AFRL-SR-AR-TR-03-

0457

REPORT DOCUMENTATION PAGE

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
01 - 08 - 2003	FINAL PERFORMANCE REPORT	01May2001 - 30Apr2003

4. TITLE AND SUBTITLE

Critical Infrastructure Protection and Information Assurance Science and Engineering Augmentation Awards for Fellows (CIPIAF) FY 2001.
(1) Statistical Techniques for Detecting Intrusions
(2) Hostile Data Stream Testing

5a. CONTRACT NUMBER
F49620-01-1-0294

5b. GRANT NUMBER
Does Not Apply

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Jorgensen, Alan, Albert
Rekab, Kamel
Whittaker, James, A.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Florida Institute of Technology
150 West University Boulevard
Melbourne, Florida 32901-6975

8. PERFORMING ORGANIZATION REPORT NUMBER
Index 2504 / Org CSC300 / Fund 200441
Program 620

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Same as #7 Above

10. SPONSOR/MONITOR'S ACRONYM(S)
FL Tech

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
2504

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Florida Tech hired two Fellows for two years under the CIPIAF program. They acted as co-Principal Investigators working with Dr. James Whittaker in the areas of (1) statistical methods for intrusion detection and (2) automatic methods for hostile data stream testing. Our findings were successful in both projects. Dr. Kamel Rekab was the Principal on the first project and his results showed that attack traffic possesses a strong statistical signature and logistic regression analysis can be used to distinguish between legitimate and attack traffic. Dr. Rekab has been retained as a full faculty in our department to continue this research. Dr. Alan Jorgensen was the Principal on the second project and used his technique to find a zero-day exploit in Macromedia Flash. He subsequently received funding from Macromedia and is currently working as a private consultant using his techniques on behalf of a number of software vendors.

15. SUBJECT TERMS

Information assurance, Zero-day exploits, Intrusion detection, Security Testing, Information warfare

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. James A. Whittaker
U		SAR		8	19b. TELEPHONE NUMBER (Include area code) (321) 674-7638

CIPIAF Fellows Final Report

Florida Institute of Technology
PI: James A. Whittaker
Fellows: Alan Jorgensen, Kamel Rekab

Abstract: Florida Tech hired two Fellows for two years under the CIPIAF program. They acted as co-Principal Investigators working with James Whittaker in the areas of (1) statistical methods for intrusion detection and (2) automatic methods for hostile data stream testing. Our research results were very positive in both projects. Dr. Kamel Rekab was the Principal on the first project and his results showed that attack traffic possesses a strong statistical signature and logistic regression analysis can be used to distinguish between legitimate and attack traffic. Dr. Alan Jorgensen was the Principal on the second project and used his technique to find a zero-day exploit in Macromedia Flash.

Training of the Fellows: Each of the Florida Tech Fellows attended a number of conferences and seminars on computer security and worked closely with experts at Florida Tech. Dr. Rekab is a statistician and required retraining in both computer science and computer security. Dr. Rekab has been retained as a full faculty in our department to continue based on his outstanding performance on this grant and is now actively collaborating with Dr. Gerald Marin and Dr. William Allen who are both researchers in network security.

Dr. Jorgensen's degree was in computer science and he came up to speed on security quickly. He was able to apply his knowledge of testing to the security problem and with Dr. Whittaker, they invented novel techniques for buffer overrun testing that has resulted in a number of zero-day exploits. Dr. Jorgensen received funding from Macromedia during his Fellowship and is currently working as a private consultant using his techniques on behalf of a number of software vendors.

Statistical Analysis for Intrusion Detection Based on Internet Protocol Characteristics

This study seeks to predict intrusions based on the internet protocol characteristics. In this paper, we propose a test for anomaly detection based on standard statistical tests. These include the Logistic Regression and Receiver Operating Characteristic Curve. We were able to predict anomalies while minimizing false alarms and maximizing intrusion detection.

We created the best model that predicts intrusions based on a set of 7911 packets. The model's performance was then measured by comparing the predicted intrusions and the observed intrusions.
The observed intrusions were based on 8127 packets that were not used to determine the model.

Among the 8127 packets, there were 4315 intrusions that were perfectly predicted, with a false alarm of .24 percent.

Recent Work

The literature in computer intrusion detection is too vast to survey here. Earlier work was based on signature detections. That is, matching patterns in network traffic to the patterns of known attacks. An alternative approach is anomaly detection, which models normal traffic and signals any deviation from this model as suspicious. Among the most recent techniques in anomaly detection, Data Mining techniques are used to discover consistent patterns of system features that describe program and user behavior, and use the set of relevant system features to compute classifiers that can recognize anomalies and known intrusions. For more details. See Lee and Stolfo (1).
Previous approaches are based on univariate investigations which can not be useful in detecting anomalies that are present mainly because of the interaction between fields. Our approach seeks to answer the following questions:

1. What combination of fields has an impact on detecting anomalies?
2. What is the model that relates the most important fields and the probability of an anomaly?

Logistic Model

The logistic regression model was performed to predict the probability that an intrusion occurs given a set of internet protocol characteristics. The model Chi-square **(I)** is very significant
(P value=.0000). It shows that our predictive model performs very well. See **(IV)**.

The logistic model given by the prediction equation **(III)** was used to predict intrusions on a new set of data that consists of 8127 packets. The 8127 packets had 4315 intrusions. Our model was able to predict every intrusion. See **(II)**.

The Receiver operating characteristic curve also shows that our predictive model is nearly perfect. The probability of perfection is **100 percent.** See **(V)**

References

1. Du Mouchel , W. "Computer Intrusion Detection Based on Bayes Factors for Comparing Command Transition Probabilities". National Institute of Statistical Science, 1999.

2. Lee, W., and Stolfo, J. S. " Data Mining Approaches for Intrusion Detection. Proc. 1998 7th USENIX Security Symposium, 1998.

3. Hosmer, D. W. and Lemeshow, S. Applied Logistic Regression, John Wiley, 1989.

Appendix A: Predictive Model Based on 7911 Packets

```
Total number of cases:      16260 (Unweighted)
Number of selected cases:   8000
Number of unselected cases: 8260

Number of selected cases:                    8000
Number rejected because of missing data:     88
Number of cases included in the analysis:    7912


Dependent Variable..   Y

Beginning Block Number  0.  Initial Log Likelihood Function

-2 Log Likelihood   10863.465

* Constant is included in the model.

Beginning Block Number  1.  Method: Enter

Variable(s) Entered on Step Number
1..        IP_TOS
           IP_LEN
           IP_ID
           IP_FLAGS
           IP_TTL
           IP_CSUM
           TCP_DPOR
           TCP_SEQ
           TCP_ACK
           TCP_WIN
           TCP_RES
           TCP_OFF
           TCP_CSUM

Estimation terminated at iteration number 21 because
Log Likelihood decreased by less than .01 percent.

 -2 Log Likelihood          240.505
 Goodness of Fit         317932.274
 Cox & Snell - R^2           .739
 Nagelkerke - R^2            .990

                     Chi-Square     df Significance

 Model                10622.961     13         .0000
 Block                10622.961     13         .0000
 Step                 10622.961     13         .0000
```

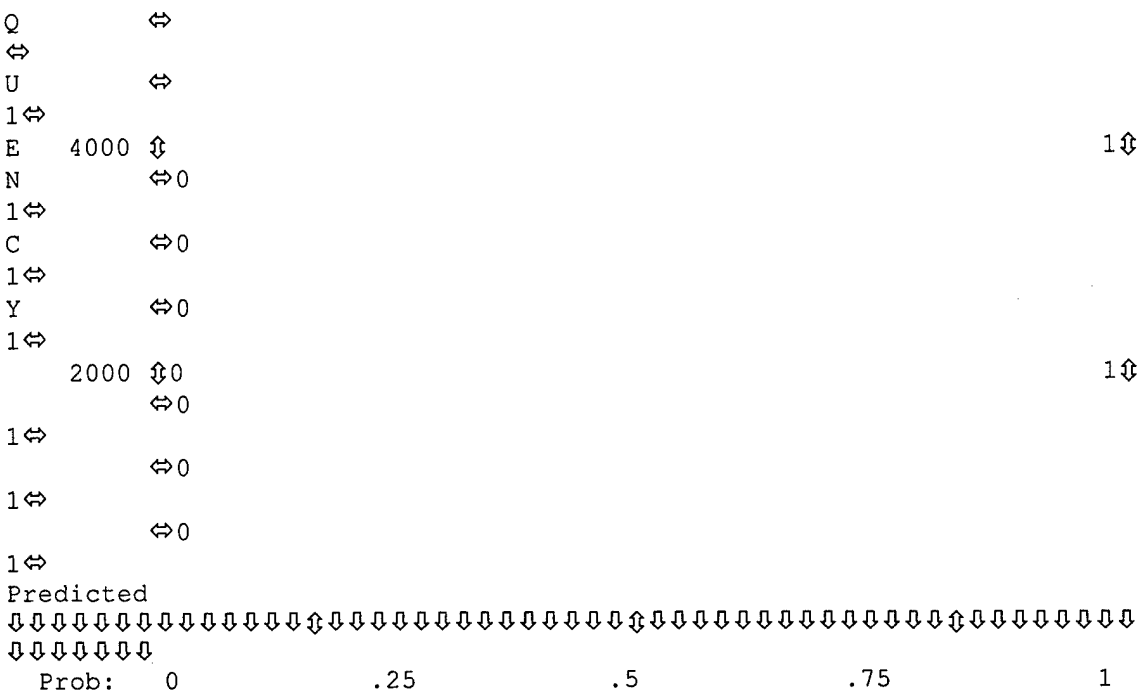
Prediction Equation using Logistic Regression

---------- Variables in the Equation ----------

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
IP_TOS	.7321	4244.5529	.0000	1	1.0000	.0000	2.0795
IP_LEN	-1.2557	3.1835	.1556	1	.6932	.0000	.2849
IP_ID	-.0039	.0006	45.9527	1	.0000	-.0636	.9961
IP_FLAGS	7.7248	675.1952	.0001	1	.9909	.0000	2263.8422
IP_TTL	-.0350	.0089	15.4674	1	.0001	-.0352	.9656
IP_CSUM	-7.7E-06	1.179E-05	.4314	1	.5113	.0000	1.0000
TCP_DPOR	-.0004	6.258E-05	42.2585	1	.0000	-.0609	.9996
TCP_SEQ	-3.6E-10	1.805E-10	4.0011	1	.0455	-.0136	1.0000
TCP_ACK	4.97E-10	1.912E-10	6.7692	1	.0093	.0210	1.0000
TCP_WIN	-.0002	3.345E-05	22.1949	1	.0000	-.0431	.9998
TCP_RES	.4372	14.6540	.0009	1	.9762	.0000	1.5484
TCP_OFF	7.7803	12.7257	.3738	1	.5409	.0000	2392.9250
TCP_CSUM	5.13E-06	1.163E-05	.1945	1	.6592	.0000	1.0000
Constant	-21.1926	1645.3527	.0002	1	.9897		

Accuracy of the Model

Observed Groups and Predicted Probabilities



1 new variables have been created.

Name	Contents
PRE_1	Predicted Value

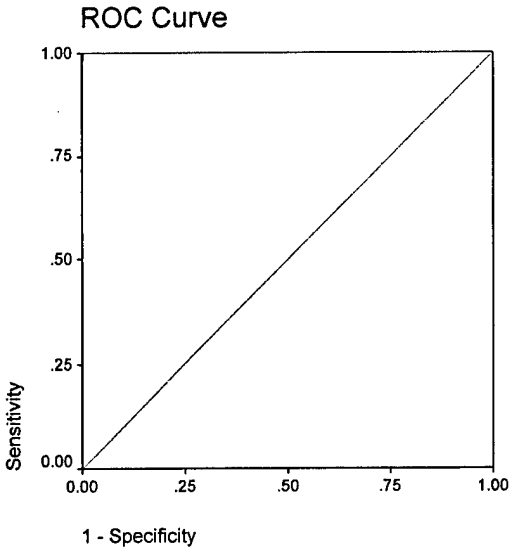
Appendix B: Receiver Operating Characteristic Curve

Case Processing Summary

Y	Valid N (listwise)
Positive[a]	8726
Negative	7313
Missing	221

Larger values of the test result variable(s) indicate stronger evidence for a positive actual state.

a. The positive actual state is 1.



Area Under the Curve

Test Result Variable(s): Predicted Value

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.999	.000	.000	.998	.999

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Hostile Data Stream Testing

As a Senior Research Scientist funded by U.S. Air Force Grant # F49620-01-1-0294, James A. Whittaker, Principal Investigator, during the past year I have improved upon the development of the Hostile Data Stream Software Testing Technology. (Jorgensen, A. "Testing with Hostile Data Streams." *ACM Software Engineering Notes*, March 2003. See http://www.cs.fit.edu/~tr/cs-2003-03.rtf .) This technology has proven that viruses and other software security compromises can be initiated by means of otherwise passive data file formats such as Portable Document Format (PDF) and Shock Wave Format (SWF). For examples, see http://cs.fit.edu/~ajorgens/CrashCases/Acrobat.htm. Over 750,000 test cases have been applied to Adobe Acrobat Reader and Macromedia Flash Player resulting in the discovery of over 40 unique symptoms of buffer overrun. Buffer overruns are a major source if internet security vulnerabilities. I obtained a small grant ($25,000) from Macromedia to fund application of the Hostile Testing Technology to Macromedia Flash Player. The vulnerabilities discovered by this technique have been reported to the respective vendors and in one instance reported to the Computer Emergency Response Team Coordinating Committee (CERT/CC). Some of these defects (though by no means most) have been repaired with associated security risk

announcements.
(http://www.infoworld.com/article/03/03/04/HNmacromedia_1.html)

A proposal for additional research funding has been submitted to the National Science Foundation to research such questions as: 1) How can this technique be applied to server side software applications to discover security vulnerabilities in server software? 2) Is there a distinction between risks posed by different buffer overruns and can a buffer overrun with apparently low risk be "promoted" to pose a more severe security risk? 3) How can this method be applied to encrypted or encoded data streams? 4) How can this method be applied to complex protocols?

The testing system has been named SHoTS (Software Hostile Testing System). Current research activities that need additional funding to complete include a redesign of the "Driver" function to create a generic driver. In the current system, a custom driver must be constructed for each new application to be tested because of problems such as dealing with various methods of recovery (or failure to recover) from corrupt file detection. This new driver (still under test as of this writing) appears to recover from all situations encountered by an application dealing with hostile (malformed) data.

Some work has been done on item 2) above that indicates that a single symptom of a buffer overrun is not sufficient to determine the severity of the risk to security of the underlying defect. (In two instances I have been able to "promote" a buffer overrun symptom from a less severe security risk to a more severe risk. This is anecdotal and more research is necessary to establish the general case.) Current practice is to perform "triage" on failure reports based upon a single symptom. More research is needed to establish that this is a poor practice when applied to symptoms of software security risk. Though no funding exists to support this activity, some students have been volunteering to apply this testing technique to other web enabled applications. In every instance where this technique has been applied, we have found that user systems are vulnerable to security attacks because of software defects detected by this testing system.

The paper, "On the Security Risks of Not Adopting Hostile Data Stream Testing Techniques" by Alan A. Jorgensen and Scott Tilley was presented at ACSE 2003: 3rd International Workshop on Adoption-Centric Software Engineering, in Portland, Oregon, May 9, 2003.