Using File Relationships in Malware Classification

Nikos Karampatziakis, Cornell/Microsoft
Jay Stokes, Microsoft Research
Anil Thomas, Mady Marinescu, Microsoft Corp.

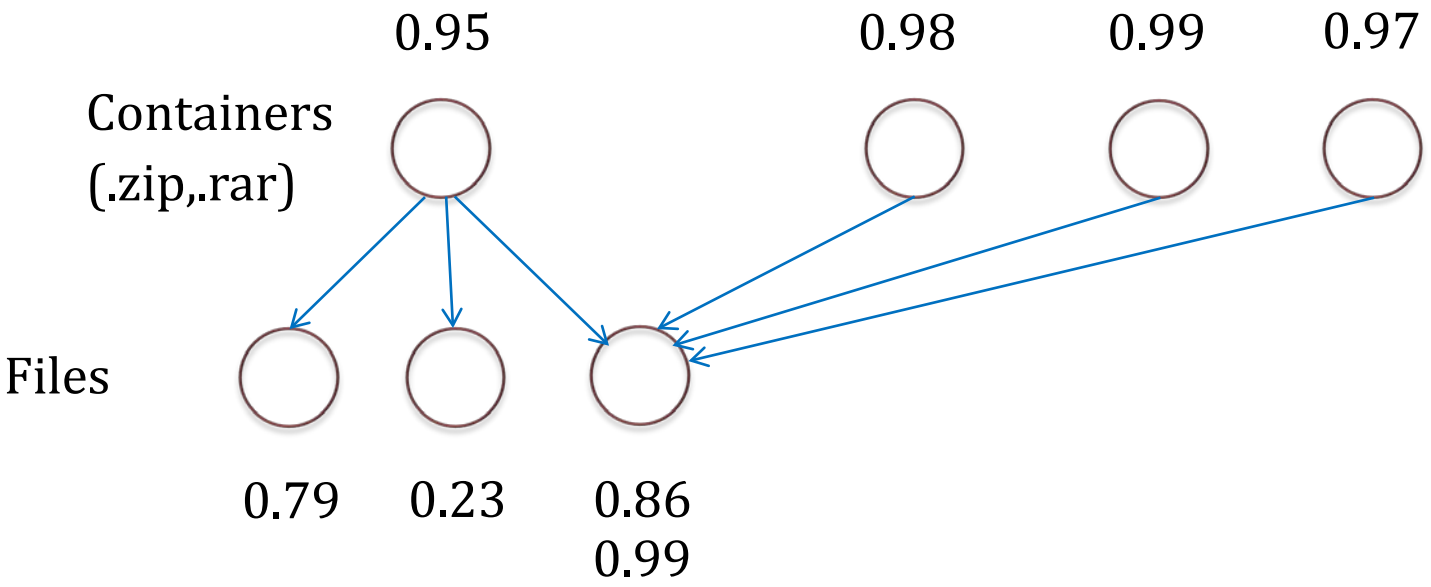
Outline

- Motivation
- Baseline Classifier
- Container Classification
- Improved File Classification
- Results

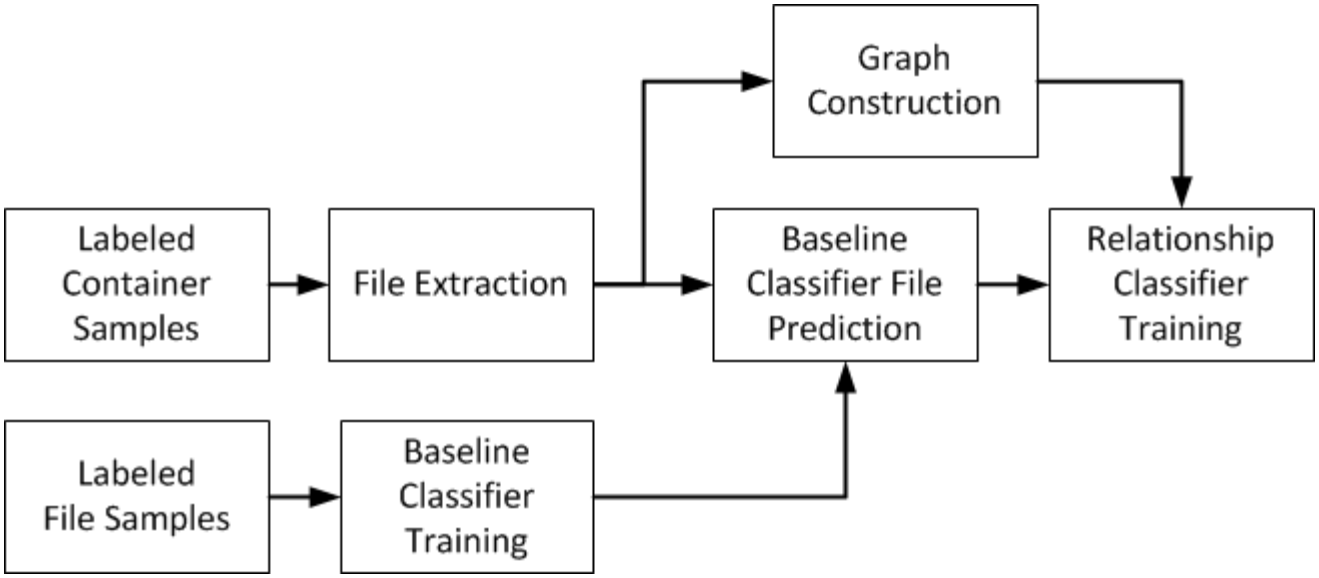
Motivation

- Automated malware classification
  - Needed to combat today's malware
- Full automation
  - Requires classifiers very low FP rate with an acceptable FN rate
- Typically files are analyze in isolation
- Recent work considers file/machine relationship [Chau2011],[Ye2011]
- Can file container relationships help?

Using File Relationships

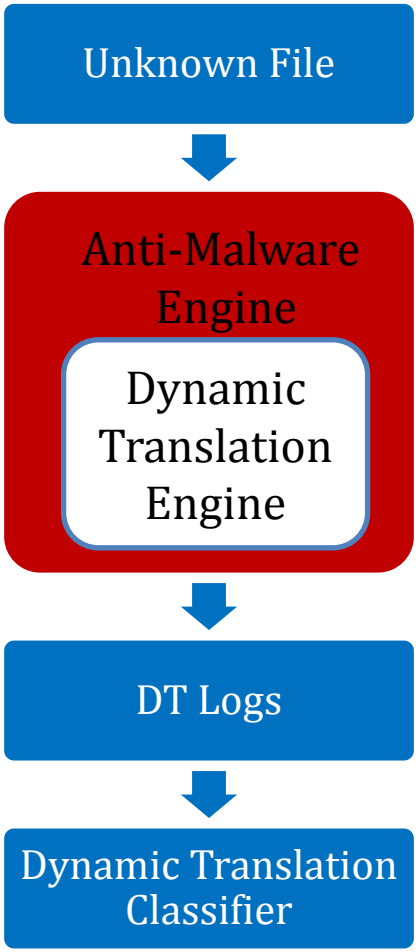


File Relationship Classifier Training



Baseline Classifier

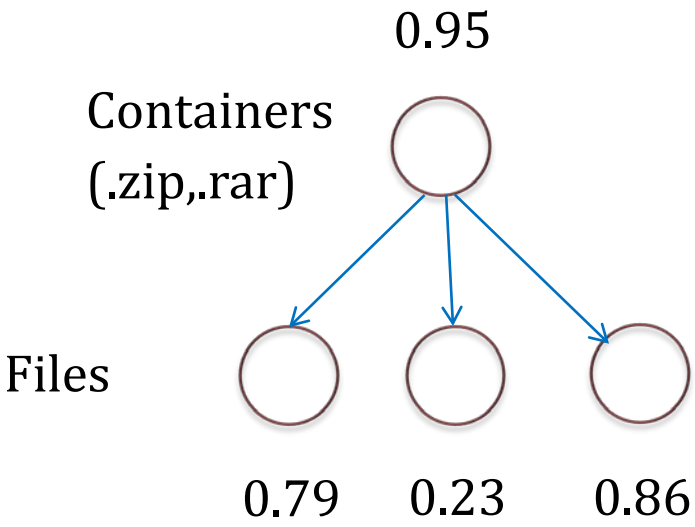
- Dynamic Translation runs files in a sandbox in the AM engine
- Baseline Classifier Features
  - API w/ parameters
  - API tri-grams
  - Unpacked strings
  - Static analysis
- Main Goal: Train a classifier to determine if an unknown file is malware or benign



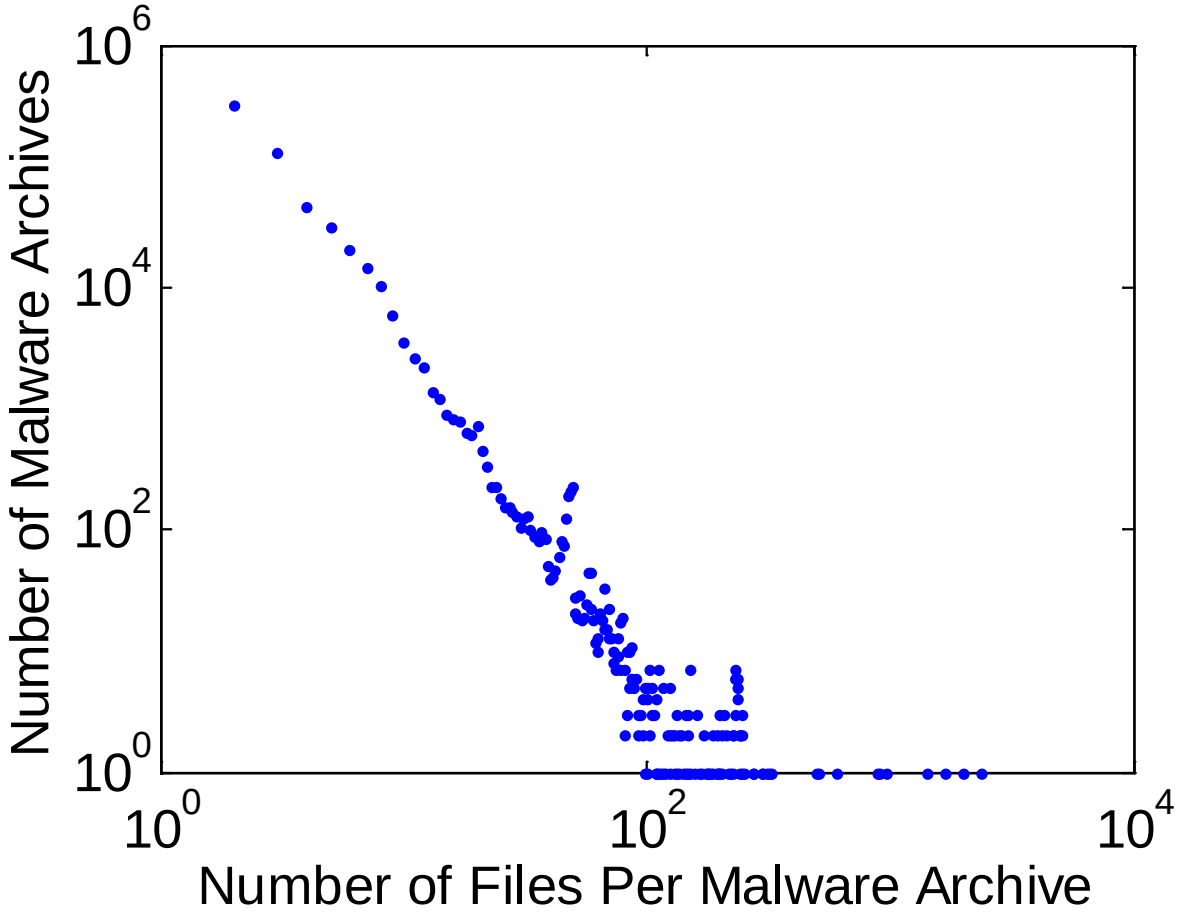
Baseline Classifier Training

- Trained with 2.6M labeled files
  - 1.8M malware
  - 0.8M benign
- 179K Sparse Binary Features
- 134 malware families, general malware and benign classes
- Multi-class logistic regression
- 1.3% FP rate, 0.7% FN rate

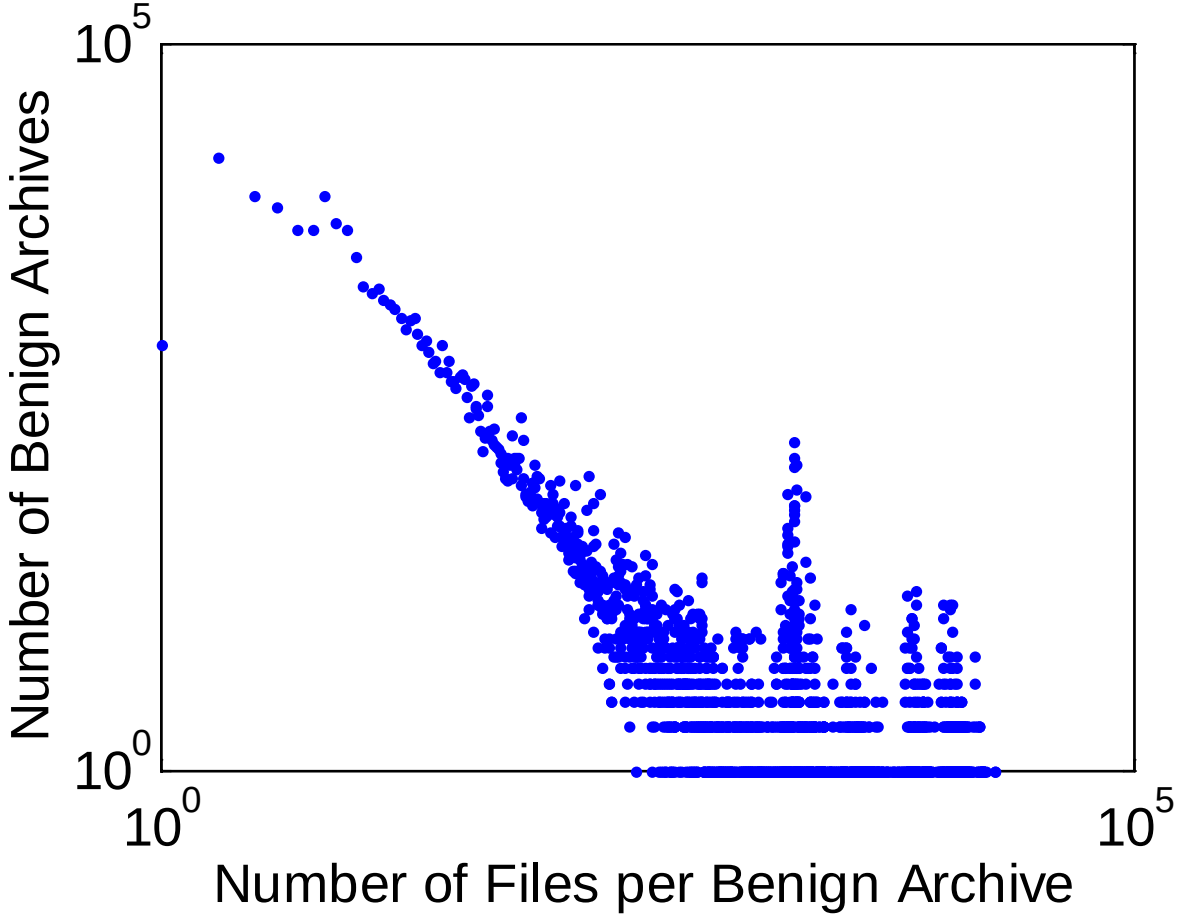
Container Classification



Distribution of Files in the Malware Containers



Distribution of Files in the Benign Containers



Container Classification

- Union Bound

$$p_c(y_i = 1) = 1 - p_c(y_i = 0) = 1 - \prod_{f=0}^{F-1}(1 - p_{b,f})$$

- Max Neighborhood

$$p_c = p_{b,max}$$

- Biased Logistic Regression Model

$$p_c(y_i = 1|x) = v^T x + b$$

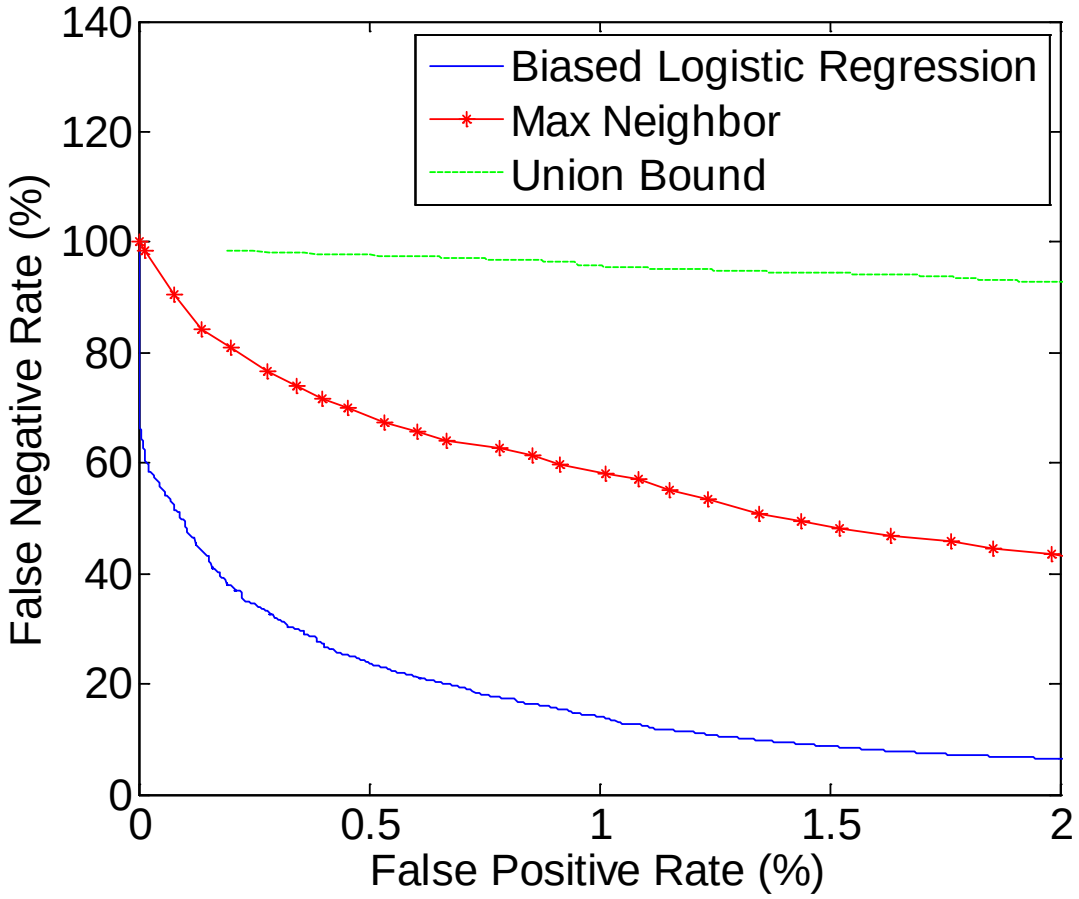
Biased Logistic Regression Features

- Histogram of contained file probabilities
  - Split the interval [0,1] into 20 equally sized bins
  - Features 2j and 2j+1: fraction and logarithm of the number of contained files predicted to be malware
  - Absolute and relative numbers that may affect our decision
  - Similar histograms for benign and inconclusive files
- Three additional features
  - $\log \frac{p_{b,max}}{1-p_{b,max}}$
  - $\log N_f$
    - N_f - number of files in the container
  - Product of the first additional features
    - Captures interactions between the number of files and the maximum file probability

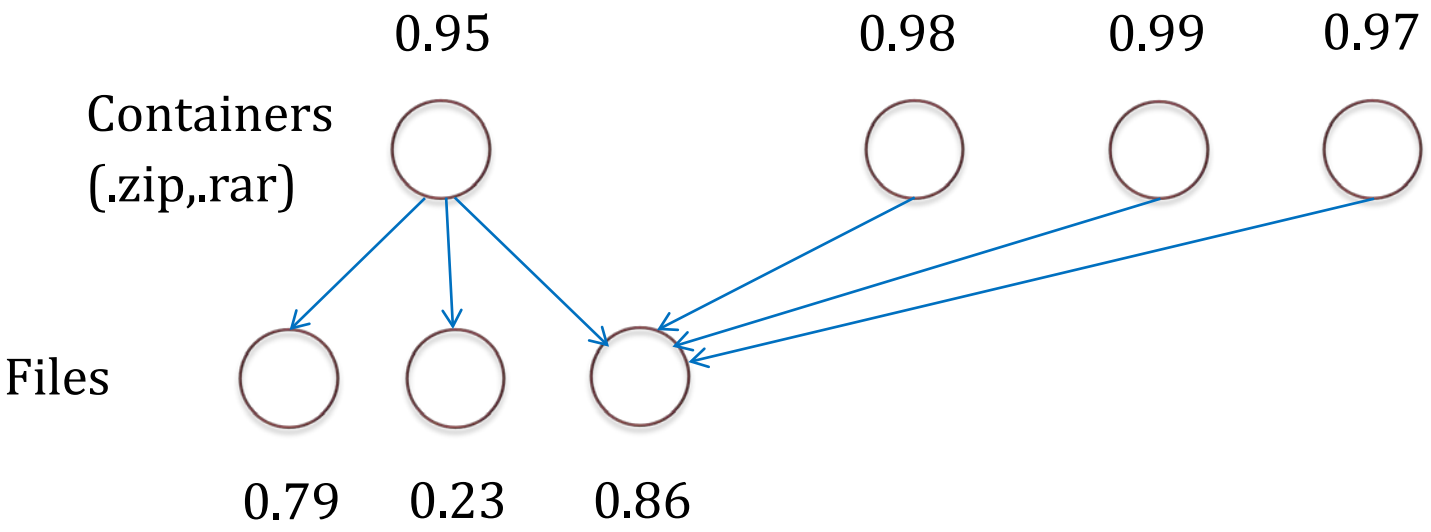
Container Classifier Training

- Bi-partite graph
  - 4.1M nodes, 24.0M edges
- 719K containers
  - 604K malicious, 115K benign
- 3.4M files
  - 482K malicious, 2.9M benign

Container Classification



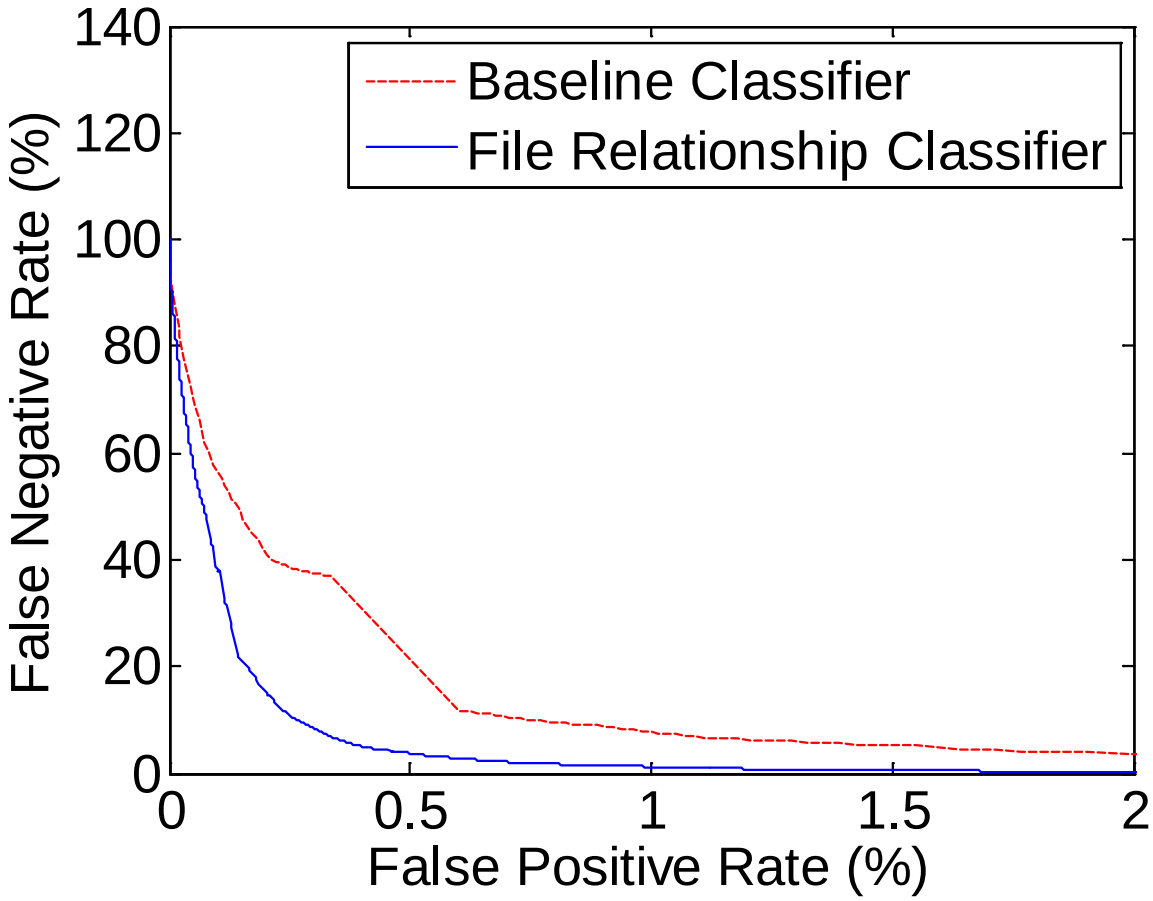
File Relationship Classification



File Relationship Classifier

- Biased Logistic Regression
  - $p_r(y_i = 1|x) = u^T x + b$
- Features
  - Histogram of container probabilities which include the file.
  - Features 2j and 2j+1: fraction and logarithm of the number of containers.
  - Histograms malicious and benign containers
  - $\log \frac{p_b}{1-p_b}$

Comparison of Baseline and File Relationship Classifier



2.exe Malware Example

Name	Determination	#Scanner Detections	#Submissions
. . . Norton Antivirus . . . 2007 .rar	Malware Container	15	2
. . . ba52.bin	Malware Container	15	4
. . . z0ffzvk .rar.part	Malware Container	14	2
. . . dc11.rar	Malware Container	14	2
. . . regcure 1.0.0.43.1.3a1400.efw	Malware Container	14	2
. . . Registry Mechanic . . . .rar	Malware Container	14	2
. . . CyberLink PowerDVD 7.0.rar	Malware Container	15	2

- Trojan variant in the Vundo family.
- Included in 8 containers labeled “Malware Container”
- Detected by at least 14 scanners.
- Baseline malware classifier failed to correctly identify the file as malicious
- The relationship classifier raised the probability 33% to 98.37%.
- Relationship classifier can help to correctly identify malicious files even when the baseline classifier misclassifies them

calleng.dll Benign Example

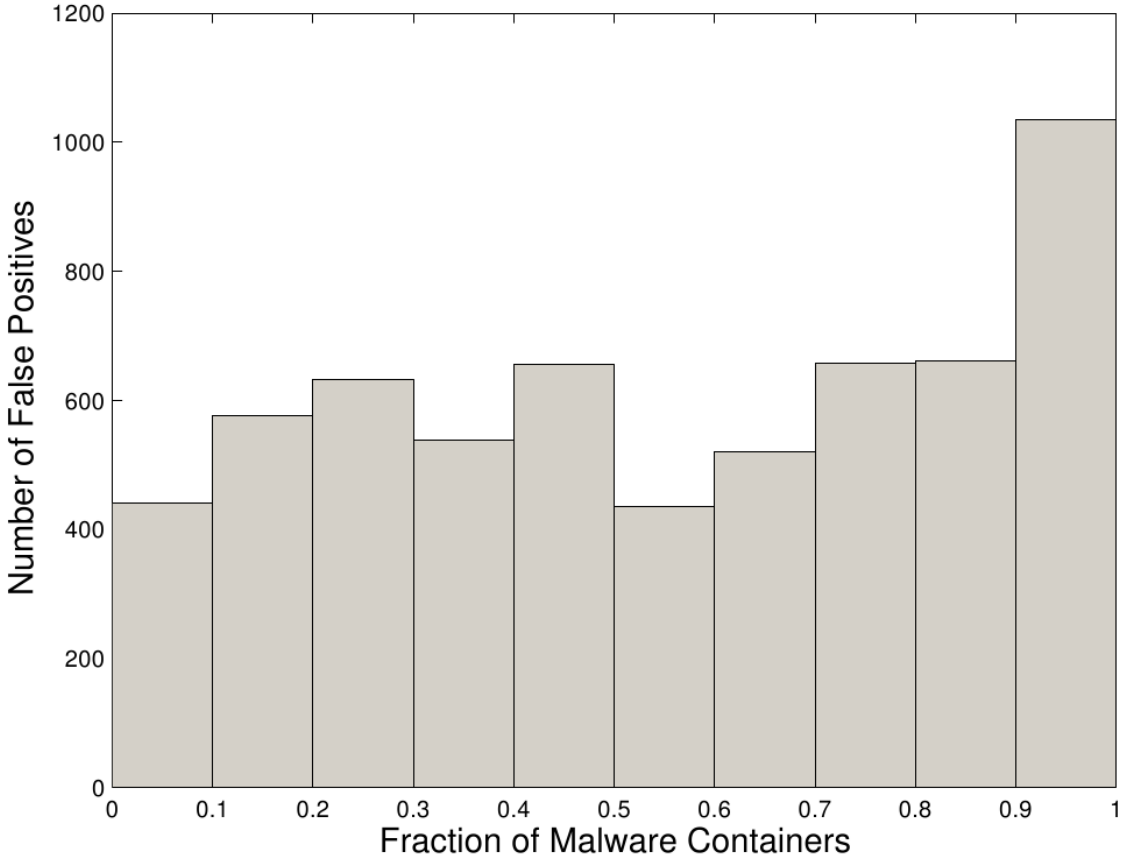
Name	Determination	#Scanner Detections	#Submissions
0d. . . bc.rar	No Determination	13	2
d3. . . 39.rar	No Determination	9	2
ec. . . da	No Determination	3	2
(RarSfx)	No Determination	0	2
(RarSfx)	No Determination	7	4
(RarSfx)	No Determination	9	4

- Manually determined to be benign
- Baseline malware classifier
  - 0% that this file is malware
- Originally distributed as part of the legitimate social networking software
- (RarSfx) on row 4 with no detections is the legitimate PalTalk.
- While calleng.dll itself is not malicious
  - Appears to be commonly used by malware authors in some manner
- After running the relationship classifier on calleng.dll the malware probability increased to 44.9%
- Not sufficient to be classified as malware

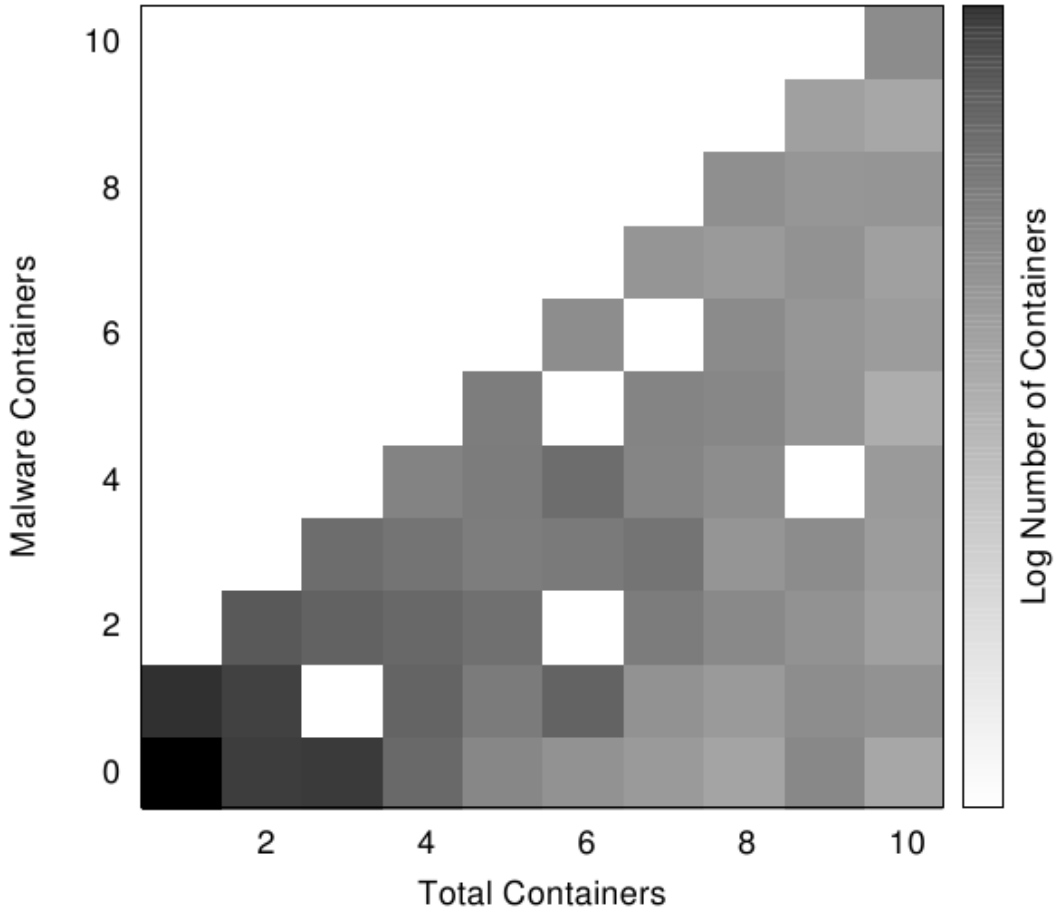
File Relationship Classification

FP Rate	Label	$p_b \leq t_b$ $p_r \leq t_r$	$p_b > t_b$ $p_r \leq t_r$	$p_b \leq t_b$ $p_r > t_r$	$p_b > t_b$ $p_r > t_r$
1.0%	Malware Benign	6269 2,909,583	161 15,561	32,170 14,590	480,548 14,959
0.5%	Malware Benign	183,454 2,950,180	15,406 1,556	109,043 1,546	211,245 1,411

False Positive Histogram



Heatmap of Containers Associated with False Positives



Summary

- Container relationship can improve an individual file malware classifier
- Biased logistic regression leads to good container classification
- Improved relationship classification
  - Better FN rates at low FP rates
- Orthogonal to baseline classifier
- Improvements in the baseline classification
  - Lead to improvements in the classification of files in containers