# An Artificial Immune System for Phishing Detection

Xing Fang, Nicholas Koceja, Justin Zhan, Gerry Dozier
Department of Computer Science
North Carolina A&T State University
Greensboro, NC 27411
xfang,nfkoceja,zzhan,gvdozier@ncat.edu

Dasgupta Dipankar
Department of Computer Science
University of Memphis
Memphis, TN 38152
dasgupta@memphis.edu

***Abstract*—The amazing nature of biological immune systems on protecting humans from pathogens inspired people to develop artificial immune systems. Designed to simulate the functionalities of biological immune systems, artificial immune systems are suggested to be mainly applied in the domain of computer security. In this paper, we propose an artificial immune system for phishing detection. The system is to detect phishing emails through memory detectors and mature detectors. The memory detectors are generated from the training data set, which, in turn, contains the phishing emails previously seen by the system. The immature detectors are reproduced through the system's mutation process. To the best of our knowledge, this is the first time such a system is ever proposed. We believe that the system is more adaptive than any other existing phishing detection techniques.**

***Keywords-artificial immune system; phishing detection; memory detector; immature detector***

## I. INTRODUCTION

The nature of biological immune systems prevents humans from being infected by pathogens that cause diseases. Without a functional immune system, even minor infections can be fatal. A typical immune system provides two kinds of defensing functions, namely innate immunity and adaptive immunity. The innate immunity is determined by the genes that a person inherits from her parents, whereas the adaptive immunity is adaptively formed when some pathogens cannot be effectively destroyed by the system's innate immunity [14].

Not only do immune systems protect people against diseases, but the systems' capability of massively-parallel adaptive information processing has become a subject of great research interest for many years [7, 8]. Specifically, an immune system has many properties that are valuable to incorporate into artificial systems because it is diverse, distributed, error tolerant, dynamic, self-monitoring and adaptable [7]. Under such observation, Hofmeyr and Forrest introduced a detailed architecture for an Artificial Immune System (AIS) [8]; the system was designed to simulate the functionalities of biological immune systems and to incorporate the properties. They suggested that the application domain of the AIS is mainly computer security (host-based intrusion detections, computer virus detections, etc.). Contending that the problem of anomalous behavior characterizing in a computer network environment can be very complex, Dasgupta and González [3] proposed an AIS-based technique to improve the situation by only maintaining a normal profile built up from positive data. Apart from computer security, AISs are also being applied in many other areas. Andrews and Timmis [1] integrated AIS's diversity into aiNet, a previously introduced algorithm for data clustering, to improve the algorithm's searching capability. Rahman et al. [16] developed an AIS based technique to optimize economic dispatch, which is the procedure of ensuring economic operation of a power system.

Phishing, by definition, is the way of deceiving individuals into disclosing sensitive personal information through deceptive computer-based means [15]. One of the most common forms of phishing is via email. Unlike spam (usually conveys unsolicited information), phishing emails can be potentially more harmful due to deliberately designed phishing schemes. Once recipients are hooked, their sensitive personal information will be suffering from leakage.

In this paper, we propose an artificial immune system for phishing detection. The contribution of the paper is as the following:1) To the best of our knowledge, this is the first time that an AIS-based phishing detection system is ever proposed; 2) The capability of the system to generate diverse mutated detectors improves the effectiveness of phishing detection; 3) Compared to any other existing phishing or spam detecting method, our system is more adaptive to the changing phishing patterns; 4) The system offers flexibility on system configurations in order to suffice different user requirements.

The remainder of the paper is organized as follows: We briefly review some related work on existing artificial immune systems and phishing detecting methods in section 2; in section 3, we propose the entire system design; in section 4, we present our detailed

experiment process and results of the AIS performance evaluations; we form a discussion towards our system and present some future work in section 5; we conclude our paper in section 6.

## II. Related Work

In this section, we propose a brief review of some related work including existing artificial immune systems and spam/phishing detection techniques.

### A. Artificial Immune Systems

Just as human beings are vulnerable to pathogens, computers are subject to infections from computer viruses. For the sake of self-protection, biological immune systems prevent people from being infected by those pathogens. This biological truth inspired computer scientists; a design for a computer immune system was in demand. Kephart [11] first addressed this problem by providing an artificial immune system design for computers. The system has the capability of self-evolving; whenever it encounters an unknown virus, the system will be able to recognize the virus' pattern and save the pattern to its virus database for future detections.

Hofmeyr and Forrest [7, 8] pushed AIS design one step further by taking negative selection [5] into account. The whole idea was based upon the simulation of immature T-lymphocytes' lifecycle: As long as an immature detector (T-lymphocyte) matches any self-pattern (benign data flow), that detector will be eliminated. Figure 1 depicts the process of negative selection.

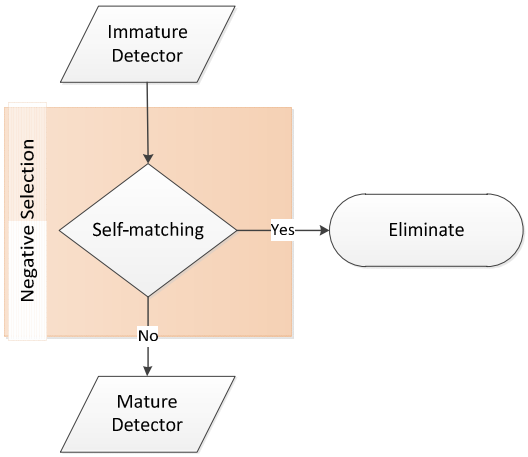


Figure 1: The process of Negative Selection

Those detectors that successfully passed the negative selection process automatically became mature detectors for non-self (anomalous data flow)-matching. A similar selecting process is adopted for non-self-matching: If a mature detector is able to match any non-self-pattern, the detector will be saved as a memory detector for further usage; otherwise, the detector will be deleted.

In a nutshell, an artificial immune system consists of a collection of memory detectors. The evolution of the system includes the generation and the maturation of immature detectors, which are formed and used to detect previously unknown intrusions. If we allow the AIS to repeatedly evolve, the system's immunity will keep being enhanced because of the increase in the memory detectors' numbers.

### B. Spam and Phishing Detection

Spam, by definition, is a type of unsolicited email usually of a commercial nature that is indiscriminately sent to multiple mailing lists, individuals, or newsgroups [17, 18]. Other names used to describe spam include junk e-mail, unsolicited commercial email, unsolicited bulk email, and unsolicited automated email. The negative impact brought by spam is somewhat annoying: Users may frequently receive spam, which will quickly take up their inbox if the spam were not deleted in time; once the storage room was depleted, legitimate emails can no longer reach users' inbox. As was mentioned before, phishing can be more harmful than spam. The deliberately conceived phishing schemes are commonly targeting people's sensitive information, such as name, date of birth, social security number, bank account number, password, or credit/debit card number. The losses due to phishing attacks are also ranging from money loss to identity theft. Before proceeding, we define the following email categories:

$$Email \begin{cases} Legitimate(Ham) \\ Spam \\ Phishing \end{cases}$$

#### 1) Naïve Bayesian Classifier for Spam Detection

One straightforward way to detect spam is to simply check if a current incoming email would be classified as spam. For instance, if the email contains many words or characters that appear frequently in previous spam the email is then classified as spam. According to Graham [6, 19], every single word or meaningful character within an email is treated as a token. The spam filer itself is constructed based on the Naïve Bayesian Classifier [19], such that each token is able to be tagged with two probability scores, the probability of a spam token and the probability of a ham token. Finally, the filter classifies the email as spam or ham via some weighted key-token values. This statistical approach opens a door for constructing Bayesian filters.

#### 2) The Stochastic Learning Weak Estimator

The stochastic learning weak estimator, also referred to as a weak estimator, was originally introduced by Oommen [13] for non-stationary environment estimations. It was later expanded by Zhan and Thomas [20] by using a weak estimator for phishing detection. Their weak estimator approach is to integrate the weak

estimator, λ, with probabilities of phishing tokens computed via Naïve Bayesian method. Specifically, given a weak estimator λ and a certain token's phishing probability, $P(t = ph)$, for any given document, $d$, they have:

$$P(t = ph) = \begin{cases} \lambda P(t = ph), if\ t\ is\ in\ d \\ 1 - \lambda P(t = ph), if\ t\ is\ not\ in\ d \end{cases}$$

## III. PHISHING DETECTION WITH THE AIS

In this section, we present our artificial immune system for phishing detection. Specifically, it includes phishing email analysis, detector generation/mutation, and general system design.

### A. Phishing Email Analysis

Information that can be extracted from a single phishing email includes the following four aspects: sender's (phisher's) email address; subject and body; attachment(s); link(s). An observation, in a phishing scenario, towards these four aspects is given as the following:

#### 1) The email address

To deceive recipients, phishing emails have to appear as legitimate ones. To achieve this, phishers always make up their email addresses as the ones from certain authorities. However, email addresses ending with .edu, .gov, or .org are not likely to be phishers' email addresses compared to the ones ending with .com.

#### 2) The subject and body

The information extracted from an email's subject and body (the main portion) is text information. In reference to phishing email, the text information contains deliberately conceived phishing schemes via human languages. Similar to Graham's Law, we have the following equation to examine every single token:

$$P(token = phishing\ token) = \frac{AP/NP}{AP/NP + AN/NN}$$

And each of the equation is defined as:

$P(token = phishing)$: The probability that the token is a phishing token.

$AP$: The number of appearances of the token in phishing emails.

$NP$: Total number of phishing emails collected.

$AN$: The number of appearances of the token in non-phishing emails.

$NN$: Total number of non-phishing emails collected.

The equation grants every token a probability score varying from 0 to 1. Based on Graham's Law, if a token's probability score is more than 0.5 (50 percent), the token will be considered as a phishing token; otherwise, it is a non-phishing token.

#### 3) The attachment and link

Compared to non-phishing emails, phishing emails usually do not have any attachments, but usually have at least one link.

### B. Detector Generation and Mutation

We formally define the format of detectors in our artificial immune system:

$$Detector_i \stackrel{\text{def}}{=} \{addr_i, btb_i, link_i\}$$

The artificial immune system is a set of detectors, where the i[th] detector, denoted as $Detector_i$, is a collection of factors extracted from the i[th] phishing email in our corpus. The factors are: the phisher's email address ($addr_i$), the bad (phishing) token bag ($btb_i$), and the number of links ($link_i$) that the i[th] email has.

The detectors formed based upon actual phishing emails are memory detectors. In a biological immune system, memory detectors only provide innate immunity that is inherited from the host's parents. Just like the innate immunity is not strong enough to protect the host from various pathogens, the memory detectors themselves are insufficient to detect various phishing emails. Therefore, adaptive immunity is a requirement for the AIS. The biological process for adaptive immunity generation is called mutation. For the sake of consistency, we also name our process of reproducing mutated detectors as mutation:

$$Detector_i \stackrel{mutate}{\Longrightarrow} Detector_i' = \{addr_i', btb_i', link_i'\}$$

The mutation process reproduces new factors that are used to replace the old ones in a memory detector. A detector is mutated at least one of its factors has been replaced.

### C. The AIS Design

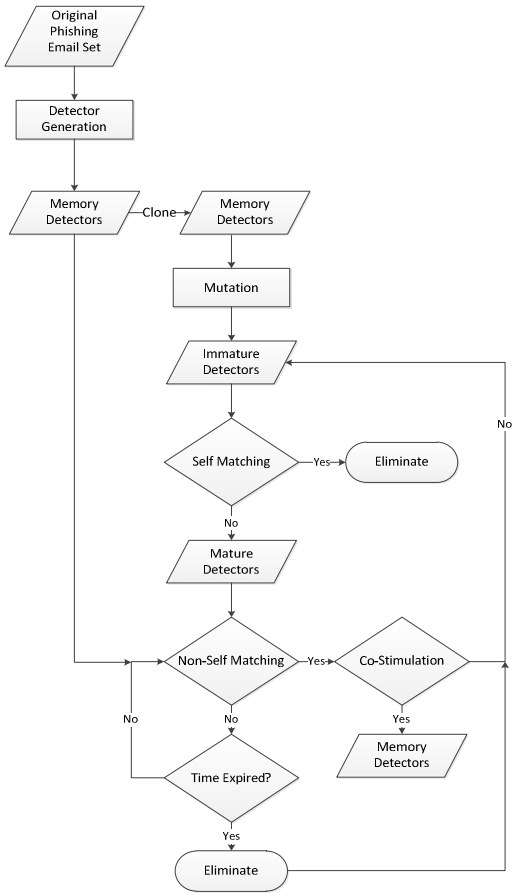


Figure 2: The Design of AIS

Figure 2 depicts the AIS design. The system starts with a set of phishing emails; the basis for the generation of memory detectors. A copy of the memory detectors is saved for future detections while the original is mutated (as immature detectors). Next, the process of Negative Selection is adopted in order to transform the immature detectors to mature ones that do not match any self-pattern (non-phishing emails). The mature detectors are then used to detect incoming phishing emails (Non-Self Matching) together with the aforementioned memory detectors. A detector fires whenever a non-self-matching occurs. In this case, only TP (True Positive) and FP (False Positive) can cause detectors to fire (Table 1). Detector firing caused by non-self matches from any mature detectors will be examined in the process of Co-Stimulation: If the match is a TP, the detector will be saved as a memory detector; otherwise, the detector degenerates to an immature detector. Meanwhile, the mature detectors that fail to match any phishing emails after a predefined time period will be eliminated.

| | DP | DNP |
|---|---|---|
| AP | TP | FN |
| ANP | FP | TN |

Table 1: The Confusion Matrix for Phishing Detection

- AP -- Actual Phishing
- ANP -- Actual Non-Phishing
- DP -- Detected as Phishing
- DNP -- Detected as Non-Phishing

Similar to biological immune systems, our artificial immune system is able to reproduce memory detectors as long as the system remains "infected". We believe that the increasing number of memory detectors and a number of in-system immature detectors can provide strong capabilities for phishing detection to the system.

## IV. Datasets and Experiments

In this section, we present the experiment conducted to detect phishing emails using our artificial immune system. We first describe the preparation process of the experiment and then show the experimentation results.

### A. Preparations

#### 1) Data Collection

For the experiment, we collected 500 emails in total as our corpus. Within the corpus, 100 phishing emails were collected from the Honey Trap database of millersmiles.co.uk [12] which is currently holding 2084338 scams. Also 400 non-phishing emails were downloaded from the Enron datasets [4].

#### 2) The Mutation and The Weighted Fire Method

Regarding all phishing emails in the corpus, we maintain a Black List, BL, for all of the email addresses, such that the BL is defined as:

$$\text{BL} = \{\bigcup_{i=1}^{100} addr_i\}$$

We implemented the Naïve Bayesian method to figure out all the phishing tokens from 100 phishing emails and we constructed a Bad Token List, BTL, defined as:

$$\text{BTL} = \{\bigcup_{i=1}^{100} btb_i\}$$

To mutate the $i^{th}$ memory detector, the system replaces its email address with a different address, $addr_i'$, randomly selected from BL and the system replaces some phishing tokens with ones randomly selected from BTL to form a new bad token bag, $btb_i'$.

The number of links remains unchanged for this experiment.

For matching, a weight value is assigned to each detector component, such that

$w_{addr} + w_{btb} + w_{link} = 1$ . Therefore, the final matching score for the i[th] detector is computed as:

$$w_{addr} * B(addr_i) + w_{btb} * P(btb_i) + w_{link} * B(link_i)$$

$B(addr_i)$ and B($link_i$) are two binary values, each of which use 1 to indicate matched and 0 otherwise; $P(btb_i)$ is the probability of a bad token matching; calculated as:

$$P(btb_i) = \frac{\#\ of\ matched\ bad\ tokens}{total\ \#\ of\ bad\ tokens\ in\ btb_i}$$

If the final score is equivalent to or larger than a user defined fire-threshold value, the detector is treated as fired. We then define a Fire-Alarm-Range value, *FAR*, such that, if the number of *FAR* fired detector(s) is (are) generated, an alarm will be caused. For instance, if the *FAR* is set to 1, one detector firing can then cause the alarm; if the *FAR* is set to 2, the alarm has to be caused by at least two detectors firing. In this case, the AIS has control of its fire-alarm rate by regulating the magnitude of *FAR*. In our experiment, we tested our AIS by adopting different *FAR* values.

### *B. The Experimental Environment and Results*

The experiment is designed to test the performances of the AIS, which is evaluated via its true positive (TP) rate and false positive (FP) rate. The corpus is equally divided into 5 sub-datasets, each of which contains a training set and a testing set. A training set has 80 phishing emails and a testing set has 20 phishing emails and 80 non-phishing emails. The training set is used for memory-detector-generations and immature-detector-reproductions. The testing set is the actual data set used to conduct the evaluation. The performances are evaluated under different *FAR* values: Initially the *FAR* value is set to be 3. It then is changed to be 2 and 1, respectively, in order to make comparison. The experiment was deployed on a desktop with the CPU of Intel i-7, 4GB RAM, 500GB hard drive, and Windows 7 64-bit operating system.

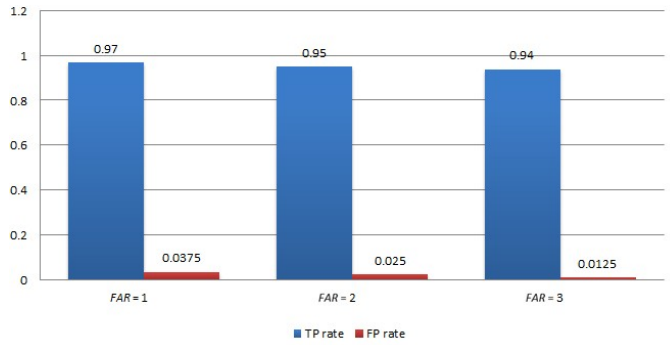


Figure 3: Performance of the AIS

Figure 3 depicts the system performance result. As the *FAR* values decrease from 3 to 1, the average system TP rate increases from 94% to 97%; the average system FP rate increases from 1.25% to 3.75%. We also employed *F*-score to evaluate the performance trend, where:

$$F = \frac{2 * P * R}{P + R}$$

P is the precision and R is the recall. Figure 4 shows that as the TP rate increased, the *F* score also increases from 96.3% to 96.6%.


Figure 4: *F*-scores of the AIS

Figure 5 shows the comparison of the *F*-scores between our AIS and the Weak Estimator method. The AIS has the score of 96.6% which is almost the same as the Weak Estimator's 96.5%.

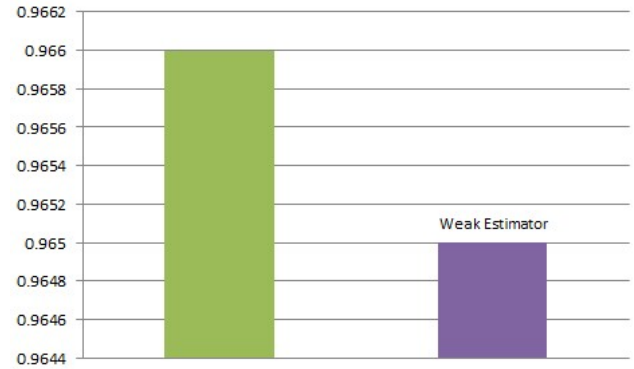

Figure 5: *F*-score Comparison

## V. DISSCUSION AND FUTURE WORK

Our experiment shows the proposed artificial immune system performs well in phishing detections. In addition, the system has the potential to gain immunity by reproducing new memory detectors. This can be achieved by adding previously undetected phishing emails to the system's training set. The AIS also allows people to adjust system configurations for different user requirements. For instance, if users can tolerate a higher FP rate, they will benefit from higher TP rates by implementing our system. On the other hand, if they prefer a low false-alarm rate, they may set a higher fire-threshold value for each detector or choose a higher *FAR* value for the entire system to accomplish the goal.

In our current system design, all detectors share a static fire-threshold value, which may not work well under some circumstances. Therefore, in our future work, we will improve the system design by assigning a dynamic fire-threshold value to each detector. Similarly, a dynamic *FAR* value can be also introduced to the system for the sake of adaptation enhancement.

## VI. CONCLUSIONS

Artificial Immune Systems have gained much attention in many research fields due to their robust ability to adapt along with several other valuable properties. The AISs are being applied on various security-related systems, such as intrusion detection or virus detection systems. In this paper, we propose an AIS in the domain of phishing detection. The system is designed to detect phishing emails through memory detectors and immature detectors. The memory detectors are generated from the training data set, which, in turn, is the phishing emails previously seen by the system. The immature detectors are reproduced through the system's mutation process. In this case, any incoming phishing emails that are similar to the ones in the training set will be detected by the memory detectors, while the incoming phishing emails with new patterns can be detected by the mutated immature detectors. Additionally, our system offers flexibilities on system configurations by allowing users to adjust the fire-threshold value and the *FAR* value to suffice different requirements under different user scenarios.

## ACKNOWLEDGMENTS

This research was funded by the National Science Foundation (NSF) Science & Technology Center: Bio/computational Evolution in Action CONsortium (BEACON). The authors would like to thank the NSF for their support of this research.

## REFERENCES

[1] P. Andrews and J. Timmis, On Diversity and Artificial Immune Systems: Incoporating a Diversity Operator into aiNet, *Neural Nets, LNCS 3931*, Apolloni et al. (Eds.), Springer, 2005.

[2] A. Bhattarai, V. Rus, and D. Dasgupta, Characterizing Comment Spam in the Blogosphere through Content Analysis, In *the proceedings of the IEEE Symposium Series on Computational Intelligence*, March 30-April 2, 2009.

[3] D. Dasgupta and F. González, An Immunity-based Technique to Characterize Intrusions in Computer Networks, Journal of the *IEEE Transactions on Evolutionary Computation*, vol. 6(3), June 2002, pp. 281-291.

[4] Enron Datasets, [Available Online]: http://www.aueb.gr/users/ion/data/enron-spam/

[5] S. Forrest, A. Perelson, L. Allen, and R. Cherukuri, Self-nonself Discrimination in a Computer, In Proceedings of *the IEEE Symposium on Research in Security and Privacy*, Los Alamos, CA, USA, 1994.

[6] P. Graham, A Plan for Spam, [Available Online]: http://www.paulgraham.com/spam.html

[7] S. Hofmeyr and S. Forrest, Immunity by Design: An Artificial Immune System, In Proceedings of *the Genetic and Evolutionary Computation Conference*, vol. 2, 1999, pp. 1289-1296.

[8] S. Hofmeyr and S. Forrest, Architecture for an Artificial Immune System, *Journal of Evolutionary Computation*, vol. 8(4), December 2000.

[9] H. Hou and G. Dozier, Immunity-based Intrusion Detection System Design, Vulerability Analysis, and GENERTIA's Genetic Arms Race, *the ACM Symposium on Applied Computing*, Santa Fe, NM, USA, March 13-17, 2005, pp. 952-956.

[10] H. Hou, J. Zhu, and G. Dozier, Artificial Immunity using Constraint-based Detectors, In Proceedings of *the 5th Biannual World Automation Congress*, Orlando, FL, USA, 9-13, June 2002, pp. 239-244.

[11] J. Kephart, A Biologically Inspired Immune System for Computers, In Proceedings of *the 4th International Workshop on Synthesis and Simulation of Living Systems*, 1994, pp. 130-139.

[12] MillerSmiles, [Available Online]: http://www.millersmiles.co.uk/

[13] B. Oommen and L. Rueda, Stochastic Learning-based Weak Estimation of Multinomial Random Variables and its Applicationsto Pattern Recognition in non-Stationary Environments, *the Journal of Pattern Recognition*, vol. 39(3), March 2006.

[14] P. Parham, The Immune System, 2nd Edition, *Garland Science, New York, USA*, 2005.

[15] Phishing, *National Information Security Glossary*, [Available Online]: http://www.cnss.gov/Assets/pdf/cnssi_4009.pdf

[16] T. Rahman, S. Suliman, and I. Musirin, Artificial Immune-based Optimization Technique for Solving Economic Dispatch in Power System, *Neural Nets, LNCS 3931*, Apolloni et al. (Eds.), Springer, 2005.

[17] Spam, *The American Heritage Dictionary of the English Language*, the 6th Edition, Houghton Mifflin Company, 2006.

[18] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song, Design and Evaluation of a Real-Time URL Spam Filtering Service, In *IEEE Symposium on Security and Privacy,* Paris, France, 2011.

[19] J. Zdziarski, Ending Spam: Bayesian Content Filtering and the Art of Statistical Language Classification, *No Starch Press*, San Francisco, CA, USA, 2005.

[20] J. Zhan and L. Thomas, Phishing Detection using Stochastic Learning-based Weak Estimators, In Proceedings of *the IEEE Symposium on Computational Intelligence in Cyber Security*, Paris, France, April 2011.