SAND2011-1599P

Statistical Modeling of Malware Behavior to detect New Threats

Julie Ard

EEC 274 Winter 2011 Final Project

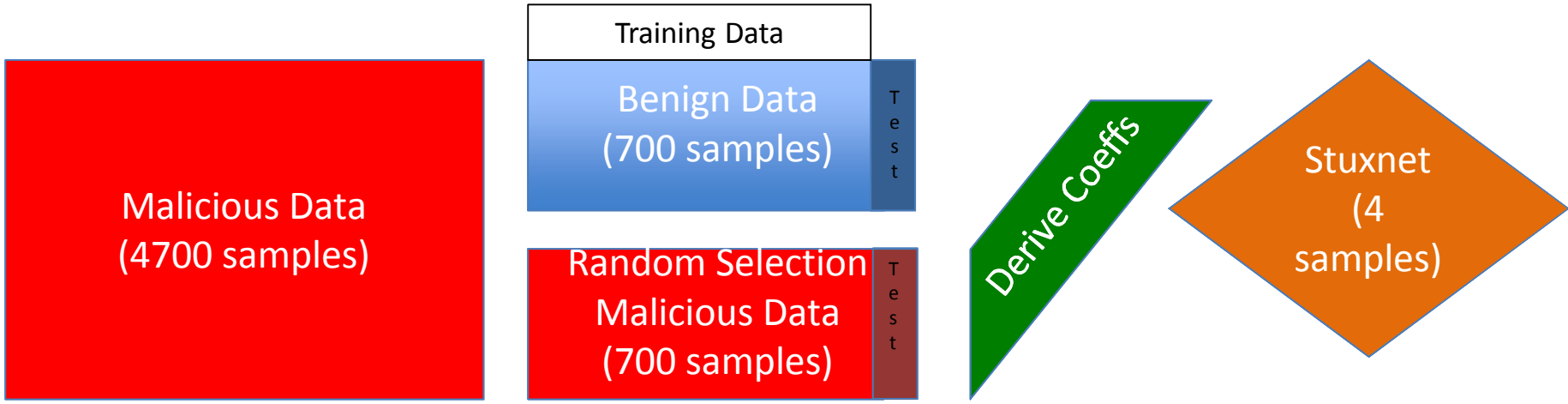
11 Mar 2011

Motivation

- Stuxnet went undetected for six months
  - Propagated via physical media and vulnerable hosts
  - Selectively infected only the hosts it wanted, stayed "under the radar"
- "The world's first precision-guided cyber munition"
- Expected that it will influence future emerging threats
  - AV tools detect signature and polymorphic variants
  - What about the next Stuxnet?

Problem Statement

- Two main detection categories
  - Signature Scanning
  - Anomaly Detection
- Can statistical models detect a malicious file not included in the original data set?



Related Work

- ESET and Symantec have performed detailed analyses of known Stuxnet variants
- 32 files collected from Offensive Computing
- Detection focus is on AV signature scanning

Approach/Methodology

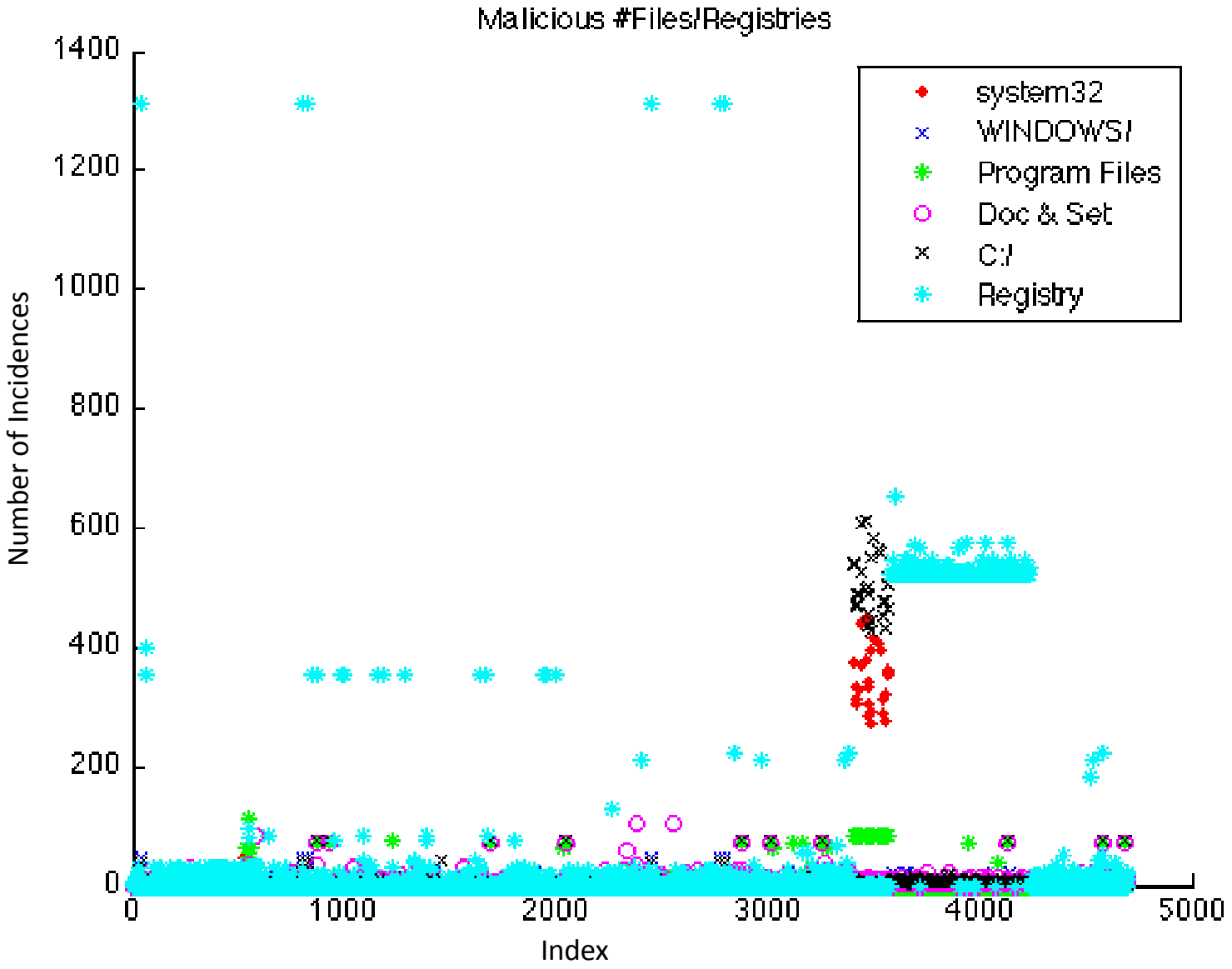
- Use known malicious and benign software behavioral data to derive coefficients of chosen behavioral variables
- Test models on randomly selected test data (10%) not included in the training set
- Test them on Stuxnet data
  - Not present in either the training or test set
  - Verified via MD5 hash

Variables

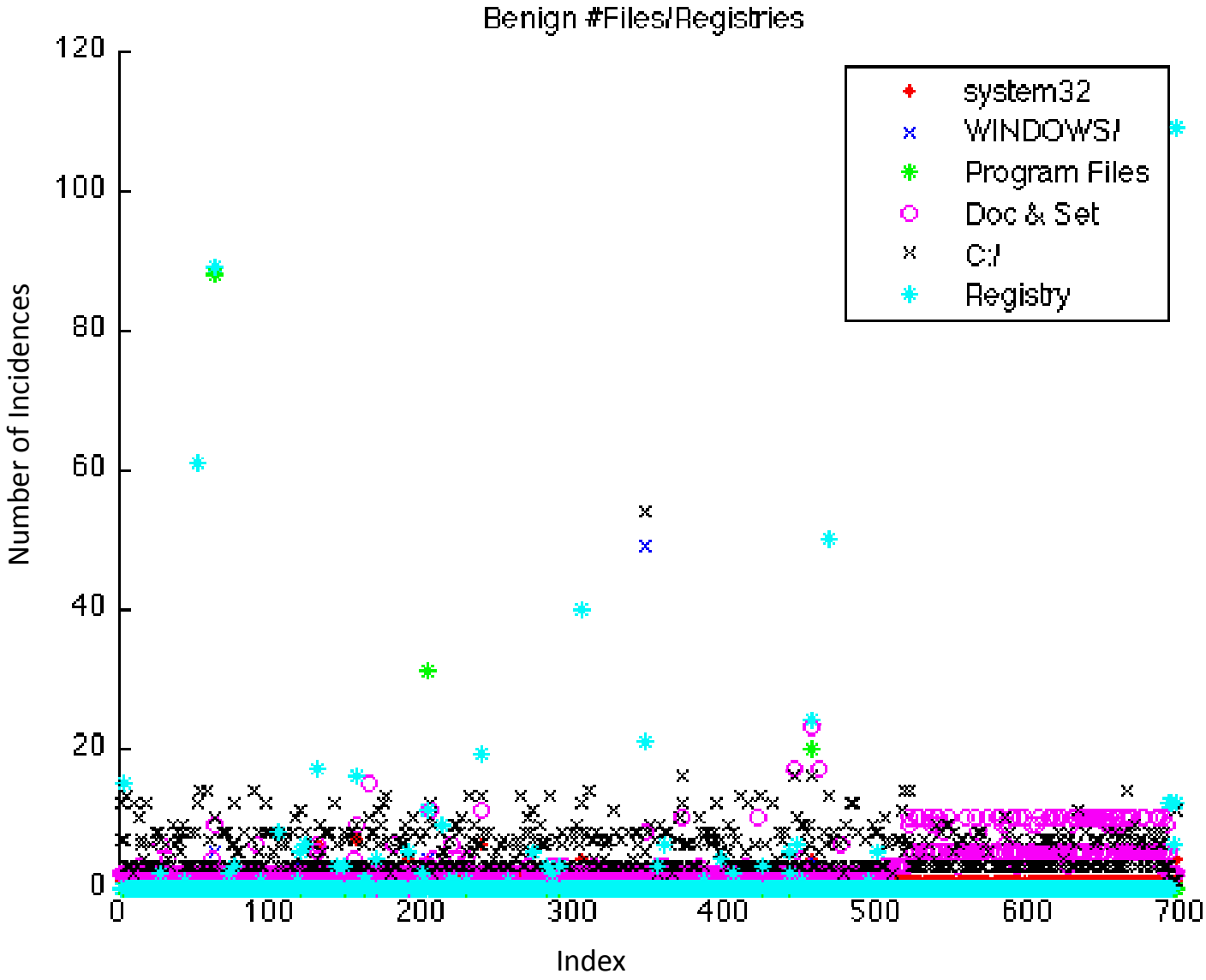
1. # of files created or modified in the C:/WINDOWS directory (excluding system32)
2. # of files created or modified in the C:/WINDOWS/system32 directory
3. # of files created or modified in the C:/Program Files directory
4. # of files created or modified in the C:/Documents and Settings directory
5. # of files created or modified in the root C:/ directory
6. # of registries read, created, or modified

7. # of DNS queries
8. # of tcp connections
9. # of http connections
10. # of udp connections

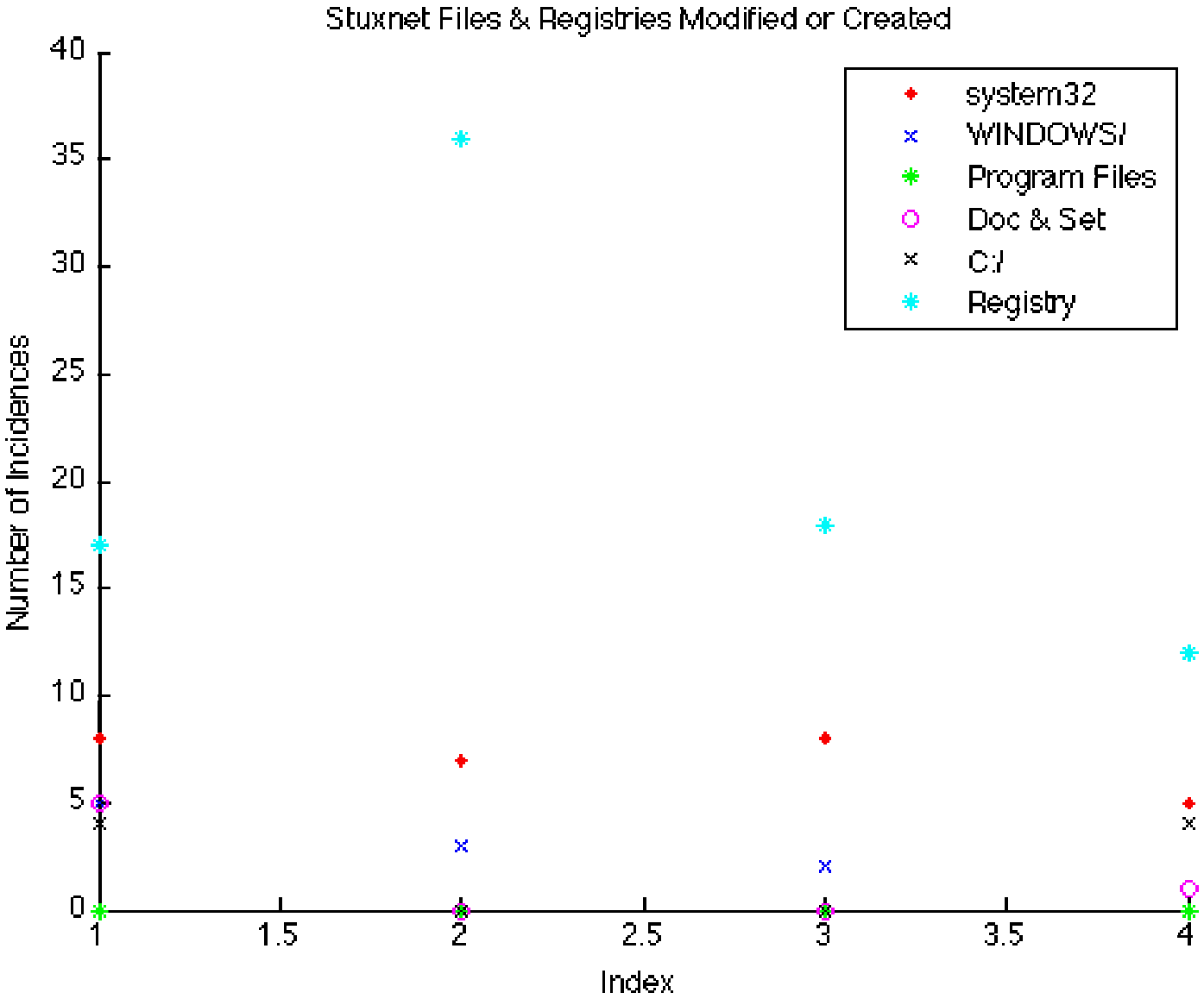
Raw Data – Filesystem – Malicious



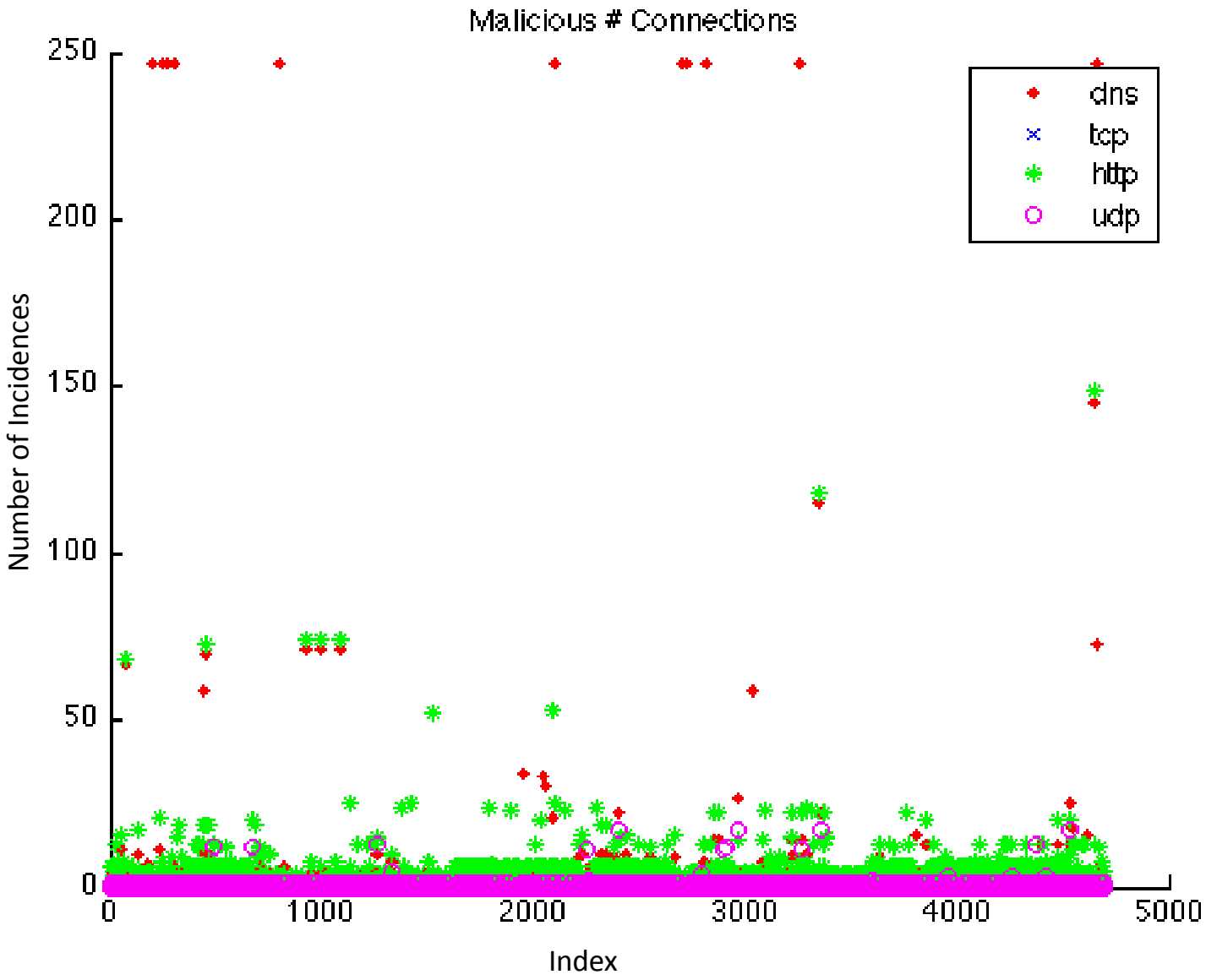
Raw Data – Filesystem – Benign



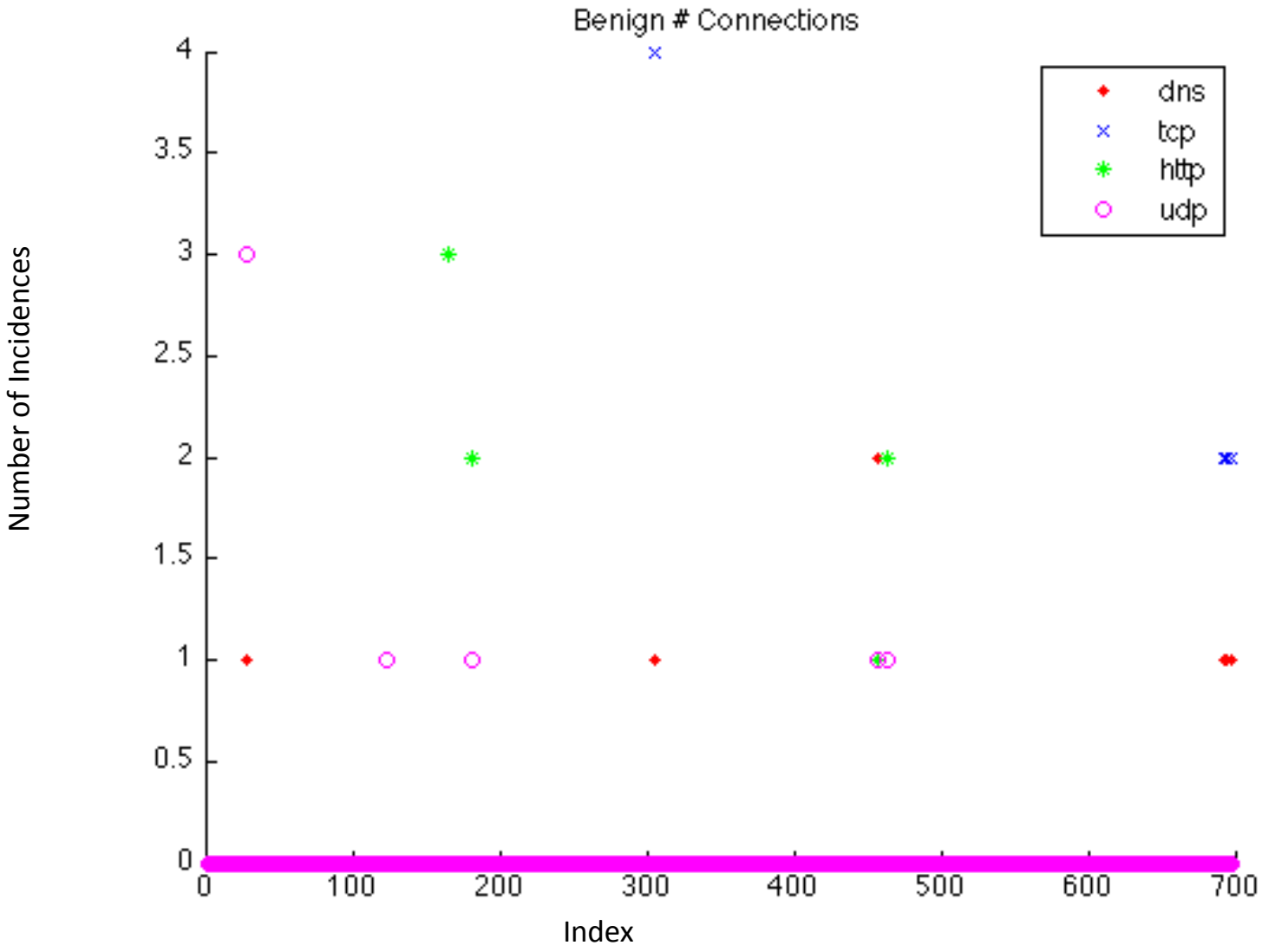
Raw Data – Filesystem – Stuxnet



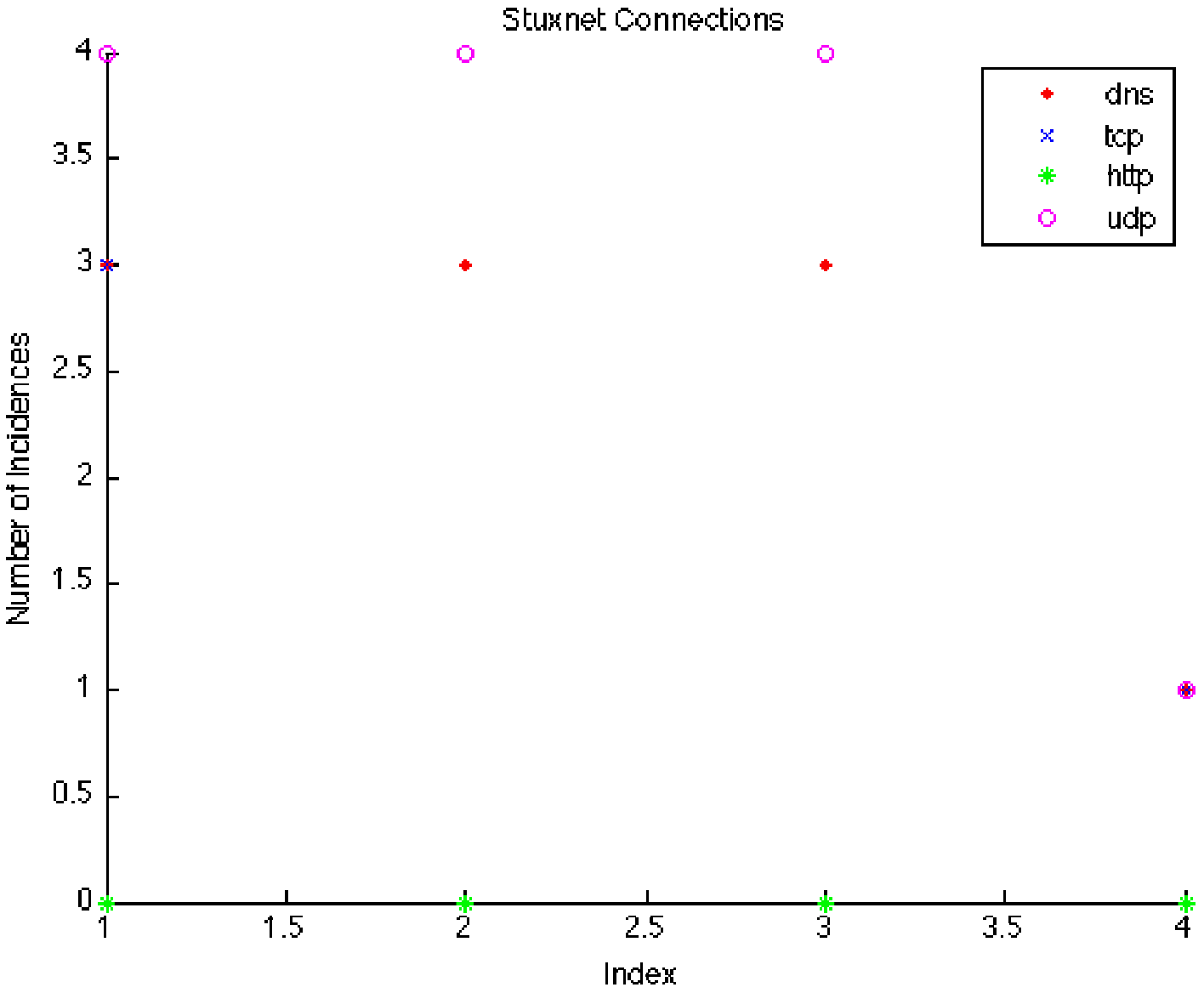
Raw Data – Connections – Malicious



Raw Data – Connections – Benign



Raw Data – Connections – Stuxnet

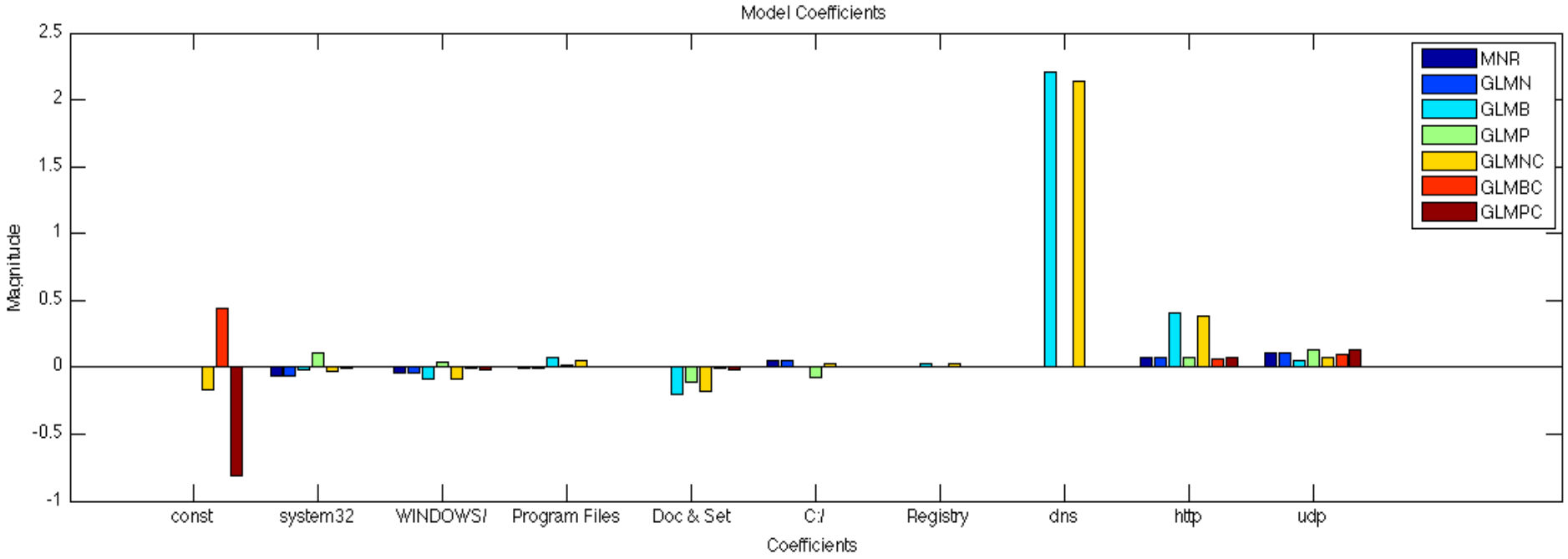


Models

- Output b's where X is the data, Y is 0's or 1's of predicted responses
- Sans constant coefficient
  - $b_1*X_1 + b_2*X_2 + \ldots + b_{10}*X_{10} = Y$
- With constant coefficient
  - $b_0 + b_1*X_1 + b_2*X_2 + \ldots + b_{11}*X_{11} = Y$

Model ID	Description	Distribution	Constant Coeff
MLR	Multilinear Regression	n/a	No
GLMN	Generalized Linear Model	Normal	No
GLMB	Generalized Linear Model	Binomial	No
GLMP	Generalized Linear Model	Poisson	No
GLMNC	Generalized Linear Model	Normal	Yes
GLMBC	Generalized Linear Model	Binomial	Yes
GLMPC	Generalized Linear Model	Poisson	Yes

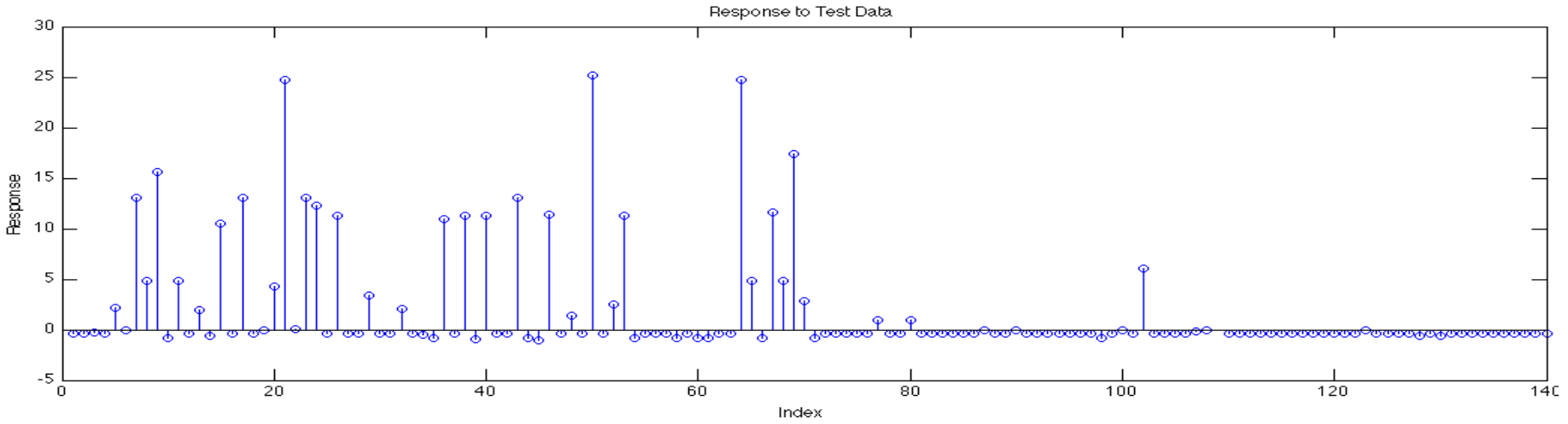
Results: Coefficients



- Negative coefficients associated with benign data (training score=0)
  - Constant, file activity, and TCP (not pictured)
- Positive coefficients imply malicious behavior (training score=1)
  - DNS, HTTP, UDP

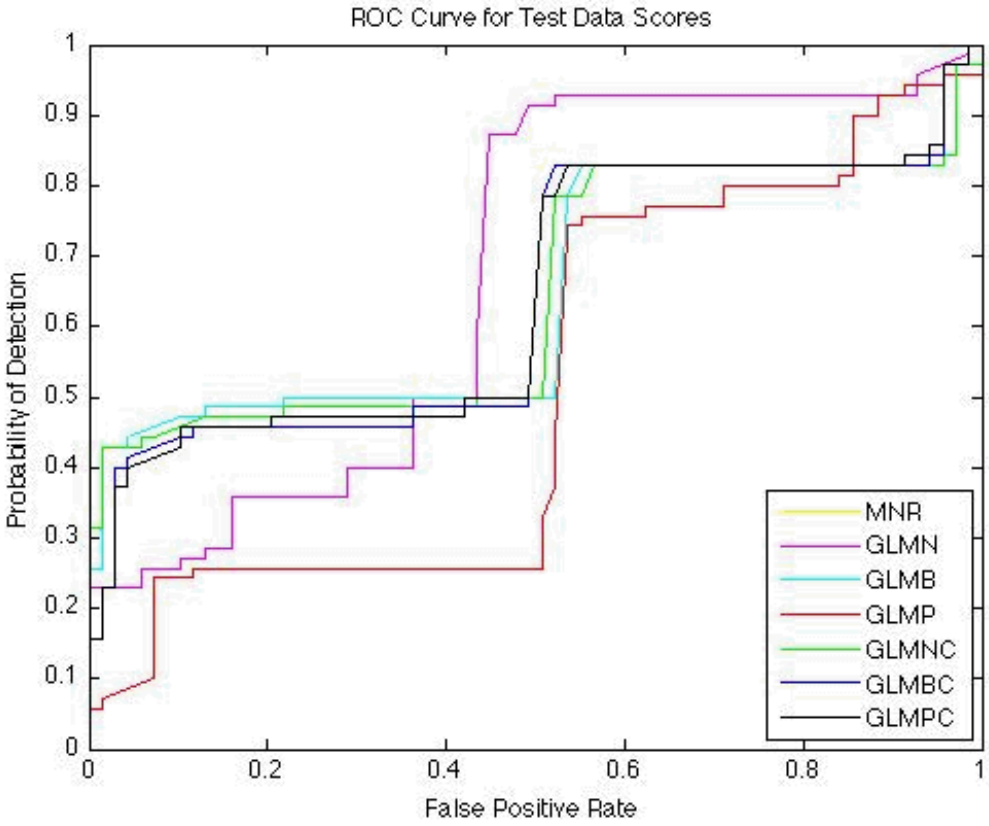
Results: Response to Training Data

- GLMB response pictured
- First half (1:70) malicious data
- Second half (71:140) benign data



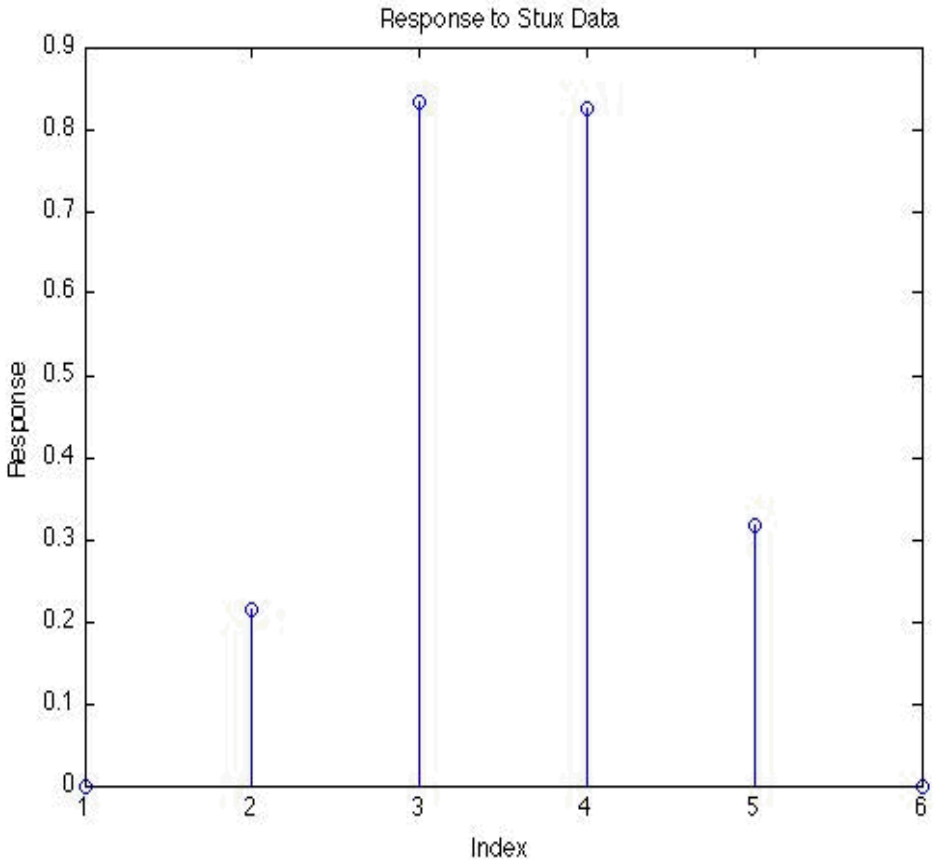
Results: ROC

- Threshold filtering does not leverage data
- GLMN/MLR
  - 91% Pd
  - 48% FAR



Results: Stuxnet

- The GLMN and MLR models scored all 4 as benign
- All other models scored #2 &3 as malicious but #1 & 4 as benign



GLMNC Response to Stux

Conclusions

- Small influence of Registry variable consistent across all models
  - Stuxnet creates or modifies 21 on average
  - Training malware creates or modifies 3 (overall mean=77)
  - Known benign software creates or modifies 1
- High influence of network activity correctly classified 2 of 4 as malicious
- Large negative TCP coefficient
  - Malware max is 2
  - Benign max is 4
  - Stuxnet max is 3

Variable	Character
Constant	n/a
WINDOWS	n/a
system32	Malicious
Program	n/a
Docs and Set	Benign
C:/	Malicious
Registries	Malicious
DNS	Malicious
TCP	Benign
HTTP	n/a
UDP	Malicious

Conclusions

Stux	MLR	GLMN	GLMB	GLMP	GLMNC	GLMBC	GLMPC
1	B	B	B	B	B	B	B
2	B	B	M	M	M	M	M
3	B	B	M	M	M	M	M
4	B	B	B	B	B	B	B

- Benign data issues
  - Analysis not comparable with that of malware
  - Benign files ask permission and require user interaction; malware does not

Future Work

- More variables
- String analysis
- Variable-length data
  - Source ports, dest ports for each connection
- More model types (multivariate, higher order, nonlinear)
- Assignment of training scores / centering data
- Malicious data classification
- Conditional probabilities (events)

References

- Arbor Networks, <http://arbornetworks.com>
- Offensive Computing, <http://offensivecomputing.net>
- PortableApps.com - Portable software for USB, portable and cloud drives. Rare Ideas, LLC. 6 3 2011 <http://portableapps.com>.
- Van Randwyk, Jamie, et al. "Farm: An automated malware analysis environment." 42nd Annual IEEE International Carnahan Conference on Security Technology. IEEE ICCST, 2008. 321-325.
- Matrosov, Aleksandr, et al. Stuxnet Under the Microscope. ESET. Revision 1.3.1. <http://www.eset.com>.
- Falliere, Nicolas, et al. W32.Stuxnet Dossier. Symantec Security Response. Version 1.4 (February 2011).