On the Effectiveness of Risk Prediction Based on Users Browsing Behavior

Davide Canali
EURECOM, France
canali@eurecom.fr

Leyla Bilge
Symantec Research Labs, France
leylya_yumer@symantec.com

Davide Balzarotti
EURECOM, France
balzarotti@eurecom.fr

ABSTRACT

Users are typically the final target of web attacks: criminals are interested in stealing their money, their personal information, or in infecting their machines with malicious code. However, while many aspects of web attacks have been carefully studied by researchers and security companies, the reasons that make certain users more "at risk" than others are still unknown. Why do certain users never encounter malicious pages while others seem to end up on them on a daily basis?

To answer this question, in this paper we present a comprehensive study on the effectiveness of risk prediction based only on the web browsing behavior of users. Our analysis is based on a telemetry dataset collected by a major AntiVirus vendor, comprising millions of URLs visited by more than 100,000 users during a period of three months. For each user, we extract detailed usage statistics, and distill this information in 74 unique features that model different aspects of the user's behavior.

After the features are extracted, we perform a correlation analysis to see if any of them is correlated with the probability of visiting malicious web pages. Afterwards, we leverage machine learning techniques to provide a prediction model that can be used to estimate the risk class of a given user. The results of our experiments show that it is possible to predict with a reasonable accuracy (up to 87%) the users that are more likely to be the victims of web attacks, only by analyzing their browsing history.

Categories and Subject Descriptors

K.6.m [**Management of Computing and Information Systems**]: Miscellaneous—*Insurance, Security*

Keywords

risk prediction, profiling, web browsing

ASIA CCS'14, June 4–6, 2014, Kyoto, Japan.

http://dx.doi.org/10.1145/2590296.2590347.

1. INTRODUCTION

The World Wide Web is one of the main vectors used by cyber criminals to reach their victims. Malicious or compromised web pages are routinely used to infect user machines, spread malware, steal user credentials, and perform other phishing and scamming operations.

A large amount of research has been conducted on the tools and techniques adopted by attackers, to automatically identify and mitigate software vulnerabilities, or to protect web browsers from exploitation. Despite this effort, the percentage of web pages that are either malicious or that have been compromised to serve malicious content is steadily increasing [11, 29, 34]. Even though this is certainly an alarming phenomenon, these global figures are computed on the entire Internet population, and therefore fail to express what is the real risk for a single user to encounter a malicious page on her daily activity. The increasing number of dangerous sites does not necessarily affect everyone in the same way. For instance, it is possible that the majority of users only navigate in "safe neighborhoods" where malicious pages are still extremely rare. In this case, it should be possible to associate to each user, based on her usual online behavior, a certain *risk* profile. In other words, there should be a correlation between the *browsing habits* of a user and the probability she has to visit potentially harmful pages. This scenario is particularly attractive in the area of cyber-insurance [6], in which user profiling is an important step toward an accurate risk evaluation. In the physical world, insurance rates are normally computed based on a risk classification. For example, car insurances are more expensive in large cities or for inexperienced drivers – because this conditions are known to be positively correlated to the probability of car accidents. Unfortunately, an equivalent measurement of risk factors in the virtual world is still missing.

While the hypothesis of a correlation between the risk and the browsing behavior is reasonable, when dealing with the analysis of user browsing behaviors, there are also other factors that one has to take into account. For instance, independently from their daily activity, users are often socially engineered to click on links sent to them over email. As a consequence, it is possible that other attributes such as the user experience in computer science, as discussed by Onarlioglu et al. [22] could be more important to determine the risk factor of a user than her browsing habits.

Unfortunately, few works have tried to answer this question and understand if there are certain behaviors or certain characteristics that may influence the probability of users to

visit malicious web pages. As discussed in Section 7, some works have tried to answer similar questions by performing field studies on the computer usage of a limited number of subjects [13]. Others have speculated whether certain behaviors may be related to higher chances of being compromised, such as the relation between browsing porn sites and being subject to infections [35]. However, no study has so far been general enough to build user profiles and analyze this information in order to assess if there is any relation between specific user habits and the probability of visiting malicious web pages.

In this paper, we conduct the first comprehensive study in this area by using the telemetry data collected by Symantec. In particular, we analyzed the webpages visited by 160,229 users over a period of 3 months (92 days). Using anonymized information, we first identified two classes of users: the *safe* ones who never visited malicious webpages during our experiments, and the ones *at risk* who visited several malicious sites in the same timespan. Our goal was to determine what kind of behavior can be used to differentiates the two classes. For this reason, we identified and extracted 74 attributes that can be used to summarize the user browsing behavior, and we correlated each of them with the users' class. It is important to note that correlation does not necessarily imply the presence of a causality relationship. In our study we only analyzed the *voluntary* browsing activity performed within a browser, and we did not include any URL that did not originate from user actions (such as the ones visited automatically by malware to contact the Command and Control infrastructures). Therefore, the fact of being already infected could not affect the data we collected.

Our experiments confirm that the volume of user activity is one of the best indicators for the level of risk. The more pages a person browses everyday, and the more diverse is the set of pages, the more likely she would be to come across a malicious website. We also show that malicious pages are more likely to be encountered during the weekend and that people in the risk class are more active during the night than users who belong to the safe class. Looking at the website categories, we found that some of them – such as adult content and shortened URLs – are positively correlated to the probability of being at risk. Finally, the results of the experiments we performed indicate that it is possible to combine all this information and train a classifier to predict whether a user is at risk of infection, just by analyzing her browsing profile.

The rest of the paper is structured as follows: In Section 2, we present the dataset we used for our experiments and how we labeled it. In Section 3, we provide a discussion about time and geographical trends adopted by different user categories. In Section 4, we explain features we employ to define user profiles. In Section 5, we analyzed the correlations between the features and being at risk and, present how we predict each category of users using the 74 features. We then discuss our findings in Section 6, summarize related work in Section 7, and briefly conclude the paper in Section 8.

2. DATASET AND EXPERIMENTS SETUP

We performed our analysis using a telemetry dataset collected by Symantec. This data is obtained from clients that voluntarily opt-in to let their computers share information on usage statistics and encountered threats. AntiVirus (AV) vendors typically employ this kind of client feedback with the purpose of identifying new threats and improve their products and services.

The dataset we used in our experiments consists of a 3-month snapshot of the web browsing history of a subset of clients that had opted in to allow the company collect information on their browsing activity. The dataset covered all the web requests issued by approximately 160,000 distinct client machines in a three-month period, from August 1st, 2013 to October 31st, 2013. This consisted in a total of 202,306,687 URL visits, covering a total of 37,797,151 distinct URLs. The data collected by Symantec included only URLs of websites visited through the HTTP protocol. All information was provided in an anonymized form, and no private client information was available to us, with the exception of the client's country. It is important to note that customers who agreed in sharing their browsing history are aware that the company stores this information in anonymized form, and that client identifiers are anonymized too. This means it is not possible for the AV company to link back the collected data to the client from which the requests originated. The specific fields we have analyzed in our study include only the unique client identifier, the timestamp of the visit, and the URL of the web site. Moreover, to further improve the privacy of the users, we anonymize each URL by removing the path and eventual URL parameters – limiting our analysis to the fully qualified domain name.

Since our main goal is to perform a statistical analysis of the dataset, we focused our study on those clients who visited at least 100 web pages during our timeframe. This prevents clients whose information has low statistical significance to pollute, or bias, our measurements. We believe the threshold of 100 pages over 3 months to be conservative enough to include almost all regular user behavior, while excluding those machines that are only sporadically used to browse the Web. Visiting less than 100 web pages over three months means, in average, around one URL per day: it would be very difficult to build a user profile based on such a limited browsing history.

2.1 Data Labeling

To be able to estimate if there is a correlation between risk and user behavior, we first need to define what the definition of *risk* is for our study. As explained in Section 3, we define the risk categories by setting an experimentally chosen threshold for the number of times a user visits distinct malicious URLs or domain names during the experiment period.

We constructed our labeled set of malicious URLs from URLs detected to be malicious either by the Norton SafeWeb service [30] or by Google SafeBrowsing [27]. We further collected malicious domain names from several public services that provide a list of domains involved in various malicious activities, such as drive-by-download, phishing, and scam web sites. In particular, we built the list by merging information collected by malware domain list [15], abuse.ch [3] and malc0de [18].

All this information allowed us to label each URL in our dataset as either *Benign*, *Malicious*, or hosted in a *Blacklisted* domain. We decided to keep this last class separated from the malicious URLs because domains have a larger granularity and therefore provides a less accurate classification. Please note that the labeling phase was performed

in an automated way on Symantec's servers, thus prior to discarding the full URL path. Once the matching was completed, the rest of the analysis only operated, in an aggregated form, on the anonymized URLs.

2.2 Risk Categories

One of the goals of this paper is to answer the question of whether it is feasible to identify a category of people that, while surfing the Internet, incurs in a higher chance of visiting malicious web pages, when compared to other users. To be able achieve this goal, we first need to separate users in different risk categories.

Following a classical insurance approach, we separate users based on their past experience. With a good approximation, users that never ended up visiting a malicious page during our three-month observation period can be considered *safe* users. We noticed, however, that the contrary is not necessarily true. Indeed, given the high number of factors contributing to the maliciousness of a website and the delay in updating popular blacklists, misclassifications are not too rare. For example, it happens even to trusted websites to serve malicious ads or to become victims of DNS hijacking attacks [33]. Thus, when looking at our classification scores, a certain noise margin has to be taken into account. To handle this problem, we define a user to be *at risk* if she visited at least two distinct malicious URLs, or at least three blacklisted domains over the 3-month period. Again, the reason to use different thresholds for URLs and Domains is that the latter have a lower granularity and thus a higher probability of misclassification.

We put users who do not belong to the previous two categories into an *uncertain* middle category. For instance, the fact that a person visits a single dangerous URL over three months (with multiple visits to the same URL counting as one) may be just due to an error in classifying the URL. This is not sufficient for us to conclude that the user has a risky browsing behavior.

Table 1 shows the average number of different types of URLs visited by each category of users. Users who are "not at risk" appear to browse over five times less malicious URLs than *at risk* users. This means that typically, as the table shows, users in the *uncertain* category end up on malicious websites less frequently than *at risk* ones. Another clear difference is that *at risk* users typically visit more pages than other categories of users, and this factor may be related to the chance of ending up on malicious websites (we will discuss this hypothesis in more detail in Section 5). This is also valid in relation to the "variety" of visited websites, since for *at risk* users the average number of *distinct* URLs, and *distinct* URLs per day are about twice as much as the same values for the *uncertain* group. Finally, the table highlights that roughly one user out of five in our dataset belongs to the *at risk* category. If we consider the total number of users who are exposed at least once to malicious websites, then, this ratio increases to half of the entire user population. This is more than what found by a recent study on Australian customers by one major AV company [12], that observed that one customer every eight was exposed to web threats.

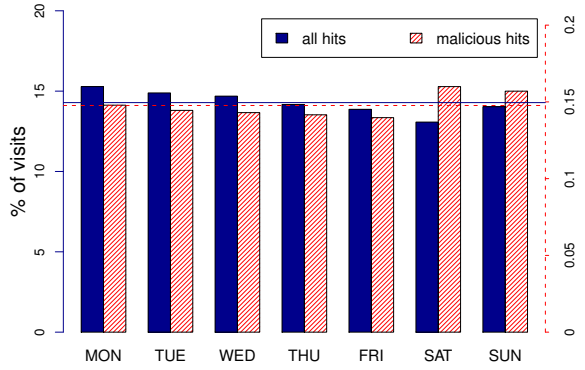

Figure 1: Global daily distribution of URL hits. The percentage of malicious hits is expressed as a fraction of the total hits on the same day.

3. GEOGRAPHICAL AND TIME-BASED ANALYSIS

This section describes the preliminary analysis we have conducted on our dataset, providing details about time and geographical trends.

3.1 Daily and Weekly Trends

We start our analysis of users' browsing behavior by looking at the weekly and hourly trends emerging from our dataset. First of all, Figure 1 shows that, as we expected, people surf less during the weekend. This trend is valid, with slight variations, all over the world, and country-wise daily trends do not differ much between each other. One can notice that there is a slight but significant increase in the percentage of malicious URLs visited during the weekend, compared to the trend of malicious hits during the rest of the week. This amounts approximately to a 10% increase in the chance of incurring in a malicious URL during the weekend, compared to the risk of doing so between Monday and Friday. The average *p-value* when comparing the two distributions is 6.44×10^{-7}, which shows the difference is indeed significant (as often found in literature [28], we consider to be statistically significant those differences showing computed *p-values* of less than 0.05).

Figure 2 shows instead the hourly trends for website visits, split between the two categories of users. As the hourly trends show, browsing trends for the *safe*, and *at risk* users do not differ much, even though *at risk* users are slightly more active during the night and less active in the morning. The statistical significance of these variations if confirmed by means of the Wilcoxon Signed Rank Test, that returned *p-values* significantly lower than the typical 0.05 significance level (e.g., the *p-value* of the test between 1 and 2am was of only 2.2×10^{-16}). However, the fact that users in the *at risk* category spend more time on the Internet at night does not imply that it is more risky to browse after midnight. Therefore, in Figure 3, we look at the same hourly trends, but from the point of view of the URLs instead of

Value	Risk Category		
	safe	*uncertain*	*at risk*
Total number of visited URLs	743	1386	2411
Distinct URLs visited	231.3	452.4	873.7
Average number of URLs visited per day	16.8	23.8	36.6
Distinct URLs visited per day	5.8	8.5	14.0
Total number of malicious URLs visited	0	0.78	8.4
Total number of blacklisted domains visited	0	2.44	8.5
Distinct number of malicious URLs visited	0	0.5	4.0
Distinct number of blacklisted domains visited	0	0.9	2.8
Percentage of malicious URLs	-	0.14%	0.71%
Percentage of blacklisted domains	-	0.32%	0.4%
Number of users	80128 (50%)	49127 (31%)	30974 (19%)

Table 1: **Average** values of different indicators, for users in the three risk categories.

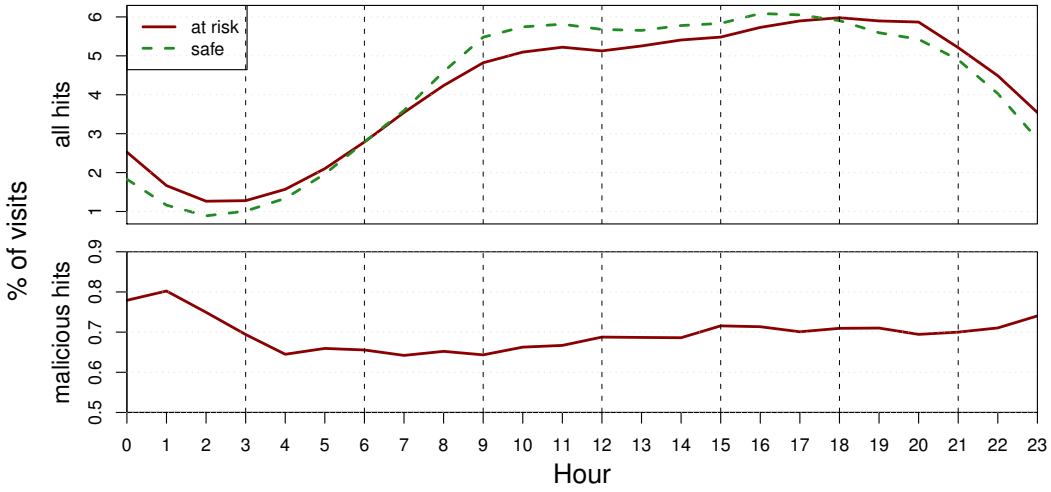

Figure 2: Hourly trends for, respectively, all the hits (upper) and malicious hits (lower) in our dataset. Malicious hits are expressed as percentage of the total hits for the same category of users, in the given hour.

users. In this case, the graph shows that hits on blacklisted domains are higher than other malicious hits between 9pm and 2am, and lower than others during business hours. Hits on malicious URLs seem instead to be prevalent in the afternoon, between 3pm and 8pm. Again, the signed rank test confirmed the differences in these distributions to be statistically significant.

Overall, these results confirm what found by a recent report on the Australian customers of a known security firm [12]. Indeed, as the mentioned study shows, also our analysis of time trends shows an increase in the percentage of malicious hits during nights and weekends. We are thus able to confirm that trends that have been reported on Australian users still hold when observing browsing statistics of users from all around the world.

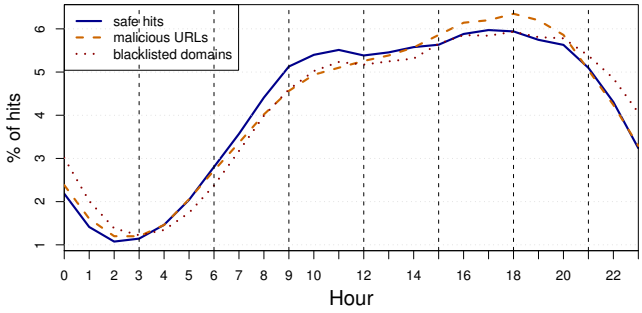

Figure 3: Hourly global trends for all hits and malicious hits in our dataset, showing also trends for the two separate sources of malicious hits

3.2 Geographical Trends

Our dataset contains information about clients located in 167 different countries. Table 2 summarizes some general statistics for those countries for which we have at least 1000 users. Simply by looking at the outliers (emphasized in bold in the table), one can notice several interesting trends.

For instance, Japan appears to have by far the lowest per-user ratio of malicious hits, and the lowest percentage of users at risk. However, the absolute value of malicious pages visited by Japanese users is in line with the ones of other countries. Percentages are just lower because average users

Country	Users	% users *at risk*	Average hits on		Visited Pages			# languages
			malicious URLs	blacklisted domains	total	distinct	domains	
US	67967	20.8	2.2 (0.22%)	2.0 (0.15%)	1250	422	194	3.6
UK	26204	17.8	1.5 (0.16%)	2.0 (0.16%)	1097	379	183	4.2
JP	16556	**10.0**	1.1 (**0.05**%)	3.1 (0.14%)	**1989**	**641**	205	3.8
CA	6798	20.9	2.0 (0.22%)	2.4 (0.17%)	1214	387	186	3.8
AU	6107	16.4	1.5 (0.17%)	1.5 (0.15%)	1007	343	173	3.7
DE	5606	22.3	2.0 (0.20%)	2.6 (0.23%)	1042	366	192	4.9
FR	4566	**29.1**	2.8 (0.27%)	3.3 (0.27%)	1127	390	209	4.5
NL	3415	15.9	1.1 (0.12%)	2.3 (0.21%)	1009	361	195	5.2
ES	1842	**28.3**	2.4 (0.23%)	3.9 (**0.33**%)	1121	391	200	**5.7**
SE	1755	15.3	1.9 (0.17%)	1.9 (0.14%)	1049	327	167	6.4
IT	1665	**27.4**	1.8 (0.18%)	**7.0** (**0.69**%)	1097	350	186	5.4
BE	1454	21.3	2.2 (0.21%)	2.5 (0.20%)	1126	396	208	5.5
NO	1208	**11.8**	1.1 (**0.10**%)	2.5 (0.11%)	1219	341	166	**6.1**

Table 2: **Average** values of several indicators, for users in the top 13 countries appearing in our dataset.

in Japan browse twice as many pages as their counterparts in other countries. At the other end of the scale, we have several Mediterranean countries (notably France, Spain, and Italy) that share similar high values of several risk indicators. These countries have a percentage of *a*t risk users ranging between 27% and 29%, three times higher than Japan, and approximately twice as much as other top countries.

Finally, the last column of the table shows the average number of languages of web pages visited by users in each country. As it can be expected, users from English speaking countries appear to visit pages in a limited number of languages compared to those visited by users in non-english speaking countries. In average, over the 3-months period, users in English speaking countries appeared to browse pages written in less than 4 different languages, while users from other countries visited pages in an average of 5.3 different languages. This fact seems however not to have any clear relation with the percentage of *a*t risk users in each country.

4. FEATURE EXTRACTION FOR USER PROFILING

After looking at time patterns and geographical trends, we decided to focus in more detail on the behavior of users. The basic idea motivating this work is that we expect users that belong to the risky category (i.e., those that regularly visit malicious web pages) to behave differently, when browsing the Internet, than users who are safe. We thus define a user profile as a sort of a multi-dimensional template such that we can characterize the behavior of each user group. In particular, we model each user profile by using a combination of 74 unique variables, or *features*, designed to precisely capture many aspects of a user's browsing habit.

Due to space limitations we cannot discuss each individual feature. Instead, in the following paragraphs we will describe the different categories that compose our features set, based on which aspect of the user behavior they are meant to represent.

How Much a User Surfs the Web.

The first set of features are designed to capture the volume of user activity. The rationale behind this is that, as common sense suggests, the more time a person spends browsing the Web, the more likely she would be to encounter a malicious web site. For instance, this category of features includes the number of total (*hits*) and unique (*distinct_urls*) hits over the three month period, the average number of hours (*hours_per_day*) and web pages (*hits_per_day*) visited in a day, and the percentage of days in which the client was active during the period of the experiment (*days_active_perc*).

In Which Period of the Day a User is More Active.

When looking at time trends (see Section 3.1), we noticed that the distribution of malicious URLs is proportionally higher during the night and in the weekend. Therefore, we added a set of features to capture this difference. In particular, we introduced the percentage of the client's hits issued, respectively, during night time (*hits_night_perc*), business hours (*hits_bh_perc*), and the evening (*hits_evening_perc*). We consider midnight to 6 am as night, 6 am to 7 pm as business hours, and 7pm-midnight as evening. These time windows were chosen in order to be as conservative as possible in the evaluation of business hours, as these differ significantly between countries in the world, and thus we may otherwise wrongly categorize this value.

How Diversified is the Set of Websites Visited by a User.

Another possible root cause for being at risk when browsing could be related to browsing a very broad and diversified class of web sites as this might increase the likelihood of landing on a malicious page. We model this aspect by counting the number of visited hostnames (*hostnames*), domain names (*domains*), and the number of visited unique top level domains (TLDs) (*tlds*).

Among these features we also consider the percentage of distinct URLs calculated over the total number of hits for the given user (*distinct_urls_perc*) and the percentages of unique hostnames and of unique domain names over the total number of hits (*distinct_hostnames_perc*, *distinct_domains_perc*). The purpose of these features is to estimate if the user tends to revisit the same set of URLs or websites, or instead appears to browse a more diversified set of web sites.

Finally, we measured the number of languages of the web pages visited by a certain user (*n_languages*). The language of websites was obtained from the same service we used to obtain the category of web pages, as explained in the following paragraph.

Which Website Categories the User is Mostly Interested in.

One of the main characteristics of a user profile is the categories of the visited web pages. To label each URL with the corresponding category, we used an internal website categorization system from Symantec. This service was designed to apply a set of heuristics to extract categories and languages from the URLs visited by the AV customers. Unfortunately, website categorization being based on heuristics, in some cases the categorization engine was able to retrieve only the main language used in the web page, but failed to properly capture the category, or vice versa. Therefore, we complemented the company database by using a number of publicly available website categorization services such as Alexa [1] and Open Directory Project [24], and a number of lists of known URL shorteners, bittorrent web sites, one-click hosting providers and porn websites [16, 31, 32, 36]. We employed these lists to complement the heuristic service provided by Symantec, as it is common belief that visiting websites from these categories yields to higher chances of being infected by malicious web code. As a result, we were able to cover 76% of the web sites in our dataset (96% for the Alexa top 10,000 domains). The language coverage was instead 77% overall, and 70% for domains in the Alexa top 10,000.

Once the website categorization phase was completed, we extracted a number of features extracted from the category information in the user profile. For example, we reported the percentage of activity in each of the following 8 categories: business websites, adult websites, communications and information search, general interest, hacking, entertainment and leisure, multimedia and downloading, uncategorized (sites for which we were not able to obtain a category).

Computer Type.

The main aspect we want to capture with this class of features is the difference between office and personal computers. The assumption we make to identify office computers is that computers that do not show any activity during the weekends are very likely to be office computers. We label all computers that are silent during the weekend as office and the others as personal computers (*work_pc*). In addition, for personal computers, we also compute the percentage of activity during the weekend (*hits_we_perc*).

The remaining features that are extracted to characterize the computer type are computed using properties of the anonymized IP addresses of the devices. Note that we do not have access to the absolute value of the IP addresses and to the name of the Internet Service Providers (ISP) they belong to. The AV company keeps the hashed values of IP addresses and their corresponding ISPs such that it is possible to calculate their distinct numbers (*n_ips* and *n_isps*). The final feature in this category is the number of countries from which the user appeared to be surfing the web from (*n_countries_user*). By using features, we aim at representing the user's mobility, and helping to assess whether a person is browsing the Internet from a static IP address or a dynamic one.

How Popular are the Websites Visited by the User.

This set of features are computed to model how common the websites visited by a user are, under the assumption that malicious pages are more commonly found in less popular sites.

The first indicator we look at is the percentage of `.com`, `.net` and `.org` top level domain (TLD) hits that appear in each user's browsing history (*hits_comnetorg_perc*), and the number of visited URLs that belong to other TLDs (*no_comnetorg_tlds*). We also extracted a number of features related to the Alexa ranking of domains [2], namely the total number of hits and distinct websites visited in the Alexa top 500 (*n_hits_top500*, and *n_dist_sites_top500*), the total number of hits and of distinct websites visited in Alexa's top one million (*n_hits_top1M*, and *n_dist_sites_top1M*) and the number of hits and of distinct websites visited out of Alexa's top one million list (*n_hits_noAlexa*, *n_dist_sites_noAlexa*). All these features are computed both as absolute numbers and as percentage among all web sites visited by the user.

How Stable is the Set of Visited Pages.

To conclude our features set, we modeled the variability of a user's browsing activity. The rationale in this case is that users who always visit the same set of pages may be less at risk than users who change their targets very often. In particular, it is possible that users are mostly exposed to malicious pages when they deviate from their usual interests and temporarily browse web sites they do not know very well. In order to obtain these features, we performed a one week training over the web browsing history of each client in our dataset. We thus recorded, for each client machine, the set of web sites visited during its first 7 days of activity (*training set*). For every other day, we recorded the set of visited websites and their intersection and difference with the *traning set*. We averaged these values, and obtained the average percentage of common web sites between the daily browsing session and the initial 7-day training period (*inters_day_host*), as well as the percentage of new web sites – not visited during the 7-day training window – browsed in average every day (*delta_day_host*). We also recorded, for each client machine, the whole set of web sites visited during all its browsing activity, and calculate its intersection with the *training set*. The size of this intersection is then scaled by the size of the training set to obtain the percentage (*inters_host*) of hosts visited after the training period that cover the initial *training set*. Finally, we also computed a measure of the increment in the number of web sites visited by the client during its entire browsing history, compared to the number of web sites in the initial *training set* (*inc_host*).

5. EVALUATION

In this section, we present the results of our analysis. We first evaluate if any of the 74 features we have presented in the previous section are correlated to the fact that a user belongs or not to the *at risk* category. Afterwards, we build on top of these results to see if it is possible to use these features in a classifier, to predict the risk class of a user given her behavior.

Feature	At Risk	Safe	Percent Difference
hits	2411	742	106%
distinct_urls	873	231	116%
domains	331	88	116%
hostnames	388	108	113%
TLDs	17.5	7.9	76%
no_comnetorg_tlds	14.6	5.2	94%
n_languages	5.4	3.4	45%
days_active_perc	0.66	0.46	36%
hits_night_perc	0.09	0.07	27%
n_dist_sites_top500_perc	0.12	0.18	38%
hits_per_day	36.5	16.8	74%
hours_per_day	4.7	3.1	41%
hits_shorteners_perc	0.0034	0.0017	67%
hits_och_perc	0.0030	0.0018	50%
hits_porn_perc	0.0282	0.0112	86%
hits_downloading_perc	0.051	0.033	42%
hits_hacking_perc	0.0002	0.0000	113%
hits_business_perc	0.147	0.220	39%
hits_adult_perc	0.144	0.042	109%
inters_day_host	0.125	0.190	41%
inc_host	11.7	8.0	38%

Table 3: Comparison of the **average** values of certain features for *safe* and *at risk* users. Only features having a percentage difference greater than 25% are shown.

5.1 Feature Correlations

In the first part of our experiments, we extracted the feature values for all the users in our dataset and we used them to perform a correlation analysis with the risk class.

To start with, we compared the average values of each feature computed on *safe* and *at risk* users. While for the majority of them the difference was relatively small, in 22 features there was a difference of more than 25%. Those features, and the respective average values, are summarized in Table 3. The fact that several features clearly show up to a threefold increase between the activity of *safe* users and users *at risk*, suggested us to look at the correlation of these variables in more detail.

The correlation analysis we have adopted is based on the value of the Spearman's correlation coefficient [10]. Spearman's correlation is a statistical measure of the strength of a monotonic relationship between paired data. This coefficient by design is constrained between -1 and 1. While -1 indicates very strong negative correlation and +1 very strong positive correlation, the values close to 0 denote the absence of a monotonic relation between variables. We chose to employ the Spearman's rank correlation because, unlike the Pearson's correlation coefficient, it does not require a normal distribution in the dataset.

After calculating the Spearman's correlation coefficient, we tested the confidence of the obtained results by performing a standard significance test. As already discussed in Section 3.1, we consider that the correlation value of a feature is statistically significant if the computed *p-value* (i.e, significance level) is less than 0.05. A set of selected features for which the *p-value* was under this threshold is summarized in Figure 4.

Note that in the literature [10], a Spearman's coefficient value below 0.20 is normally considered an indication of a *very weak* correlation. Similarly, values between 0.2 and 0.4 are considered *weak*, between 0.4 and 0.6 *moderate*, between 0.6 and 0.8 *strong*, and above 0.8 *very strong* indication of a correlation between the variables. As it can be clearly seen in Figure 4, most of the features we have extracted have weak or no correlation with the fact that a user is at risk.

In the weakly correlated category we find features related to the amount of daily web activity (hits and hours per day), the number of porn and adult websites visited by a user, the number of languages, and an inverse correlation with the percentage of visited websites falling in the top Alexa 500. In the moderate correlation interval we find again some absolute measures of the amount of URLs, domains, and hostnames visited by a user. Moreover, and more interestingly, we also find a correlation between being at risk and the number of web pages with a TLD different from `.org`, `.com`, and `.net`.

Not surprisingly, these results indicate that the more a user surfs the Internet, the more she might be exposed to the risk of encountering a malicious page. The category does not seem to matter much, with very little correlation found with the percentage of usage of URL shorteners, downloading, and hacking websites – and a small negative correlation with the percentage of business sites. The only exception, as discussed in more detail in Section 6, is the higher correlation with adult and porn categories.

5.2 Predictive Analysis

The results we obtained from the correlation analysis show that some of the features we examined, although not very strongly, have some mild correlation with the fact that a user is *at risk*. Therefore, the same information might be helpful as well to predict whether a user is at risk or not. Motivated by this assumption, we have generated a number of prediction models leveraging state-of-the art machine learning techniques. Before opting for Logistic Regression [5], we experimented with many other machine learning approaches including decision trees [14,25], support vector machines [7], and Bayesian classifiers [37]. In our tests, logistic regression achieved the best results in terms of accuracy and false positive rates.

Logistic regression is a probabilistic statistical classification model that aims at predicting a category from features presenting either continuous or discrete values [5]. Compared to other classification methods, one of its advantages is that it does not explicitly require features that are not correlated with each other. Moreover, when new data is available, logistic regression can efficiently update the models with the new input.

The Receiver Operating Characteristic (ROC) curve summarizing the true and false positive rates of our classifier, applied to the entire dataset, is shown in Figure 5. The curve has an area of 0.919. For instance, if used to detect *at risk* users, the system can be tuned to have a detection rate of 74% (i.e., users *at risk* properly classified as *at risk*) with 8% false positives (i.e., users not at risk misclassified as being *at risk*). These results, both in terms of the best classifier algorithm and of the area under the ROC curve, are in line with what has been measured in previous studies about the precision of classification algorithms for financial risk prediction [23].

Since, as explained in Section 3.2, the distribution and behaviors of users' risk classes are different in different countries, we decided to retrain our classifier on each country in isolation. The results are quite similar to the overall results,

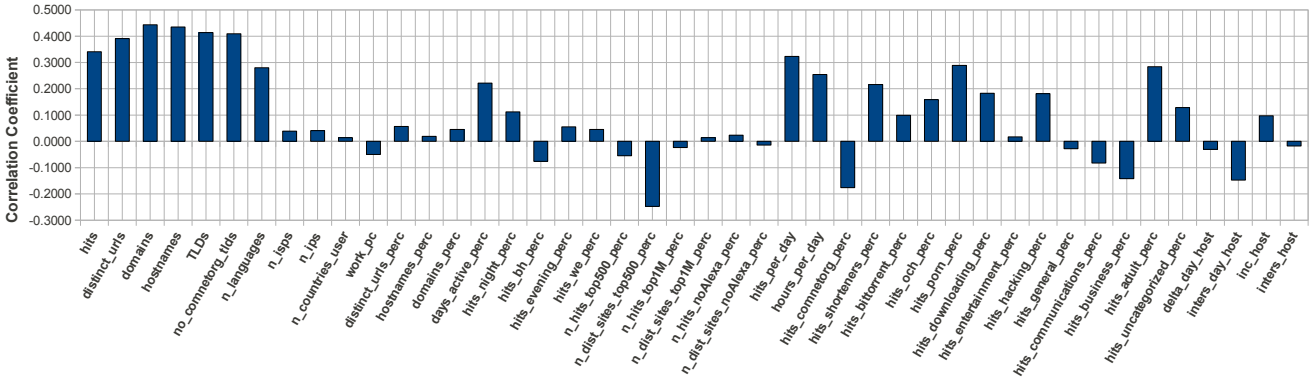

Figure 4: Spearman's Correlation Coefficient between user profile features and being *at risk*.

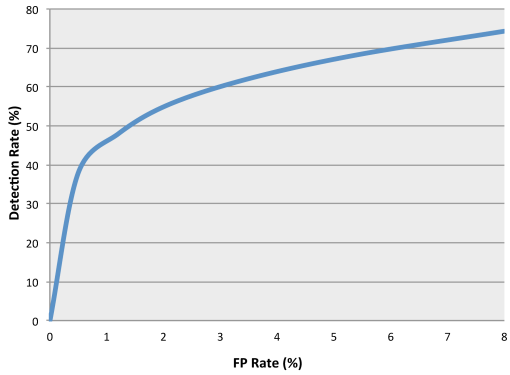

Figure 5: ROC Curve of the risk class classifier applied to the entire dataset

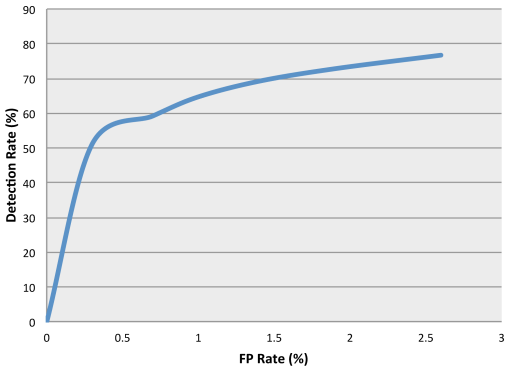

Figure 6: ROC Curve of the risk class classifier applied to the Japanese users only

with the only exception of Japan, for which the system was more precise – achieving 73% detection with 1.9% false positives, and an area under the ROC curve of 0.958, as shown in Figure 6.

6. DISCUSSION AND LESSONS LEARNED

Our experiments confirm the finding of a recent study by Levesque et al. [13] regarding the fact that the more websites a user visits the higher is her exposure to threats. However, we reach a different conclusion regarding the correlation of browsing adult content. Levesque et al. found that the amount of sport or Internet infrastructure websites visited by a user are more related to the fact of being infected than the number of porn websites [13]. However, while the absolute number of porn website may not be a good indicator, the percentage of time spent browsing porn seems to be a better feature. In fact, even though browsing adult web pages may not be a risky activity per se, from our results it looks like people who browse (in proportion) more porn and adult websites are more likely to also visit malicious pages in their daily activity. To better investigate this phenomenon, in Figure 7 we plotted the percentage of users at risk for different ranges (using a decile split) of the *hits_adult_perc* feature. The graph clearly shows that the percentage of safe users decreases as the ratio of adult hits increase.

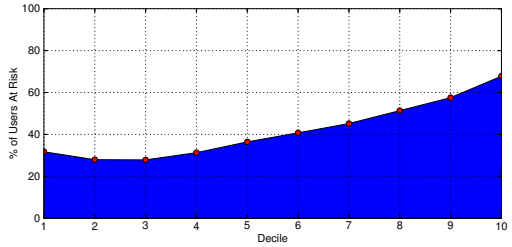

Figure 7: Decile plot for *at risk* users with respect to the percentage of hits on adult web sites.

Similar trends can be plotted also for the other variables. For instance, Figure 8 presents an even clearer picture of the relationship between the number of unique top level domains and the steep increase in the percentage of users at risk. While the large majority of safe users lie in the first half of the decile plot (more than 50% of the total number of safe users actually lie in the 1st decile – visiting less than 8 TLDs in the 3-month period), their percentage drops to less than 20% in the last three deciles of the plot, corresponding to users that visit at least 21 different TLDs. This is yet another confirmation that the variety, and not just the number, of the pages visited by a user is a strong indicator for its risk factor.

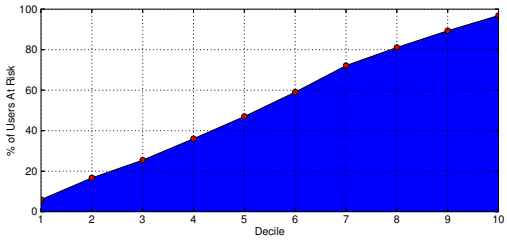

Figure 8: Decile plot for *at risk* users with respect to the number of different TLDs visited.

As mentioned in Section 3.2, also the geographical location of a user is very important. Citizens of certain countries (such as Norway and Japan) seem to be more careful in their browsing habits compared to their peers located, for example, in Italy or Spain.

While some of the features we used in our experiments were already discussed in previous works, the main finding of our study is the fact that by extracting and combining a much larger number of features from the URLs visited by a user in a certain period of time, it is possible to train a classifier to predict the risk class a user belongs to. This is a very interesting finding from several points of views. First, our results can be obtained by looking only at HTTP requests, without any access to the end-user devices. This allows companies (or even ISPs) to silently profile their users and even combine their risk class into an aggregated risk factor at a company or network level. Second, while still far from perfect, the accuracy of the extracted models is sufficient to be used in a risk prediction scenario. Comparable models are used everyday to compute the risk associated to financial operations [23], such as when processing credit card requests. This opens the door to a simple yet effective way to implement a cyber-insurance mechanism that rewards users who show a safe browsing profile.

7. RELATED WORK

Web-based attacks are one of the main vectors for cybercriminals to compromise Internet users. Therefore, there has been a considerable amount of work aimed at building defense mechanisms for web attacks [4, 9, 26]. However, due to lack of data about people's web browsing habits and experiences, the number of studies that tried to understand if there is any relation between users' behaviors or characteristics and their probability of visiting malicious web pages is still very limited. Moreover, most of the existing studies have been built upon the observation of very limited customer bases, or on clinical-style case studies based on data collection and surveys on tens or hundreds of users in a monitored environment [13].

7.1 User-based Risk Analysis

One of the first studies that sought to understand the risk factors behind user infections was carried out by researchers at the École Polytechnique de Montréal and Carleton University [13]. This study, which has a similar nature to works that evaluate medical interventions, examined the interactions among three important players: the users, the AV software and the malicious software detected on the system. The authors provided 50 users that accepted to join the experiment with a laptop that was configured to constantly monitor possible malware infections and collect information about how the users behaved in such cases. The results of the experiment show that user behavior is significantly related to infections, but that demographic factors such as age, sex, and education cannot be significantly related to risk. Furthermore, innocuous categories of websites such as sports and Internet infrastructure are associated with a higher rate of infection when compared to other categories, such as porn and illegal content sites, that common sense traditionally associates with higher risk. Similarly, and surprisingly, computer expertise seems to be one of the factors positively related to higher risk of infection.

Onarlioglu et al. [22] studied the behavior of users when they are faced with concrete Internet attack scenarios. The authors built an online experimental platform that was used to evaluate the behavior of 164 users with different backgrounds. Their findings confirm that non-technical users tend to fail in spotting sophisticated attacks, and that they are easily deceived by tricky advertising banners. On the other hand, they are able to protect themselves as effectively as technical users when dealing with simple threats.

Maier et al. [17] conducted a study of the security hygiene of approximately 50K users from four diverse environments: a large US research institute, a European ISP, a community network in rural India, and a set of dormitory users of a large US university. The paper analyzes anonymized network traces, which were collected from each observed environment for a period that ranges between 4 and 14 days, containing only the first bytes of each connection. The results of the analysis indicate that having a good security hygiene (i.e., following antivirus and OS software updates) has little correlation with being at risk. However, on the other hand, risky behaviors such as accessing blacklisted URLs double the likelihood of becoming infected with malware. Unlike our work, this paper has the advantage of being able to monitor all the Internet activity of the users, and thus is not limited to the analysis of web browsing traffic. However, it considers only malicious behaviors that overtly manifest themselves at a network level, e.g., sending spam emails, performing address scans or communications with botnet C&C servers. For this reason, attacks that produce little traffic, or install themselves on victim's machines and remain latent for long periods of time, may have been missed.

Finally, a recent report by TrendMicro and the Deakin University provided an analysis on the Australian web threat landscape [12]. The report states that, in average, 0.14% of the web browsing hits collected by the AV company are malicious in nature. One interesting finding is that Australians seem to incur in a higher percentage of daily malicious hits during holidays and weekends, when compared to weekdays. Moreover, the percentage of malicious hits rises during night time, with a peak around 4 am. As explained in Section 3.1, these findings are also confirmed as a worldwide trend by our experiments. Finally, the reported statistics show that one out of eight Australian IP addresses are exposed to web threats, on a typical day. In our study, instead, we find a higher risk of exposure to web threats for users in our dataset (19% of *at risk* users, overall).

Our work is fairly different from these studies in many respects. Compared to the majority of previous works, we perform our analysis on a much larger dataset (i.e. three

months of data generated by 160K distinct users). Moreover, our analysis does not rely on personal information about the users such as their educational background, sex and age. We significantly extend the study of the Australian threat landscape by conducting similar analysis on a *worldwide* basis. In addition, we performed a more precise and deeper analysis by building user profiles based on over 70 features, and we tried to understand if different risk categories have different probabilities to end up in malicious web sites.

7.2 User Profiling

Olejnik et al. [21] presented a study in which they evaluated the possibility of fingerprinting users given their past web browsing history. The methodology the authors adopt to fingerprint users was able to profile 42% out of approximately 380,000 users involved in the study. From the experiment they performed with only 50 web pages, they conclude that categorization information of visited websites could be a useful parameter to build more accurate user profiles. The results of our study confirm that categorization information could be used for user profiling.

The problem of user profiling has been largely studied within recommender systems [8, 19, 20], to help users find topics that are in their interest, while hiding those topics that are unattractive to them. Therefore, the goals of user profiling in recommender systems' research are completely different from ours.

8. CONCLUSIONS

In this work, we have presented a first step towards the prediction of users' risk when browsing the Internet. Our in-depth analysis of a large telemetry dataset collected by one of the major AV companies allowed us to gain a number of insights on the relation between users' browsing habits and their chances of visiting malicious web pages on the Internet. For example, we have been able to confirm some known trends, such as the fact that browsing the web late at night and during weekends is typically correlated with higher chances of ending up on malicious web sites. Another general trend confirmed by our work is that, in general, the more a user surfs the Internet, the more her chances are of ending up in some "unsafe neighborhood".

We have also been able to shed some light on whether profiling can be effectively used as a basis for predicting the risk for a user to end up on malicious web sites. By employing machine learning, we showed that user profiling could actually be employed, at least to some extent, in predicting the class of risk for a user on the web, similarly to what is currently done in the field of insurances and financial risk prediction.

In order to provide even more complete insights on human risk factors linked to visiting malicious web pages, it would be interesting to have access to users' "social" features, such as their sex, age, profession and personal interests. This would improve the completeness of users profiles we build, and could be the object of some future work.

9. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n.257007.

10. REFERENCES

[1] Alexa. Alexa Browse by Category. `http://www.alexa.com/topsites/category/Top`, 2013.

[2] Alexa. Alexa Top Websites. `http://www.alexa.com/topsites`, 2013.

[3] amada.abuse.ch. Malware Database (AMaDa) :: AMaDa Blocklist. `http://amada.abuse.ch/blocklist.php?download=domainblocklist`, 2013.

[4] A. Barth, C. Jackson, and J. C. Mitchell. Robust defenses for cross-site request forgery. In *15th ACM Conference on Computer and Communications Security (CCS 2008)*, 2008.

[5] C. M. Bishop. nformation Science and Statistics. In *Pattern Recognition and Machine Learning*. Springer, 2006.

[6] R. Böhme and G. Schwartz. Modeling cyber-insurance: Towards a unifying framework. In *Ninth Workshop on the Economics of Information Security (WEIS)*, 2010.

[7] N. Cristianini and J. Shawe-Taylor. An introduction to support vector machines and other kernel-based learning methods. In *Cambridge University Press*, 2000.

[8] J. Delgado and R. Davidson. Knowledge bases and user profiling in travel and hospitality recommender systems. In *Proceedings of the ENTER 2002 Conference*, pages 1–16. Citeseer, 2002.

[9] M. Egele, P. Wurzinger, C. Kruegel, and E. Kirda. Defending browsers against drive-by downloads: Mitigating heap-spraying code injection attacks. In *Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 88–106. Springer, 2009.

[10] Y. G.U. and K. M.G. *An Introduction to the Theory of Statistics (14th ed.)*. .Charles Griffin & Co., 1968.

[11] Kaspersky. Kaspersky Security Bulletin 2012. `http://www.securelist.com/en/analysis/204792255/Kaspersky_Security_Bulletin_2012_The_overall_statistics_for_2012`, 2012.

[12] C. Ke, J. Oliver, and Y. Xiang. Analysis of the Australian Web Threat Landscape. `http://www.trendmicro.com.au/cloud-content/au/pdfs/security-intelligence/white-papers/australian_web_threat_landscape_-v7.pdf`, May 2013.

[13] F. L. Lévesque, J. Nsiempba, J. M. Fernandez, S. Chiasson, and A. Somayaji. A Clinical Study of Risk Factors Related to Malware Infections. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, Nov. 2013.

[14] A. Liaw and M. Wiener. Classification and regression by randomforest. In *R News*, volume 2/3, page 18, 2002.

[15] M. D. List. Malware Domains List. `http://www.malwaredomainlist.com/`, 2013.

[16] URL Shortening Services - A List of URL Shorteners. `http://longurl.org/services`, 2013.

[17] G. Maier, A. Feldmann, V. Paxson, R. Sommer, and M. Vallentin. An assessment of overt malicious activity manifest in residential networks. In *Detection*

of Intrusions and Malware, and Vulnerability Assessment, pages 144–163. Springer, 2011.

[18] Malcode. Malcode. `http://malc0de.com/bl/BOOT`, 2013.

[19] D. W. McDonald and M. S. Ackerman. Expertise recommender: a flexible recommendation system and architecture. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 231–240. ACM, 2000.

[20] S. E. Middleton, N. R. Shadbolt, and D. C. De Roure. Ontological user profiling in recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):54–88, 2004.

[21] L. Olejnik, C. Castelluccia, and A. Janc. Why Johnny Can't Browse in Peace: On the Uniqueness of Web Browsing History Patterns. In *5th Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs 2012)*, Vigo, Espagne, July 2012.

[22] K. Onarlioglu, U. O. Yilmaz, D. Balzarotti, and E. Kirda. Insights into user behavior in dealing with internet attacks. In *19th Annual Network and Distributed System Security Symposium (NDSS)*, NDSS 12, January 2012.

[23] Y. Peng, G. Wang, G. Kou, and Y. Shi. An empirical study of classification algorithm evaluation for financial risk prediction. *Applied Soft Computing*, 11(2):2906 – 2915, 2011. <ce:title>The Impact of Soft Computing for the Progress of Artificial Intelligence</ce:title>.

[24] O. D. Project. DMOZ Open Directory Project. `http://www.dmoz.org/`, 2013.

[25] J. Quinlan. C4.5: Programs for machine learning. In *Morgan Kaufmann Publishers*, 1993.

[26] P. Ratanaworabhan, B. Livshits, B., and Zorn. Nozzle: a defense against heap-spraying code injection attacks. In *Proceedings of the USENIX Security Symposium*, 2009.

[27] Google Safe Browsing API. `http://code.google.com/apis/safebrowsing/`, 2008.

[28] S. Stigler. Fisher and the 5*CHANCE*, 21(4):12–12, 2008.

[29] Symantec. 2013 Internet Security Threat Report. `http://www.symantec.com/security_response/publications/threatreport.jsp`, 2013.

[30] Symantec. Norton Safe Web. `https://safeweb.norton.com/`, 2013.

[31] TBLOP - The Big List of Porn. `http://www.tblop.com/`, 2013.

[32] Torrent Sites. `http://www.torrentresource.com/`, 2013.

[33] S. J. Vaughan-Nichols. How the Syrian Electronic Army took out the New York Times and Twitter sites. `http://www.zdnet.com/how-the-syrian-electronic-army-took-out-the-new-york-times-and-twitter-sites-7000019989/`, August 2013.

[34] Websense. Websense 2013 Threat Report. `http://www.websense.com/content/websense-2013-threat-report.aspx?cmpid=prnr2.13.13`, 2013.

[35] G. Wondracek, T. Holz, C. Platzer, E. Kirda, and C. Kruegel. Is the internet for porn? an insight into the online adult industry. In *Ninth Workshop on the Economics of Information Security (WEIS)*, 2010.

[36] List of File Hosting and Sharing Websites. `http://xboxpirate.eu/forums/topic/280-list-of-file-hosting-and-sharing-websites-137-entries/`, 2013.

[37] H. Zhang. The Optimality of Naive Bayes. In *FLAIRS2004 conference*, 2004.