testing data is often modeled from the hypothesis itself. Obviously, the real data demonstrated a more realistic outcome.

7 Conclusion

The intended purpose of WAM was to define an objective model that delineates varying behavior according to the potential risk of compromise within a population of

computers. Though the simpler web location measure significantly reduces the overall relevance of WAM, we believe that the comparison, results, and the WAM model itself are still useful and important applied cyber security research.

While there is continued opportunity for improvement in terms of fidelity and fit, we believe the model we have presented in this paper presents a newly explored approach and at least an objective means to judge risky behavior. Even more important and rare, this paper uses a significant real-world data set to validate and quantify the model. WAM demonstrates a model for showing the association of compromise through differentiated web browsing behavior over a population of computers. While we also showed the WAM model does not provide additional predictive capability over unique web location counts, we have demonstrated the value of this simple count. Indeed, more web surfing does equate to a higher chance of compromise.

References

[1] BAYKAN, E., HENZINGER, M., MARIAN, L., AND WEBER, I. Purely URL-based topic classification. In *Proceedings of the 18th International Conference on World Wide Web* (New York, NY, USA, 2009), WWW '09, ACM, pp. 1109–1110.

[2] BILENKO, M., WHITE, R. W., RICHARDSON, M., AND MURRAY, G. C. Talking the talk vs. walking the walk: salience of information needs in querying vs. browsing. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2008), ACM, pp. 705–706.

[3] BORDERS, K., AND PRAKASH, A. Quantifying information leaks in outbound web traffic. In *Security and Privacy, 2009 IEEE Symposium on* (2009), IEEE, pp. 129–140.

[4] CANALI, D., COVA, M., VIGNA, G., AND KRUEGEL, C. Prophiler: a fast filter for the large-scale detection of malicious web pages. In *Proceedings of the 20th International Conference on World Wide Web* (New York, NY, USA, 2011), ACM, pp. 197–206.

[5] COCKBURN, A., AND MCKENZIE, B. What do web users do? an empirical analysis of web use. *International Journal of Human-Computer Studies 54*, 6 (2001), 903 – 922.

[6] DANTU, R., KOLAN, P., AND CANGUSSU, J. Network risk management using attacker profiling. *Security and Communication Networks 2*, 1 (2009), 83–96.

[7] DAVIS, G., GARCIA, A., AND ZHANG, W. Empirical analysis of the effects of cyber security incidents. *Risk Analysis 29*, 9 (2009), 1304–1316.

[8] DOMINGOS, P. The role of Occam's razor in knowledge discovery. *Data mining and knowledge discovery 3*, 4 (1999), 409–425.

[9] ECKNER, A. A framework for the analysis of unevenly spaced time series data. `http://www.eckner.com/papers/unevenly_spaced_time_series_analysis.pdf`, August 2012.

[10] GEER, D., J., HOO, K., AND JAQUITH, A. Information security: Why the future belongs to the quants. *Security and Privacy, IEEE 1*, 4 (2003), 24–32.

[11] GOPALAKRISHNA, R., SPAFFORD, E., AND VITEK, J. Efficient intrusion detection using automaton inlining. In *Security and Privacy, 2005 IEEE Symposium on* (2005), IEEE, pp. 18–31.

[12] GRIER, C., TANG, S., AND KING, S. Secure web browsing with the OP web browser. In *Security and Privacy, 2008 IEEE Symposium on* (2008), IEEE, pp. 402–416.

[13] HEIN, D., MOROZOV, S., AND SAIEDIAN, H. A survey of client-side web threats and counter-threat measures. *Security and Communication Networks 5*, 5 (2012), 535–544.

[14] HERDER, E. Characterizations of user web revisit behavior. In *Workshop on Adaptivity and User Modeling in Interactive Systems ABIS05* (2005), pp. 32–37.

[15] INVERNIZZI, L., COMPARETTI, P., BENVENUTI, S., AND KRUEGEL, C. EVILSEED: a guided approach to finding malicious web pages. In *Security and Privacy, 2012 IEEE Symposium on* (2012), pp. 428–442.

[16] KENT, A. D., AND LIEBROCK, L. M. Statistical detection of malicious web sites through time proximity to existing detection events. In *Resilience Week Symposium* (2013), IEEE.

[17] KIRDA, E., KRUEGEL, C., BANKS, G., VIGNA, G., AND KEMMERER, R. Behavior-based spyware detection. In *Proceedings of the 15th Conference on USENIX Security Symposium - Volume 15* (Berkeley, CA, USA, 2006), USENIX Association.

[18] KUMAR, R., AND TOMKINS, A. A characterization of online browsing behavior. In *Proceedings of the 19th International Conference on World Wide Web* (2010), ACM, pp. 561–570.

[19] LIU, C., WHITE, R. W., AND DUMAIS, S. Understanding web browsing behaviors through weibull analysis of dwell time. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2010), ACM, pp. 379–386.

[20] MA, J., SAUL, L. K., SAVAGE, S., AND VOELKER, G. M. Beyond blacklists: learning to detect malicious web sites from suspicious URLs. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2009), ACM, pp. 1245–1254.

[21] MOORE, T., AND CLAYTON, R. Examining the impact of website takedown on phishing. In *Proceedings of the anti-phishing workinggroup 2nd annual eCrime researchers summit* (New York, NY, USA, 2007), ACM, pp. 1–13.

[22] MOSHCHUK, A., BRAGIN, T., GRIBBLE, S., AND LEVY, H. A crawler-based study of spyware on the web. In *Proceedings of the 2006 Network and Distributed System Security Symposium* (2006), pp. 17–33.

[23] PROVOS, N., MAVROMMATIS, P., RAJAB, M. A., AND MONROSE, F. All your iFRAMEs point to Us. In *Proceedings of the 17th Conference on Security Symposium* (2008), USENIX Association, pp. 1–15.

[24] PROVOS, N., MCNAMEE, D., MAVROMMATIS, P., WANG, K., AND MODADUGU, N. The ghost in the browser analysis of web-based malware. In *Proceedings of the First Conference on First Workshop on Hot Topics in Understanding Botnets* (2007), HotBots'07, USENIX Association, p. 4.

[25] ROGERS, E. *Diffusion of Innovations, Fifth Edition*. Free Press, 2003.

[26] SYMANTEC. Internet security threat report, April 2012.

[27] TEEVAN, J., ADAR, E., JONES, R., AND POTTS, M. A. S. Information re-retrieval: repeat queries in yahoo's logs. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2007), ACM, pp. 151–158.

[28] VERENDEL, V. Quantified security is a weak hypothesis: a critical survey of results and assumptions. In *Proceedings of the 2009 workshop on New security paradigms workshop* (New York, NY, USA, 2009), NSPW '09, ACM, pp. 37–50.