# Predictive Defense Against Evolving Adversaries

Richard Colbaugh
Sandia National Laboratories
Albuquerque, NM USA
colbaugh@comcast.net

Kristin Glass
New Mexico Institute of Mining and Technology
Socorro, NM USA
kglass@icasa.nmt.edu

***Abstract*—Adaptive adversaries are a primary concern in several domains, including cyber defense, border security, counterterrorism, and fraud prevention, and consequently there is great interest in developing defenses that maintain their effectiveness in the presence of evolving adversary strategies and tactics. This paper leverages the coevolutionary relationship between attackers and defenders to derive two new approaches to *predictive* defense, in which future attack techniques are anticipated and these insights are incorporated into defense designs. The first method combines game theory with machine learning to model and predict future adversary actions in the learner's "feature space"; these predictions form the basis for synthesizing robust defenses. The second approach to predictive defense involves extrapolating the evolution of defense configurations forward in time, in the space of defense parameterizations, as a way of generating defenses which work well against evolving threats. Case studies with a large cyber security dataset assembled for this investigation demonstrate that each method provides effective, scalable defense against current and future attacks, outperforming gold-standard techniques. Additionally, preliminary tests indicate that a simple variant of the proposed design methodology yields defenses which are difficult for adversaries to reverse-engineer.**

***Keywords*—-**predictive analytics, adversarial coevolution, machine learning, game theory, cyber security, security informatics.

## I. INTRODUCTION

Adaptive adversaries are a primary concern in many domains, including cyber defense, border security, counterterrorism, and crime prevention [e.g. 1-3]. For instance, emerging technologies and operational practices in these domains are increasingly moving toward highly interconnected architectures with small numbers of widely-shared protocols, thereby dramatically increasing the potential impact of even a single unanticipated attack. It is therefore essential that security professionals develop defenses which are able to respond rapidly to, or even foresee, evolving attack strategies and tactics.

Recognizing these trends and challenges, several researchers have recently proposed defenses which incorporate models of adversary behavior in order to increase defense system reliability and responsiveness against adaptive opponents; applications receiving attention include cyber defense [e.g. 4-7], border and transportation security [e.g. 8-10], and improvised explosive device defense [11,12]. However, while these model-informed methods represent an important advance over standard techniques, they continue to produce reactive defense designs and thus are limited in their ability to defend against new attacks.

Very recently, security researchers have begun working to develop *predictive* defenses, in which future attack strategies are explicitly anticipated and preemptively countered [13-16]. Despite this attention, much remains to be done to place the objective of predictive defense on a scientifically-grounded and practically-implementable foundation. Fundamental issues associated with the dynamics and predictability of coevolutionary "arms races" between attackers and defenders have yet to be resolved. For instance, although the work [13-15] has demonstrated that previous attacker actions and defender responses provide predictive information about future attacker behavior, little is known about which system characteristics have predictive power or how to employ these features to form useful predictions. Moreover, even in settings where these predictability and prediction issues have been resolved, it often remains an open question how to incorporate such predictive analytics into the design of practical real-world defense systems.

This paper leverages the coevolutionary relationship between attackers and defenders to derive two predictive defense algorithms which are effective against both current and future attacks strategies. We formulate the defense task as one of behavior classification, in which innocent and malicious activities are to be distinguished, and assume only limited historical information is available regarding prior attacker behavior or attack attributes. The first method combines game theory [17] with machine learning (ML) [18] to model and predict adversary actions in "feature space", that is, in the space of observable variables that the ML algorithm uses for learning; these predictions form the basis for synthesizing robust defenses. The second predictive defense strategy involves extrapolating the evolution of defense system configurations forward in time, in the space of defense parameterizations, as a way of generating defenses which work well against evolving threats. Interestingly, formulating the attack prediction/defense synthesis problem in an abstract space (of ML features or defense parameters) enables the development of algorithms that are scalable to applications of real-world size and complexity.

To permit the performance of these methods to be evaluated, we have assembled a large collection of non-Spam and Spam emails reflecting the evolution of Spammer tactics over an eight year period. Case studies with this dataset demonstrate that each of the proposed predictive methods provides robust, scalable defense, outperforming gold-standard Spam filters. Additionally, preliminary tests suggest that a simple "randomized feature" variant of the proposed design methodology generates defenses which are difficult for adversaries to reverse-engineer.

978-1-4673-2104-4/12/$31.00 ©2012 IEEE

## II. Predictive Defense Via Game-Based Learning

### A. Problem Formulation

As indicated in the Introduction, there is significant interest in developing *predictive* approaches to defending against adaptive adversaries, in which opponents' evolving strategies are anticipated and these insights are employed to counter new attacks. This section considers the following concrete instantiation of the predictive defense problem: given some history of attacker actions, design a defense system which performs well against both current and future attacks.

It is reasonable to expect that concepts and techniques from game theory might be helpful in understanding adversary coevolution, and indeed such approaches have been explored in a variety of domains [5,10,19]. These investigations have revealed several challenges to successfully using game-theoretic methods for predictive defense, and we mention two that have been particularly daunting. First, the space of possible attacker actions is typically very large in realistic environments, and because the complexity of most game models increases exponentially with the number of actions available to players, this has often made these models intractable in practice [19]. Second, it has proved difficult to derive models that capture evolving attacker behavior in any but the most idealized situations.

We overcome these two challenges by developing a game-based model for adversary adaptation within an ML framework, enabling effective defense in realistic settings. Crucially, the proposed approach seeks to derive the optimal defense for new attacks, rather than to predict these attacks perfectly, and therefore enjoys robust performance in the presence of (inevitable) prediction errors. We approach the task of countering adversarial behavior as an ML classification problem, in which the objective is to distinguish innocent and malicious activity. Each instance of activity is represented as a feature vector $x \in \Re^{|F|}$, where entry $x_i$ of x is the value of feature i for this instance and F is the set of instance features. In what follows, F is a set of "reduced" features, obtained by projecting measured feature vectors into a lower-dimensional space. While feature reduction is standard practice in ML [18], we show below that *aggressive* reduction allows us to efficiently manage the complexity of our game models. Behavior instances x belong to one of two classes: positive/malicious and negative/innocent (generalizing to more than two behavior classes is straightforward [18]). The goal is to learn a vector $w \in \Re^{|F|}$ such that classifier orient = $\text{sign}(w^T x)$ accurately estimates the class of behavior x, returning +1 (−1) for malicious (innocent) activity.

It is useful to assess the predictability of a phenomenon before attempting to predict its evolution; for example, such an analysis permits identification of measurables that possess predictive power [20]. There has been limited theoretical work assessing predictability of adversarial dynamics, but existing studies suggest attack-defend coevolution often generates predictable dynamics. For instance, although [21] finds that certain player strategies lead to chaos in a simple repeated game, [22] shows that large sets of player strategies and repeated games exhibit predictable adversarial dynamics. Here we supplement this theoretical work by conducting an empirical investigation of predictability, and select as our case study a cyber security problem – Spam filtering – which possesses attributes that are representative of many adversarial domains.

To conduct this investigation, we first obtained a large collection of emails from various publicly-available sources for the period 1999-2006, and added to this corpus a set of Spam emails acquired from B. Guenter's Spam trap for the same time period. Following standard practice, each email is modeled as a "bag of words" feature vector $x \in \Re^{|F|}$, where the entries of x are the frequencies with which the words in vocabulary F appear in the message. The resulting dataset consists of ~128,000 emails composed of more than 250,000 features. We extracted from this collection of Spam and non-Spam emails the set of messages sent during the 30 month period between January 2001 and July 2003 (email in other periods exhibit very similar evolutionary dynamics). Finally, the dimension of the email feature space was reduced via a singular value decomposition (SVD) analysis [18], yielding a reduction in feature space dimension of four orders of magnitude (from ~250K to 20).

We wish to examine, in a simple but meaningful way, the predictability of Spam adaptation, and propose two intuitively reasonable criteria with which to empirically evaluate predictability: *sensibility* and *regularity* (obviously more comprehensive, mathematically-rigorous frameworks can be derived for defining and assessing predictability [e.g.,20]). More specifically, and in the context of Spam, it would be *sensible* for Spammers to adapt their messages over time in such a way that Spam feature vectors $x_S$ come to resemble the feature vectors $x_{NS}$ of legitimate emails, and *regularity* in this adaptation might imply that the values of the individual elements of $x_S$ approach those of $x_{NS}$ in a fairly monotonic fashion.

To permit convenient examination of the evolution of feature vectors $x_S$ and $x_{NS}$ during the 30 month period under study, the emails were first binned by quarter. Next, the average values for each of the 20 (reduced) features was computed for all the Spam emails and all the non-Spam emails (separately) for each quarter. Figure 1 illustrates the feature space dynamics of Spam and non-Spam messages for one representative element (F1) of this reduced feature space. As seen in the plot, the value of feature F1 for Spam approaches the value of this feature for non-Spam, and this increasing similarity is a consequence of changes in the composition of Spam messages (the value of F1 for non-Spam emails is essentially constant). The dynamics of the other feature values (not shown) are analogous.

Observe that the Spam dynamics illustrated in Figure 1 reflect *sensible* adaptation on the part of Spammers: the features of Spam email messages evolve to appear more like those of non-Spam email, making Spam more difficult to detect. Additionally, this evolution is *regular*, with feature values for Spam approaching those for non-Spam in a nearly-monotonic fashion. Thus this empirical analysis indicates that coevolving Spammer-Spam filter dynamics possesses some degree of predictability, and that the features employed in Spam analysis may have predictive power; this result is in general agreement with the conclusions of the theoretical predictability analysis reported in [22]. Moreover, because many of the characteristics of Spam-Spam defense coevolution are shared by other adversarial systems, this result suggests these other systems may have exploitable levels of predictability as well.

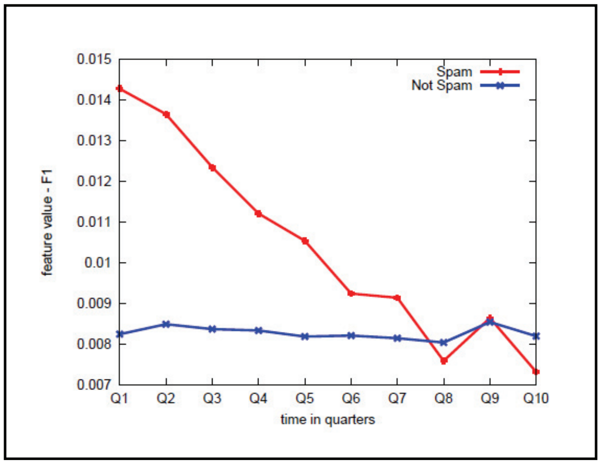

**Figure 1.** Spam/non-Spam evolution in feature space. The plot depicts evolution of feature F1 for Spam (red) and non-Spam (blue) over time (horizontal axis).

### B. *Predictive Defense Algorithm*

The proposed approach to designing a predictive defense system which works well against both current and future attacks is to combine ML with a simple game-based model for adversary behavior. In order to apply game-theoretic methods, it is necessary to overcome the complexity and model-realism challenges mentioned above. We address problem complexity by modeling adversary actions directly in an aggressively-reduced ML feature space, so that the (effective) space of possible adversary actions which must be considered is dramatically decreased. The difficulty of deriving realistic representations for attacker behavior is overcome by recognizing that the actions of attackers can be modeled as attempts to *transform* data (i.e., feature vectors x) in such a way that malicious and innocent activities are indistinguishable. (This is in contrast to trying to model the attack instances "from scratch"). It is possible to model attacker actions as transformations of data because, within an ML problem formulation, historical attack data are available in the form of training instances.

We model adversarial coevolution as a sequential game, in which the attacker and defender iteratively optimize the following objective function:

$$\min_{w} \max_{a} \left[ -\alpha\|a\|^3 + \beta\|w\|^3 + \sum_{i} \text{loss}\left(y_i, w^T(x_i + a)\right) \right] \qquad (1)$$

In (1), the loss function represents the misclassification rate for the defense system, where $\{y_i, x_i\}^n_{i=1}$ denotes pairs of currently-observed activity instances $x_i$ and their labels $y_i$ and w parameterizes the defense (recall the defense attempts to distinguish malicious and innocent activities using the classifier orient = sign($w^T x$)). The attacker attempts to circumvent the defense by transforming the data through vector $a \in \Re^{|F|}$, and the defender's goal is to optimally counter this attack through specification of the appropriate classifier vector $w \in \Re^{|F|}$. The terms $-\alpha||a||^3$ and $\beta||w||^3$ define "regularizations" imposed on attacker and defender actions, respectively, as discussed below.

Observe that (1) models the attacker as acting to increase the misclassification rate with vector a, subject to the need to limit the magnitude of this vector (large a is penalized via the term $-\alpha||a||^3$). This model thus captures in a simple way the fact that the actions of the attacker are in reality always constrained by the goals of the attack. For instance, in the case of Spam, the Spammer tries to manipulate message x in such a way that it "looks" enough like legitimate email to evade the Spam filter. However, the transformed message x+a must still communicate the desired information to the recipient or the attacker's goal will not be realized, and so the transformation vector a cannot be chosen arbitrarily.

The defender attempts to reduce the misclassification rate with an optimal choice for vector w, and avoids "over-fitting" through regularization with the $\beta||w||^3$ term [18]. Notice that the formulation (1) permits the attacker's goal to be modeled as counter to, but not exactly the opposite of, the defender's goal, and this is consistent with many real-world settings. Returning to the Spam example, the Spammer's objective of delivering messages which induce profitable user responses is not the inverse of an email service provider's goal of achieving high Spam recognition with a very low false-positive rate.

The preceding development can be summarized by stating the following predictive defense (PD) algorithm:

**Algorithm PD**

1. Collect historical data $\{y_i, x_i\}^n_{i=1}$ which reflects past behavior of the attacker and past legitimate behavior.
2. Optimize objective function (1) to obtain the predicted actions a* of the attacker and the optimal defense w* to counter this attack.
3. Estimate the status of any new activity x as either malicious (+1) or innocent (−1) via orient = sign($x^T w^*$).

Observe that Step 2 of this algorithm can be interpreted as first predicting the attacker strategy through computation of attack vector a*, and then learning an appropriate countermeasure w* by applying ML to the "transformed" data $\{y_i, x_i+a^*\}^n_{i=1}$.

### C. *Algorithm Evaluation*

This case study examines the performance of Algorithm PD for the Spam filtering problem. We use the Spam/non-Spam email dataset introduced above, consisting of ~128,000 messages that were sent during the period 1999-2006. The study compares the effectiveness of Algorithm PD, implemented as a Spam filter, with that of a well-tuned naïve Bayes (NB) Spam filter [15]. Because NB filters are widely used and work very well in Spam applications, this filter is referred to as the gold-standard algorithm. We extract from our dataset the 1000 oldest legitimate emails and 1000 oldest Spam messages for use in training both Algorithm PD and the gold-standard algorithm. The email messages sent during the four year period immediately following the date of the last training email are used as test data. More specifically, these emails are binned by quarter and then randomly sub-sampled to create balanced datasets of Spam and legitimate emails for each of the 16 quarters in the test period.

Recall that Algorithm PD employs aggressive feature space dimension reduction to manage the complexity of the game-based modeling process. This dimension reduction is accomplished here through SVD analysis, which reduces the dimension |F| of feature vectors from ~250K to 20) [18]. (The orthogonal basis used for this reduction is derived by performing SVD analysis using the 1000 non-Spam and 1000 Spam training emails.) We have found that good classification accuracy can be obtained with a wide range of (reduced) feature space dimensions. For example, we achieve a filtering accuracy of ~97% with the training data when using an NB classifier implemented with feature dimension ranging from |F|=100,000 to |F|=5.

The gold-standard strategy is applied as described in [15]. Algorithm PD is implemented with parameter values $\alpha = 0.001$ and $\beta = 0.1$, and with a sum-of-squares loss function. To evaluate the utility of these defenses against evolving adversaries, we train Algorithm PD and the gold-standard algorithm *once*, using the (oldest) 1000 non-Spam/1000 Spam dataset, and then apply the filters without retraining to the four years of emails that follow these 2000 messages.

Sample results from this study are depicted in Figure 2. Each data point in the plots represents the average accuracy over ten trials (two-fold cross-validation). It can be seen that the Spam filter based upon Algorithm PD significantly outperforms the gold-standard method: the predictive defense experiences almost no degradation in accuracy over the four years of the study, while the gold-standard method suffers a substantial drop in accuracy during this period. These results suggest that combining ML with simple game-based models offers an effective means of defending against adaptive adversaries .

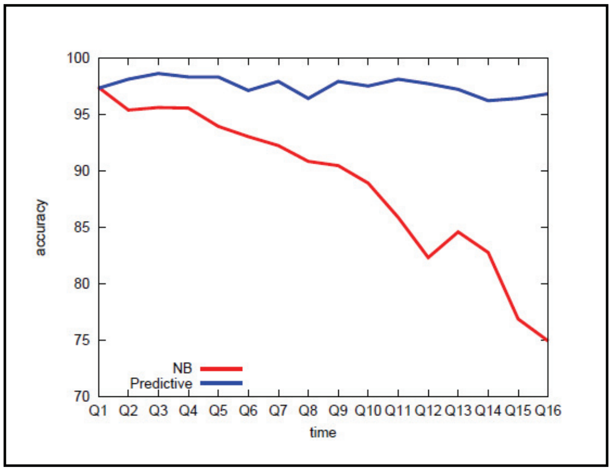

**Figure 2.** Results for the predictive defense case study. The plot shows how Spam filter accuracy (vertical axis) varies with time (horizontal axis) for the gold-standard NB filter (red) and Algorithm PD filter (blue).

### *D. Randomized Feature Learning*

An important consideration when applying ML techniques in adversarial settings is the extent to which adversaries can reverse-engineer the learning algorithm and use this knowledge to circumvent the classifier [3]. One way to increase the difficulty of the adversary's reverse-engineering task is to employ "randomized feature" learning [23]. Here we explore a very simple three-step implementation of this idea: 1.) divide the set of available features into randomly-selected, possibly overlapping subsets; 2.) train one classifier for each subset of features; and 3.) alternate between classifiers in a random fashion during operation. The fact that good classifier performance is often obtainable with only a few features (see the Spam example above) suggests the feasibility of employing multiple small subsets of randomly-selected features in a suite of classifiers.

To test the effectiveness of this strategy, we use a variant of the optimization process specified in (1). More specifically, we first use training data $\{y_i, x_i\}^n{}_{i=1}$ to computed the classifier vector w in two ways: 1.) using the full set of (reduced-dimension) features F, 2.) using two subsets of features randomly selected from set F; the resulting classifier vectors are denoted $w_F$ and $\{w_{F1}, w_{F2}\}$. (1) is then employed to compute the optimal attack against classifier vector $w_F$, denoted $a_F$, and to compute the optimal attack against the defense consisting of randomly alternating classifiers $w_{F1}$ and $w_{F2}$, denoted $a_{F12}$.

Applying this evaluation process to the 2000 email training dataset described in Section IIC suggests that randomized feature leaning may be an effective way to reduce the efficacy of adversary reverse-engineering methods. We define F to be the set of 20 features with largest singular values (in the SVD reduction process), and build sets F1 and F2 by randomly sampling F (with replacement) until each subset contains 10 features. The classification accuracy of $w_F$ against *nominal* data (i.e., with a=0) is superior to that provided by a classifier which randomly alternates between classifiers $w_{F1}$ and $w_{F2}$, but the difference is modest – the respective accuracies are 98.4% and 96.2% (two-fold cross-validation). Crucially, however, the randomized feature classifier is substantially more robust against *attack* data (i.e., data corresponding to $a=a_F$ or $a=a_{F12}$). Indeed, the accuracy of classifier $w_F$ is only 66.1% against attack data, while the accuracy of filter $\{w_{F1},w_{F2}\}$ is 86.8%, in the attack setting (two-fold cross-validation, see Figure 3).

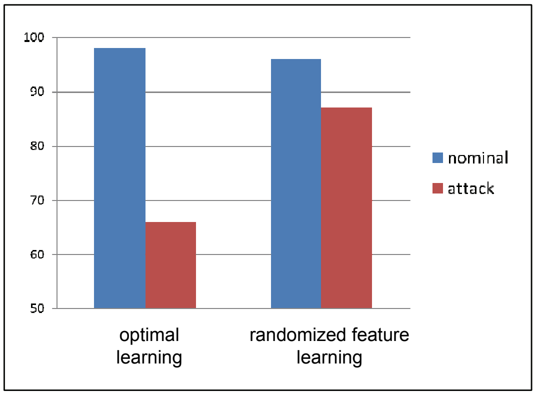

**Figure 3.** Results for randomized feature learning case study. Bar chart shows Spam/non-Spam filter accuracy for classifiers $w_F$ (left bars) and $\{w_{F1}, w_{F2}\}$ (right bars) for nominal data (blue) and "attack" data (red).

## III. Predictive Defense Via Extrapolative Learning

### A. *Problem Formulation*

The previous section derives a predictive defense system in the "feature space" of observable variables that characterize adversary activity. In this section we adopt a complementary perspective, proposing a simple technique for developing proactive defenses in "defense space", that is, in the space of defense system parameterizations. The specific problem of interest may be stated as follows: given a (possibly limited) history of defense system configurations, design a new defense which performs well against both current and future attacks.

As noted above, it is useful to examine the predictability of a phenomenon of interest before attempting to predict its evolution [20]. Here we conduct an empirical investigation of the predictability of defense system dynamics through a case study which employs the same Spam/non-Spam email dataset introduced in Section II. The present study focuses on those messages sent during the three year period 2001-2004 (other periods exhibit very similar behavior). We assess defense system predictability in terms of the *sensibility* and *regularity* of the observed dynamics. More specifically, and in the context of Spam defense, it is *sensible* for a Spam filter to adapt to compensate for the way Spammers modify their messages over time, and in a *regular* adaptation the values of defense system parameters might change approximately monotonically.

To examine the dynamics of Spam filter configurations associated with our dataset, we first binned the messages by quarter and performed aggressive feature-space dimension reduction via SVD analysis, retaining the five features with largest singular values. Next, separate NB filters were trained for each quarter, and the filter weights {w1, w2, w3, w4, w5} corresponding to features F1-F5 were recorded. Figure 4 depicts the values of the NB filter weights for quarters 1, 5, 9, and 13 (filter weights for the other quarters are consistent with those shown in the plot and are suppressed for clarity).

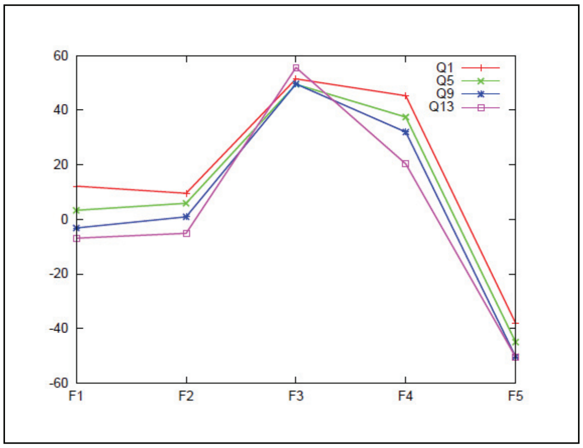

**Figure 4.** Spam filter evolution in defense space. The plot depicts values of the Spam filter weights corresponding to features F1-F5 for the first quarters of 2001 (red), 2002 (green), 2003 (blue), and 2004 (magenta).

Inspecting the evolution of filter weights depicted in Figure 4 reveals that defense adaptation is sensible. For example, by comparing Figures 1 and 4 it is seen that, as feature F1 evolves to become less predictive of Spam (Figure 1), the Spam filter places less emphasis on this feature (Figure 4); similar behavior is observed for the other weights. Additionally, the dynamics of the feature weights is regular, with most of the weights exhibiting monotonic adaptation. Thus the empirical analysis indicates that Spam filter dynamics possesses some degree of predictability, and that filter parameters may have predictive power. These results suggest the possibility that defenses in other domains may have exploitable levels of predictability as well.

### B. *Extrapolative Defense Algorithm*

The proposed approach to designing a predictive classifier that works well against both current and future attacks is to simply extrapolate the sequence of observed defense systems forward in time. Note that this strategy is motivated by the results of the empirical predictability analysis summarized above. Sequences of defense system parameterizations can often be obtained directly, for example from the system "owners". Alternatively, if historical attack data are available, these data can be used to learn associated defense sequences (as illustrated above).

There are many ways to extrapolate a given sequence of defense system parameterizations $\{w_1, w_2, \ldots, w_p\}$ into the future, and thereby generate predictions for useful future defenses. We adopt the following linear strategy:

$$w_{p+T} = \Sigma^{p}_{i=1} \beta_i w_i \qquad (2)$$

where the $w_i$ and $\beta_i$ are defense parameterizations and extrapolation coefficients, respectively, and T is the time horizon for which a prediction is desired. The coefficients $\beta_i$ are ordinarily specified so that $|\beta_i| \geq |\beta_j|$ if $i > j$, so more recent observations are emphasized. Appropriate values for the $\beta_i$ can be estimated in various ways, including statistical inference from historical data [18] or consultation with domain experts [15].

The preceding discussion can be summarize by sketching an algorithm for predicting a classifier vector $w_{p+T}$ which may be expected to be useful at future time $t=p+T$:

**Algorithm ED (Extrapolative Defense)**

1. Collect a sequence of defense system parameterizations $\{w_1, w_2, \ldots, w_p\}$ (e.g., from historical defense data or by learning appropriate defenses from historical attack data).
2. Estimate the coefficients $\beta_i$ in (2) (e.g., using ML).
3. Compute classifier vector $w_{p+T}$ from (2), and estimate the status of any new activity as either malicious (+1) or innocent (−1) via orient = $\text{sign}(x^T w_{p+T})$.

### C. *Algorithm Evaluation*

This case study examines the performance of Algorithm ED for the Spam filtering problem. We use the Spam/non-Spam email dataset described above, consisting of all emails sent during the 54 month period from early 2001 to mid-2005. The study compares the effectiveness of Algorithm ED, implemented as a Spam filter, with that of two NB Spam filters trained in different ways. As in the previous case studies, we first binned the emails by quarter, and then randomly sampled each quarter to create balanced datasets for all 18 quarters in the study period.

To provide a demanding test, we extracted from our dataset the emails sent during quarters Q1, Q5, and Q9 for use in training Algorithm ED. This procedure is intended to reflect the common situation in which opportunities for observation may arise only sporadically. The messages sent during the 18 month period from quarters Q13 to Q18 serve as test data. (The quarters closest to the training period, Q10 through Q12, are not included in the test set to increase the difficulty of the task.)

Algorithm ED is implemented by first training NB filters on data from quarters Q1, Q5, and Q9, yielding defense parameterizations $\{w_1, w_5, w_9\}$, and then using (2) to extrapolate these defenses. More specifically, we compute predicted defense $w^*$ using (2) with $\beta_1 = 0$, $\beta_5 = -1$, and $\beta_9 = 2$ (a simple Euler-like extrapolation). The first NB filter used for comparison employs $w_9$, that is, the filter derived from the most recent training data. The second NB filter examined in this case study is permitted to use "future" data during training: when attempting to distinguish Spam and non-Spam emails in quarter Qm, for $m \in \{13, 14, \ldots, 18\}$, this filter is trained on Qm data. Because the latter NB filter has access to future data, which is unavailable to the other defense systems, the performance of this filter is expected to be an upper bound for that of a predictive filter, and we refer to this NB filter as the gold-standard. All three filters – Algorithm ED, nominal NB, and gold-standard – are applied using an aggressively-reduced feature space of dimension $|F|=5$.

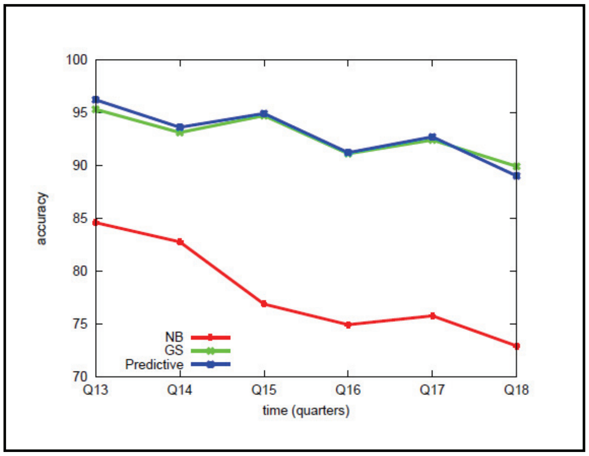

**Figure 5.** Results for the extrapolative defense case study. The plot shows how Spam filter accuracy (vertical axis) varies with time (horizontal axis) for the nominal NB filter, (red), gold-standard NB filter (green), and Algorithm ED filter (blue).

Sample results from this study are shown in Figure 5. Each data point in the plots represents the average accuracy over ten trials (two-fold cross-validation). It is seen that the filter based upon Algorithm ED significantly outperforms the nominal NB method. Moreover, the accuracy of Algorithm ED is comparable to that achieved by the gold-standard NB method, despite the fact that the latter filter is trained on "future" data not available to Algorithm ED. These results suggest that simple defense system extrapolation offers an effective means of defending against evolving adversary behavior.

## Acknowledgements

This work was supported by the Laboratory Directed Research and Development Program at Sandia National Laboratories. We thank Chip Willard and Curtis Johnson for numerous helpful discussions on aspects of this research.

## References

[1] *Proc. 2010 IEEE ISI*, Vancouver, BC Canada, May 2010.

[2] *Proc. 2011 IEEE ISI*, Beijing, China, July 2011.

[3] "Machine learning in adversarial environments", P. Laskov, R. Lippmann, Eds, Special Issue, *Machine Learning,* Vol. 81, 2010.

[4] Zhang, Q., D. Man, and W. Yang, "Using HMM for intent recognition in cyber security situational awareness", *Proc. IEEE KAM*, Wuhan, China, November 2009.

[5] Parameswaran, M., H. Rui, and S. Sayin, "A game theoretic model and empirical analysis of Spammer strategies", *Proc. CEAS 2010*, Redmond, WA, July 2010.

[6] Ahmadinejad, S., S. Jalili, and M. Abadi, "A hybrid model for correlating alerts of known and unknown attack scenarios and updating attack graphs", *Computer Networks*, Vol. 55, pp. 2221-2240, 2011.

[7] Zakrzewska, A. and E. Ferragut, "Modeling cyber conflicts using an extended Petri Net formalism", *Proc. IEEE CICS*, Paris, France, April 2011.

[8] Kaza, S., Y. Wang, and H. Chen, "Enhancing border security: Mutual information analysis to identify suspect vehicles", *Decision Support Systems*, Vol. 43, pp. 199-210, 2007.

[9] Gkonis, K. and H. Psaraftis, "Container transportation as an interdependent security problem", *J. Transportation Security*, Vol. 3, pp. 197-211, 2010.

[10] Pita, J. et al., "GUARDS: Game theoretic security allocation on a national scale", *Proc. AAMAS '11*, Taipei, Taiwan, May 2011.

[11] Williams, E., *Surveillance and Interdiction Models: A Game Theoretic Approach to Defend Against VBIED*, Thesis, Naval Postgraduate School, June 2010.

[12] Smith, A., "Improvised explosive devices in Iraq, 2003-09", *The Letort Papers*, US Army War College, April 2011.

[13] Colbaugh, R., "Does coevolution in malware adaptation enable predictive analysis?", *IFA Workshop: Exploring Malware Adaptation Patterns*, San Francisco, CA, May 2010.

[14] Bozorgi, M., L. Saul, S. Savage, and G. Voelker, "Beyond heuristics: Learning to classify vulnerabilities and predict exploits", *Proc. ACM KDD '10*, Washington DC, July 2010.

[15] Colbaugh, R. and K. Glass, "Proactive defense for evolving cyber threats", *Proc. 2011 IEEE ISI*, Beijing, China, July 2011.

[16] Cipriano, C. et al., "NEXAT: History-based approach to predict attacker actions", *Proc. ACSAC*, Orlando, FL, December 2011.

[17] Peters, H., *Game Theory*, Springer, Berlin, 2008.

[18] Hastie, T., R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Second Edition, Springer, New York, 2009.

[19] Dalvi, N. et al., "Adversarial classification", *Proc. ACM KDD '09*, Seattle, WA, August 2004.

[20] Colbaugh, R. and K. Glass, "Predictive analysis for social processes I: Multi-scale hybrid system modeling, and II: Predictability and warning analysis", *Proc. 2009 IEEE MSC*, Saint Petersburg, Russia, July 2009.

[21] Sato, Y., E. Akiyama, and J.D. Farmer, "Chaos in learning a simple two-person game", *Proc. National Academy of Sciences USA,* Vol. 99, pp. 4748-4751, 2002.

[22] Colbaugh, R., "Arctic ice, George Clooney, lipstick on a pig, and insomniac fruit flies: Combining kd and m&s for predictive analysis", *Proc. ACM KDD '11*, San Diego, CA, August 2011.

[23] Johnson, C., Personal communication, December 2011.