Predicting Cyber Security Incidents Using Feature-Based Characterization of Network-Level Malicious Activities

Yang Liu
University of Michigan
youngliu@umich.edu

Jing Zhang
University of Michigan
jingzj@umich.edu

Armin Sarabi
University of Michigan
arsarabi@umich.edu

Mingyan Liu
University of Michigan
mingyan@umich.edu

Manish Karir
QuadMetrics, Inc.
mkarir@quadmetrics.com

Michael Bailey
UIUC
mdbailey@illinois.edu

ABSTRACT

This study offers a first step toward understanding the extent to which we may be able to predict cyber security incidents (which can be of one of many types) by applying machine learning techniques and using externally observed malicious activities associated with network entities, including spamming, phishing, and scanning, each of which may or may not have direct bearing on a specific attack mechanism or incident type. Our hypothesis is that when viewed collectively, malicious activities originating from a network are indicative of the general cleanness of a network and how well it is run, and that furthermore, collectively they exhibit fairly stable and thus predictive behavior over time. To test this hypothesis, we utilize two datasets in this study: (1) a collection of commonly used IP address-based/host reputation blacklists (RBLs) collected over more than a year, and (2) a set of security incident reports collected over roughly the same period. Specifically, we first aggregate the RBL data at a prefix level and then introduce a set of features that capture the dynamics of this aggregated temporal process. A comparison between the distribution of these feature values taken from the incident dataset and from the general population of prefixes shows distinct differences, suggesting their value in distinguishing between the two while also highlighting the importance of capturing dynamic behavior (second order statistics) in the malicious activities. These features are then used to train a support vector machine (SVM) for prediction. Our preliminary results show that we can achieve reasonably good prediction performance over a forecasting window of a few months.

IWSPA'15, March 4, 2015, San Antonio, Texas, USA.

ACM 978-1-4503-3341-2/15/03 ...$15.00.
http://dx.doi.org/10.1145/2713579.2713582.

Categories and Subject Descriptors

C.2.0 [**General**]: Security and protection; C.2.3 [**Network Operations**]: Network Monitoring; C.4 [**Performance of Systems**]: Measurement techniques, modeling techniques

General Terms

Network Security, Measurement, Management

Keywords

Network security, Network reputation, Prediction, Temporal pattern, Time-series data

1. INTRODUCTION

This study seeks to understand to what extent we can predict whether a network may suffer a cyber security incident in the near future, by applying machine learning techniques and using externally observed malicious activities associated with that network. Our prediction goal is quite broad, in the sense that we are not targeting a specific type of security incidents; it could range from data breach to webpage defacement. At the same time, the external observation we rely on is also quite broad, including spamming, phishing, and scanning, each of which may or may not have direct bearing on a specific attack mechanism or incident type. Thus the fundamental underlying question is whether the latter collectively could provide valuable information on the overall security *risk* a network is facing.

Our hypothesis is that when viewed collectively, such malicious activities originating from a network (a set of IP addresses suitably defined, e.g., according to Autonomous System (AS), prefix, or other administrative domain) are indicative of the general cleanness of a network and how well it is run, and that furthermore, collectively they exhibit fairly stable and thus predictive behavior over time. This is because the factors influencing a network's cleanness or security posture generally vary on a relatively slow time scale, including various network policy related issues such as operating systems and patch levels, firewall policies, password strength checks, the expertise and training of IT personnel, and even user awareness levels.

By contrast, the common used host reputation systems or blacklists [1–3] collect and distribute information about such externally observed malicious activities associated with individual host IP addresses. These are routinely used in filtering and blocking policies adopted by network operators.

The highly dynamic nature of IP addresses [17] can severely limit the timeliness and accuracy of these lists.

To test our hypothesis, that network level malicious activities can reveal stable and predictable behavior which can be used for incident prediction, we need to be able to describe the network level malicious activities, a dynamic temporal process, in an efficient and effective way. Toward that end, we will utilize two unique datasets in this study. The first consists of 11 commonly used IP address-based/host reputation blacklists (RBLs) collected over more than a year starting in January 2013. The second consists of a set of security incident reports collected over the same period. The goal is to see whether by training a classifier using historical RBL data (January to September 2013) and the ground truth incident data in October 2013, we can effectively predict the incidents reported in the future months starting in November 2013.

Our methodology consists of first aggregating the RBL data to form a basic per-prefix temporal signal that represents the presence of that prefix on the blacklists (in terms of the total IPs or % of IPs belonging to that prefix being listed). We then introduce a set of three features that capture the dynamics of this signal. A comparison between the distribution of these feature values taken from the incident dataset and from the general population of prefixes shows distinct differences, suggesting that these features can indeed be used to distinguish between the malicious behaviors of these two datasets. This comparison also highlights the importance of capturing dynamic behavior (second order statistics) in the malicious activities. These features are then used to train a support vector machine (SVM) [15], which is then used to predict future incidents. Our preliminary results show that we can achieve fairly good prediction performance over a forecasting window of a few months.

This study serves as a first step toward the broader notion of using advanced learning techniques to extract useful information from large quantities of Internet measurement data in the cyber security domain. Much more remains to be explored, including collecting a higher qualify incident dataset and using a more diverse set of maliciousness data.

The remainder of the paper is organized as follows. Section 2 introduces the datasets this study is based upon and provides a rationale for aggregating information at the BGP prefix level. We then define a set of features and show why they are relevant in predicting security incidents in Section 3. In Section 4 we describe how the classifier is constructed and present prediction results. We conclude our paper in Section 6.

2. THE DATASETS AND PRELIMINARIES

2.1 RBL dataset

Our RBL data consists of 11 IP address-based reputation blacklists over more than a year starting in January 2013. The sampling rate is once per day, i.e., the list content is refreshed on a daily basis. Table 1 summarizes these lists and the type of malicious activities they target. They collectively target three major categories of malicious behaviors: spam, phishing/malware, and active attacks/scanning. All combined, this dataset includes 164 million unique IP addresses.

These lists indicate malicious activities seen by the outside world, and are routinely used by spam filters. They are obviously not perfect and contain both false-positives and false negatives which are generally unknown. This however does not prevent us from examining their effectiveness in predicting security incidents in a network, which may or may not be a function of these errors.

Type	Blacklist Name
Spam	CBL [2], SBL [11], SpamCop [9], WPBL [14], UCEPOTECT [12]
Phishing/Malware	SURBL [10], Phish Tank [8], hpHosts [6]
Active attack	Darknet scanner list, Dshield [3], OpenBL [7]

Table 1: The RBL datasets

2.2 RBL data aggregation

The RBLs consist of individual IP addresses, which are hard to use directly for prediction purposes. This is due to two reasons. Firstly, as we detail next the incident reports sometimes only identify an organization, not a precise IP address to which the incident occur. Secondly, the increasingly dynamic association between IP addresses and physical machines (e.g., owing to human mobility) leads to highly dynamic list content: these malicious activities originate from physical machines and as they move around they get associated with different IP addresses. For these reasons, we postulate that it would be much more effective if we consider malicious activities in the aggregate, for an entire network (however defined). Intuition as well as the following simple experiment appear to support this argument.

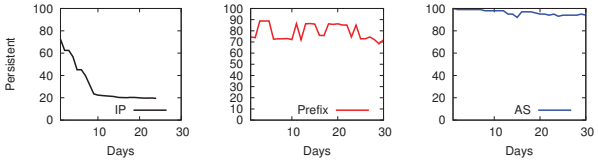


Figure 1: Persistency of malicious IPs, the worst prefixes, ASes.

We first combine lists within each type of malicious activities, resulting in a set of *lists*, or *raw lists*, more specifically referred to as the *spam list*, the *scan list* and the *phishing list*, respectively. An IP is included on a particular list on a given day if it shows up on at least one of the individual blacklists of that type. Combining all 11 lists results in a *union list*, which we use in the illustration below. We then combine entries on these lists according to their membership in the same prefix or the same Autonomous System (AS). This then allows us to rank the prefixes or ASes according to their presence on a list as measured by the percentage of its IPs being listed.

On a typical day, the 100 worst ASes have more than 70% of their respective IP addresses blacklisted; at the prefix level, the worst 9,000 (resp. 15,000) prefixes have nearly 100% (resp. 70%) of their IPs listed. Figure 1 shows the malicious IPs (or worst ASes/prefixes) on a randomly selected day (called the first day), which remains on the union list on day x as a function of x. As can be seen, these worst ASes/prefixes are much more persistent than individual IP addresses; the latter appears highly dynamic: only 20% of

Month	Oct.	Nov.	Dec.	Jan.	Feb.
Total	93	100	110	73	64
Identified	46	52	94	60	54

Table 2: Reported cyber crimes by month.

the IP addresses originally listed remain on the list during the one-month period, while 75% of the 15,000 worst prefixes and more than 90% of the worst ASes persist during the same period.

For the remainder of this paper we will use prefix as the aggregation level, as it offers a good balance between stability in behavior (and thus expected predictive power) and spatial as well as prediction resolution – after all, if we aggregate over a sufficiently large part of the Internet then an incident will happen with high probability, rendering any prediction meaningless. We do note that other levels of aggregation are also possible which remain to be investigated.

Once we aggregate at the prefix level over a particular raw list, we obtain a discrete-time *aggregate signal* for each prefix i denoted by $r_i(t)$, $t = 0, 1, 2, \cdots$. There are two types of aggregate one can define, the normalized version and the un-normalized version. For the normalized version, $r_i(t)$ is given by the fraction of the total number of IPs on the list and belonging to prefix i on day t, over the total number of addresses within prefix i. The un-normalized version of $r_i(t)$ is simply defined as the total number of IPs on the list and belonging to prefix i on day t. The dataset contains a total of 363,667 prefixes.

2.3 Incident reports dataset

To determine the predictive power of the RBL dataset, we shall rely on cyber attack incident reports from [13]. We will focus on incident reports for the months October 2013 to February 2014 to illustrate the training and testing process.

There are in all 93, 100, 110, 73, and 64 incidents in each of these months, respectively, from which we extract and identify 46, 52, 94, 60 and 54 domain names with verifiable prefix information and which are also represented in our RBL dataset. We will utilize this set, also referred to as the set of *incident prefixes* for verifying the prediction results. Some incidents have rather ambiguous description over its targets and some have domain names that are hard to pin down or have no records in our RBL database, in which case we simply discard these entries.

The majority of the incidents in our dataset are reported hacking events for websites of organizations, which further lead to leak of confidential information. Specific types of cyber attacks include website defacement and distributed Denial of Service (DDoS) attacks etc. We summarize this information in Table 2. In terms of region of origination, the vast majority of the reported incidents originated from the US; others cover fairly wide regions, including Canada, Europe, India, Peru, and Australia.

Note that among these cyber crimes, falling victim to a DDoS attack is not particularly correlated with the security quality of a network, i.e., there is nothing a network can do in terms of security practice to prevent itself from becoming the target of a DDoS attack. It thus may be supposed that including the DDoS incidents will not help build the predictor or may even hurt its performance. Accordingly, in the majority of our experiment reported in subsequent sections we tested both cases with including and excluding the DDoS incidents from this dataset; we did not observe significant difference in the results. For this reason in the results reported we have included the DDoS incidents.

3. FEATURE-BASED CHARACTERIZATION

3.1 Dynamics in the aggregate behavior

We start by looking into the dynamics of prefixes' aggregate signals. If these aggregate signals are to be used for predicting security incidents, as we set out to examine, the signals of the incident prefixes must in some way differ from those of the non-incident prefixes (referred to as the clean prefixes/set). The critical step is to determine whether there is a difference and how to capture this difference. In what follows we will first use entropy as a measure to directly compare aggregate signals from the two sets; the comparison turns out to be unsatisfactory as there isn't significant difference. We then turn to feature extraction and define several key features aimed at succinctly describing time series data. As we shall see this leads to significant observed difference between the two data sets.

3.2 An Entropy Comparison

Entropy is a common metric for assessing the dynamics embedded in a temporal signal, see e.g., studies on trending of social media, change in wind power [16]. The entropy of aggregate signal $\mathbf{r}_i$ of prefix i is defined as

$$H(\mathbf{r}_i) = -\sum_{j \in \Omega_i} p_j \log p_j \ , \qquad (1)$$

where Ω_i is the state space of $\mathbf{r}_i$, i.e., the possible number (or percentage) of IP addresses being listed as malicious. Note that this state space is finite as there is only a finite number of IP addresses in a prefix. p_j is the measured frequency of j in $\mathbf{r}_i$. In general a higher entropy indicates a more dynamic signal.

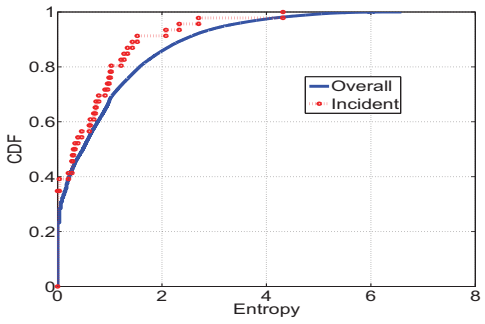


Figure 2: Cumulative distribution of entropy.

The distributions of entropy calculated over the entire set of 363, 667 prefixes and that calculated over the set of incident prefixes are plotted in Figure 2. We first observe that a large portion (30%) of prefixes in either set are extremely static in their malicious activities (with entropy close to 0). Interestingly, the incident set contains a slightly larger portion of low entropy prefixes, i.e., with more static signals. On the whole, however, these two distributions do not appear significantly different. This motivates us to seek better features to describe the aggregate signal.

3.3 Dynamic feature extraction

To strike a good balance between feature richness and complexity, we consider the following two-step approach. We start by value-quantizing the aggregate signal of a particular prefix into three regions: "good", "normal" and "bad", on a scale relative to that prefix's average magnitude. Specifically, the average magnitude of the aggregate signal is given by (for simplicity in the following we have suppressed the subscript i with the understanding that it applies to any prefix i): $r_{ave} = \frac{\sum_{t=1}^{T} r(t)}{T}$ with T being the time horizon under consideration. A point at t belongs to the "normal" region if $r(t) \in [(1-\delta)r_{ave}, (1+\delta)r_{ave}]$, the "good" region if $r(t) < (1-\delta)r_{ave}$, and the "bad" region if $r(t) > (1+\delta)r_{ave}$, where $0 < \delta < 1$ is a constant[1].

Our second step is to associate each region with three features: intensity, duration, and frequency, where intensity is the average magnitude of the aggregate signal within that region, duration is the average amount of time the signal remains in that region upon each entry (measured in days), and frequency is the rate (measured in number of times per day, a fraction since our sample rate is once per day) at which the aggregate signal enters that region. As there are three regions, each feature is a triple/vector, formally given as follows, with the indices 0, 1, and -1 denoting the normal, good and bad regions, respectively.

$$\text{intensity} \quad \lambda = [\lambda(0), \lambda(1), \lambda(-1)] \tag{2}$$

$$\text{duration} \quad \mathbf{d} = [d(0), d(1), d(-1)] \tag{3}$$

$$\text{frequency} \quad \mathbf{f} = [f(0), f(1), f(-1)] \ . \tag{4}$$

We now examine whether the feature values extracted from the two datasets, the general prefix set and the incident prefix set, exhibit statistically significant differences. The distribution comparison is shown in Figure 3. As each feature vector contains three values, each comparison consists of six cdf curves, three from the general dataset and three from the incident dataset.

Looking across all three features, we observe that for the incident data, regardless of the feature (with the exception of the un-normalized intensity), the values representing the three regions are much closer together. This is evidenced by the closer "bundling" of the three red curves in all but the second figures. By contrast, there is a much clearer separation between the values from the good region and the bad region within the general dataset, as evidenced by the larger distance between the "good" and "bad" (or "normal") blue curves in the figures. This observation is particularly prominent in the duration and frequency features. This means that a non-incident prefix likely has a much longer good duration than a bad or normal one, whereas for an incident prefix its good and bad durations are much more similar in lengths. Likewise, a non-incident prefix exhibits much higher frequency entering its good region than its bad or normal region, whereas an incident prefix shows these two as much closer in value.

More interestingly and perhaps a bit surprisingly, whereas the difference between the two sets of distributions is clearly seen in the duration and frequency features, it is much less so in the un-normalized intensity feature and not at all distinct in the normalized intensity feature – there is even a higher portion with smaller normalized intensity within the incident set in their respective bad regions. In other words, an incident prefix and a non-incident prefix tend to have similar % of their IPs listed on average. Compared to the duration and frequency features, this also suggests that the magnitude of maliciousness as a measure alone, which is what most studies focus on, see e.g., [4, 5], is insufficient in distinguishing between the two datasets. What appears to matter more is the persistence in and recurrence of a good or bad region captured by the duration and frequency features.

[1]Selecting the right value for such constants is in general non-trivial and typically done experimentally. We have used the value 0.2 in this study. This choice, while not insignificant in our method, is however not the sole determining factor of the subsequent performance, since we also separately assess the magnitude within each type of region.

4. FEATURE-BASED INCIDENT PREDICTION

The significant difference in the feature vector distribution shown in the previous section suggests that it could be utilized to distinguish one from the other; by doing so, we may be able to predict future incidents given the historical RBL data of a network entity. In this section we present a prediction method based on the above observation and the use of Support Vector Machine (SVM) [15]. We show that this method achieves reasonably good prediction results for a forecast window as small as two months.

4.1 Prediction methodology

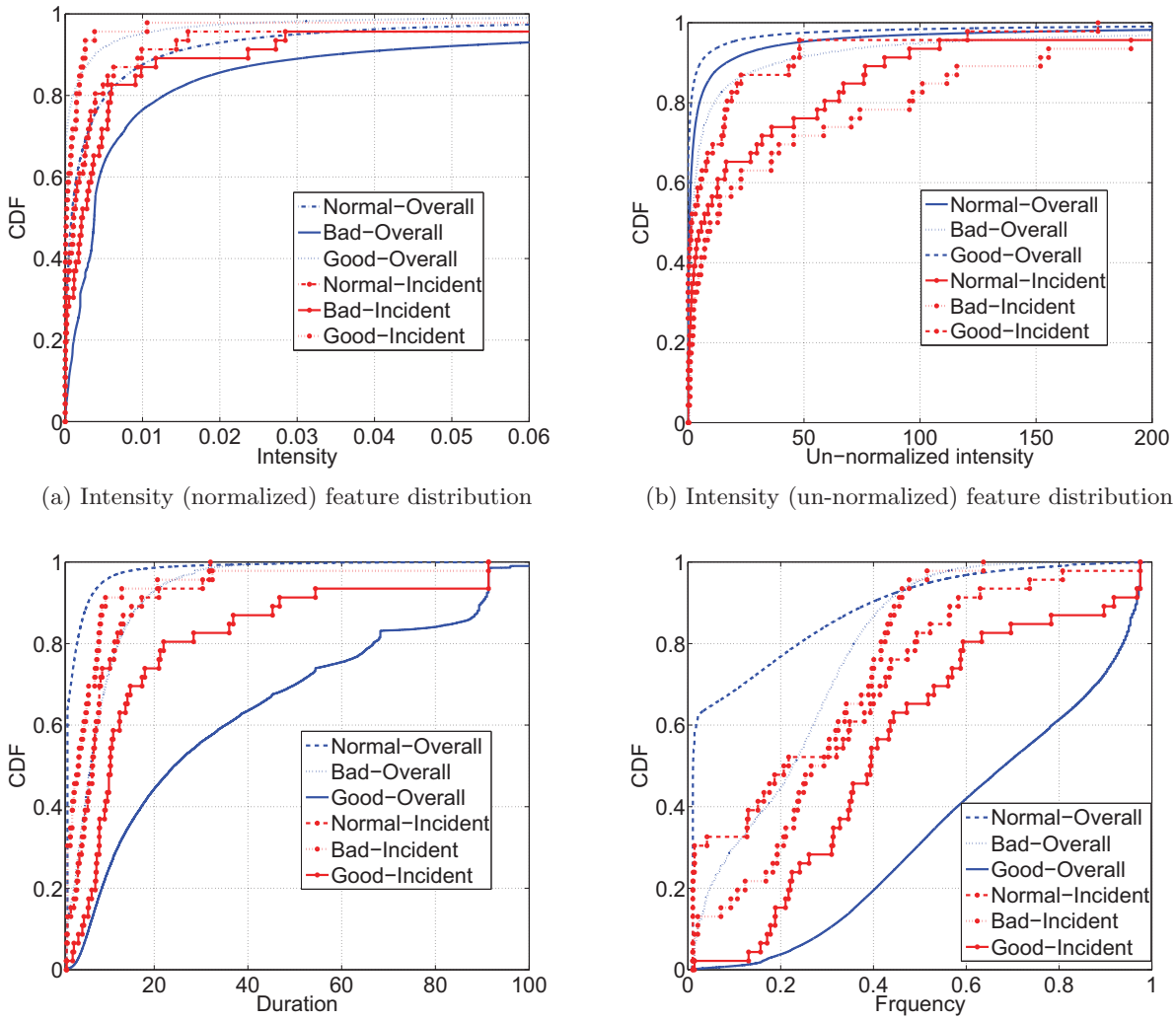
Following our observations in the previous section we will now focus on the duration and frequency features $\mathbf{d}, \mathbf{f}$ to build a predictor. This consists of a training step and a testing step. The training dataset consists of the following two sets of subjects.

1. The set of incident prefixes from the month of October 2013. This will be referred to as Group(1), or the incident group or incident prefixes.

2. A randomly selected set of prefixes ($< 1,000$) that are not associated with any incident in the month of October 2013. This will be referred to as Group(-1), or the clean group or clean prefixes. The random selection is due to the large size of our dataset. As mentioned, our overall RBL dataset includes the record of $363,667$ prefixes; compared to this, the size of verified incident dataset is very insignificant. Therefore simply combining the two sets of data would leads to the common problem of unbalanced training data in machine learning. The random selection thus serves as a subprocess to remedy this issues of imbalance. We will however repeat the random selection, each time training a new predictor/classifier.

For a prefix i belonging to either of the above two groups, we use its RBL data collected from January-September 2013 (using the union list) to calculate its feature values $\lambda_i, \mathbf{d}_i, \mathbf{f}_i$, i.e., using the data collected prior to the reported incident(s). Each prefix i also comes with a known label (or ground truth or group information): $t_i = 1$ if i belongs to the incident set and $t_i = -1$ if it belongs to the clean set. Collectively the above data constitutes the training dataset for each random trial, denoted as $\{\mathbf{x}_i, t_i\}_{i=1}^{N}$, where

$$\mathbf{x}_i = [\lambda_i, \mathbf{d}_i, \mathbf{f}_i] \ , \tag{5}$$



(a) Intensity (normalized) feature distribution

(b) Intensity (un-normalized) feature distribution

(c) Duration feature distribution

(d) Frequency feature distribution

Figure 3: CDF distribution of the feature vector values.

and N is the size of training set (the number of prefixes combining the two groups). In our experiments shown below we also show results obtained using other combination of the features, e.g., $\mathbf{x}_i = [\mathbf{d}_i, \mathbf{f}_i]$, $\mathbf{x}_i = [\lambda_i, \mathbf{f}_i]$, and so on. However for the intensity feature we will only use the un-normalized version following observations from the previous section that the normalized intensity appears to provide little useful information in separating the two sets.

4.2 Prediction result

Our test dataset again consists of two sets of subjects:

1. The set of incident prefixes identified for the months of November and December 2013.

2. A randomly selected 1,000 prefixes that are not associated with any incident in the months of November and December 2013. As in training, this set is repeatedly selected and the final result is the average over these random trials.

The resulting average true positives and false positives are calculated and plotted in Figure 4 under different combinations of the features used for training. As detailed earlier, this figure is generated by training on RBL data collected from Jan. to Sep. and incident data collected in Oct., and the prediction is for incidents occurring in Nov. and Dec.. Each point on a curve is the average over random selections of the clean set used for prediction. Different points on a curve correspond to different versions of the classifier generated by the random selections of the clean set used for training (recall for each random selection of the clean prefix set we obtain a different classifier). One of the best choices appears to be around (62%, 20%) for instance, on the duration-frequency feature curve. It is worth noting that the actual false positive rate at any operating point is likely to be lower than what is shown here as our incident dataset is by no means complete due to reasons such as under-reporting or delayed reporting of security incidents.

There is some difference between the prediction performance of classifiers trained using different combinations of

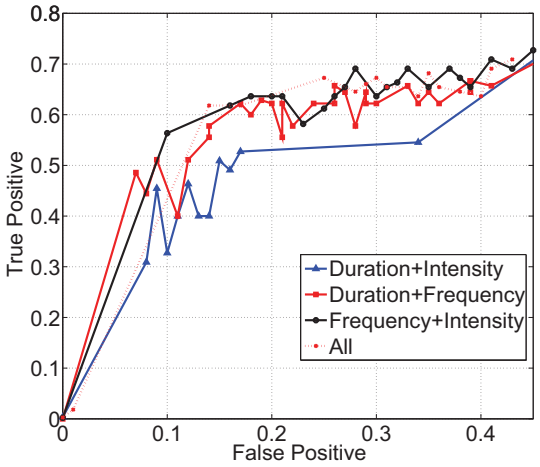


Figure 4: Incident prediction result. Training is based on the RBL data from Jan. to Sep. 2013 and incident data in Oct. 2013, using both clean and incident sets; prediction is for incidents in Nov. and Dec. 2013.

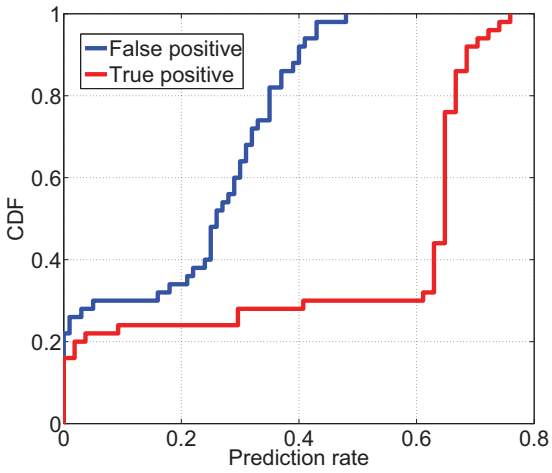


Figure 5: Distribution of the incident prediction performance of the classifiers. Each point corresponds to a random selection trace.

features. In particular, using only duration and intensity (un-normalized) appear to perform poorly, but the other combinations have very close performance. Using all features also appear to hold negligible advantage over these other combinations. This suggests that there is certain redundancy in what these features reveal. To avoid repetition, for the rest of this section we will only focus on the duration-frequency combination.

We see that different versions of the classifier present different combinations of true positive and false positive rates. The difference is attributed to the relationship between the randomly selected clean set and the incident set. For instance, if the clean set of prefixes happen to have a significant portion with feature values similar to those from the incident prefix set, then the resulting classifier will have high false positive, and by extension a relatively high true positive; the opposite is also true leading to combinations of low false positive but also low true positive. This is a consequence of the fact that the population of "clean" prefixes overwhelms the incident set, and that ours is a binary classifier that assigns two labels to a very wide range of behavioral patterns. To gain more insight into how likely each of these operating points is generated with a random selection of the clean prefixes, we plotted the distribution of the true and false positive rates in Figure 5 (using only the duration and frequency features). We see that the majority of these trials (random selection of clean training sets) have true positive larger than 60% (more than 70% of all trials) while the false positive mostly falls between 20% and 40% (around 65% of all trials). One advantage of this spectrum of possible operating regimes is that the random selection of the clean prefix set during the training phase can serve as a sub-process for finding the most desirable operation point depending on how risk-averse one prefers to be.

4.3 Forecast window size

We further examine the effect of the forecasting window size on the prediction accuracy. To do so we use the same set of classifiers trained using the union RBL list (from Jan-Sep'13) and the incident data from Oct'13, but perform tests for the three months (Nov'13-Jan'14) and four months (Nov'13-Feb'14), respectively. These results are added to the earlier curve and shown together in Figure 6. Clearly we see there is a significant performance improvement (with best operation point being around $(69\%, 20\%)$ or $(65\%, 15\%)$) when the forecasting window is extended to include Jan'14. As mentioned earlier, this is to be expected because the likelihood of something happening over a longer period of time is greater. However when we further extend to Feb'14 we see a decrease in performance. The main reason is by then the RBL dataset used for training (Jan-Sep'13) is somewhat outdated (temporal features become more and more outdated over time), and the classifier would need to be retrained (e.g., using RBL data from Mar'13 to Dec'13).

4.4 Specific incident case study

We now take a closer look at a recent major case of security breach that occurred at the University of Maryland in February 2014. Since the incident occurred recently, we extracted the duration and frequency features by using the RBL dataset over the period Mar-Dec'13, and trained the classifier using the January 2014 incident reports. The aggregated signal (normalized) for the Univ. of Maryland (prefix 128.8.0.0/16 which is used by its College Park campus) is shown in Figure 7. First again from Figure 7 we easily see a unclean network spanning the observation horizon.

We see an average short duration in the good region (3 days) while a high frequency into the bad region (0.29, or once every 3 days). These are negative signs of recurring malicious activities, despite its extremely low average (normalized) intensity. These are first pass evidence showing the network is likely to be maintained badly, despite its ex-

Samples	Reported time	Average	Duration	Frequency
U. Maryland	Feb-14	$8.7 \cdot 10^{-5}$	[3.0, 3.9, 2.9]	[0.30, 0.29, 0.40]

Table 3: Profiling statistics of selected samples : U. Maryland

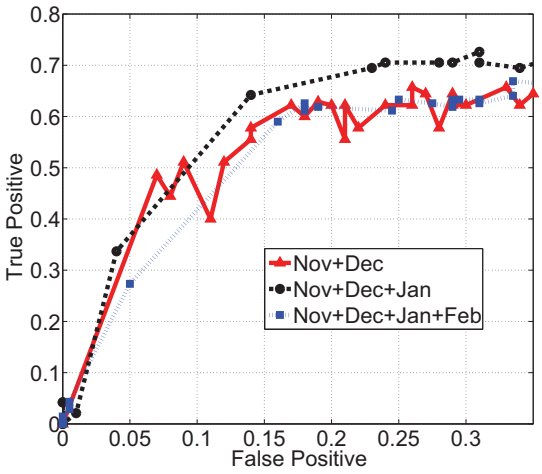


Figure 6: Incident prediction result over different forecast window sizes: two, three, and four months.

tremely low average intensity. According to our classifier, Univ. of Maryland again has been successfully indicated as being risky in the near future (Group 1).

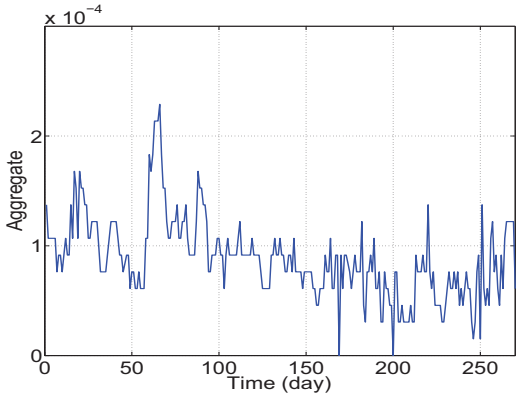


Figure 7: Univ. of Maryland's temporal evolution of maliciousness (Mar. - Dec. 2013). Prefix : 128.8.0.0/16.

5. ACKNOWLEDGEMENT

This work is partially supported by the NSF under grant CNS 1422211 and by the DHS under contract number HSHQDC -13-C-B0015.

6. CONCLUSION AND FUTURE WORK

In this paper we applied a simple machine learning technique SVM to a set of reputation blacklists (RBLs) to generate predictions for future security incidents that may happen to a network. The key to our approach is to aggregate the RBL data at the network (prefix) level and use a set of dynamic features to succinctly capture the dynamic behavior exhibited in the malicious activities. We showed that the resulting classifier is able to produce fairly accurate prediction results over a forecasting window of 2-3 months. These results have wide applications in for example providing risk evaluations, network management feedbacks etc. Further development includes building a higher-resolution classifier (as opposed to binary) that could generate prediction on the actual likelihood of an incident occurring, using more sophisticated machine learning techniques.

7. REFERENCES

[1] Barracuda Reputation Blocklist. http://www.barracudacentral.org/.
[2] Composite Blocking List. http://cbl.abuseat.org/.
[3] DShield. http://www.dshield.org/.
[4] Global Security Reports. http://globalsecuritymap.com/.
[5] Global Spamming Rank. http://www.spamrankings.net/.
[6] hpHosts for your pretection. http://hosts-file.net/.
[7] OpenBL. http://www.openbl.org/.
[8] PhishTank. http://www.phishtank.com/.
[9] SpamCop Blocking List. http://www.spamcop.net/.
[10] SURBL: URL REPUTATION DATA. http://www.surbl.org/.
[11] The SPAMHAUS project: SBL, XBL, PBL, ZEN Lists. http://www.spamhaus.org/.
[12] UCEPROTECTOR Network. http://www.uceprotect.net/.
[13] Web Hacking Incidence Reports. http://hackmageddon.com/.
[14] WPBL: Weighted Private Block List. http://www.wpbl.info/.
[15] Bishop, C. M., et al. *Pattern Recognition and Machine Learning*, vol. 1. springer New York.
[16] Deng, R., Yang, Z., Chen, J., and Chow, M.-Y. Load Scheduling With Price Uncertainty and Temporally-Coupled Constraints in Smart Grids.
[17] Xie, Y., Yu, F., Achan, K., Gillum, E., Goldszmidt, M., and Wobber, T. How Dynamic Are IP Addresses? In *Proceedings of SIGCOMM* (New York, NY, USA, 2007), ACM, pp. 301–312.