Monthly Research

Consideration for indicators of malware likeness based on static file information

Junichi Murakami

FFRI, Inc
http://www.ffri.jp

Ver2.00.01

Agenda

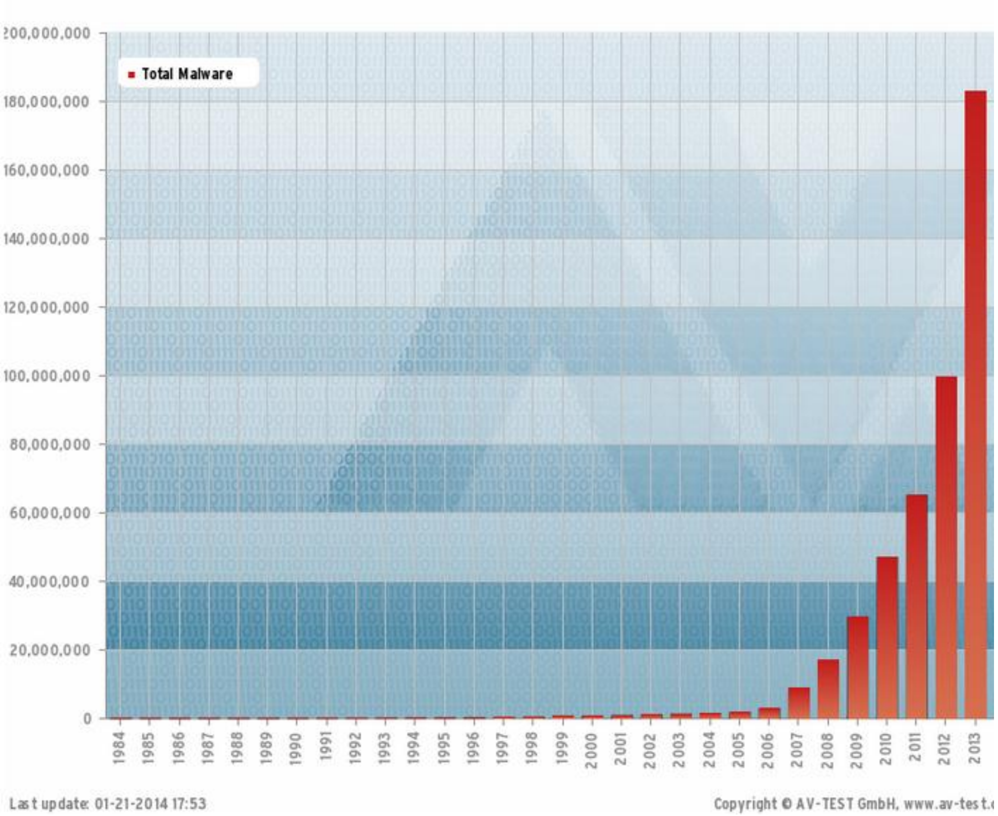
- Background and purpose
- An experiment
- The result
- Evaluation and consideration
- Conclusions

Background and purpose(1/3)

- Traditional signature matching is getting harder to detect malware due to dramatic increase of malware
- Therefore, signature-less(zero knowledge-based) detection is demanding
- Static heuristic detection is proposed and implemented as one of the method
- Most of the detection mechanisms are developed based on knowledge of experts like malware analyst
- In this slides, we consider a way to develop detection logic based on numerical indicators using regression analysis
- We summarize the overview, the steps, and the aspects of the evaluation

Background and purpose(2/3)

- In recent years, malware has been dramatically increased(Jan 2014)



http://www.av-test.org/en/statistics/malware/

Background and purpose(3/3)

- Why we use regression analysis?
  - There are other methods which can be applied to detect malware
    - Decision tree, random forest, neural network, SVM, etc.
  - However, malware detection is an area of application in which errors are not permitted relatively
  - The matter of risk for errors caused by unknown data (False Positive)
  - Capability of iterative improvement, determining a cause and explanation are required
  - Regression analysis is a prospective method in terms of these requirements
    (IMHO, appropriate to R&D rather than implementing to detection logic)

An experiment

- The goals
  - To understand which variables are how effective to determine if a file is malware or not
  - To understand which combination of variables is appropriate
- Extracting 5,000 malware and goodware for each randomly from dataset which we collected
- Analyzing the files above by applying reported features in “Attributes of Malicious Files”
  - (SANS Institute InfoSec Reading Room) https://www.sans.org/reading-room/whitepapers/malicious/attributes-malicious-files-33979
- Applying logistic analysis(LR) for the analysis above
- Using following tools:
  - R 3.0.2, python, pefile-1.2.10-139 (http://code.google.com/p/pefile/)

Overview of “Attributes of Malicious Files” (1/2)

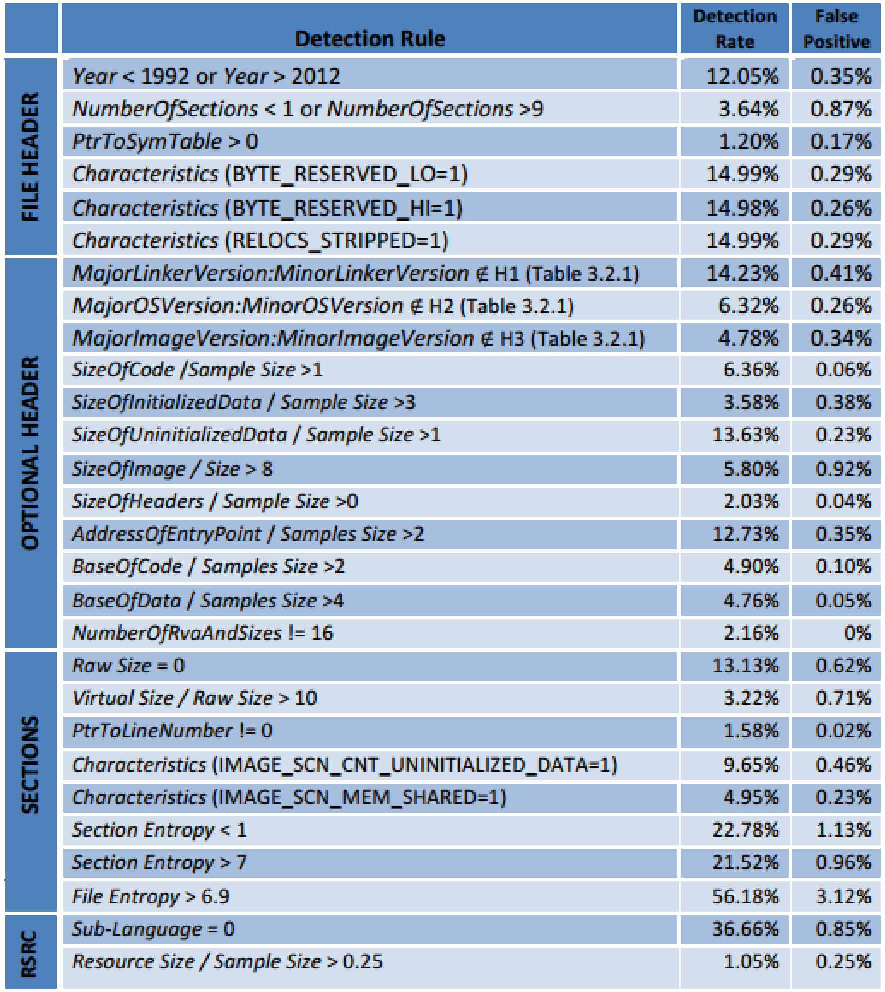
- Sampling 2.5M malware and 65,000 goodware
- Examining trends of various field values in PE header and reporting following information
  - Trends of field values which are appeared in malware frequently
  - Detection rules based on the trends above
  - The results of TPR/FPR by applying the rules to the samples
- Ex.)
  - There are malware whose TimeDateStamp in PE header is manipulated intentionally by setting before 1992 or a date of future(#)
  - Making detection rules based on those facts and the result of the evaluation is as below

Year	# of matched godoware	# of matched malware	diff.
< 1992	0.01%	11.72%	11.71%
1992-2012	99.98%	87.93%	-
>2012	0.00%	0.35%	0.35%

#The report is published in 2012

Overview of 「Attributes of Malicious Files」(2/2)

- Proposing 28 rules in the conclusions as the right table
- They are evaluated on individually and combination of them are not mentioned



	Detection Rule	Detection Rate	False Positive
FILE HEADER	*Year* < 1992 or *Year* > 2012	12.05%	0.35%
	NumberOfSections < 1 or *NumberOfSections* >9	3.64%	0.87%
	PtrToSymTable > 0	1.20%	0.17%
	Characteristics (BYTE_RESERVED_LO=1)	14.99%	0.29%
	Characteristics (BYTE_RESERVED_HI=1)	14.98%	0.26%
	Characteristics (RELOCS_STRIPPED=1)	14.99%	0.29%
OPTIONAL HEADER	*MajorLinkerVersion:MinorLinkerVersion* ∉ H1 (Table 3.2.1)	14.23%	0.41%
	MajorOSVersion:MinorOSVersion ∉ H2 (Table 3.2.1)	6.32%	0.26%
	MajorImageVersion:MinorImageVersion ∉ H3 (Table 3.2.1)	4.78%	0.34%
	SizeOfCode /Sample Size >1	6.36%	0.06%
	SizeOfInitializedData / Sample Size >3	3.58%	0.38%
	SizeOfUninitializedData / Sample Size >1	13.63%	0.23%
	SizeOfImage / Size > 8	5.80%	0.92%
	SizeOfHeaders / Sample Size >0	2.03%	0.04%
	AddressOfEntryPoint / Samples Size >2	12.73%	0.35%
	BaseOfCode / Samples Size >2	4.90%	0.10%
	BaseOfData / Samples Size >4	4.76%	0.05%
	NumberOfRvaAndSizes != 16	2.16%	0%
SECTIONS	*Raw Size* = 0	13.13%	0.62%
	Virtual Size / Raw Size > 10	3.22%	0.71%
	PtrToLineNumber != 0	1.58%	0.02%
	Characteristics (IMAGE_SCN_CNT_UNINITIALIZED_DATA=1)	9.65%	0.46%
	Characteristics (IMAGE_SCN_MEM_SHARED=1)	4.95%	0.23%
	Section Entropy < 1	22.78%	1.13%
	Section Entropy > 7	21.52%	0.96%
	File Entropy > 6.9	56.18%	3.12%
RSRC	*Sub-Language* = 0	36.66%	0.85%
	Resource Size / Sample Size > 0.25	1.05%	0.25%

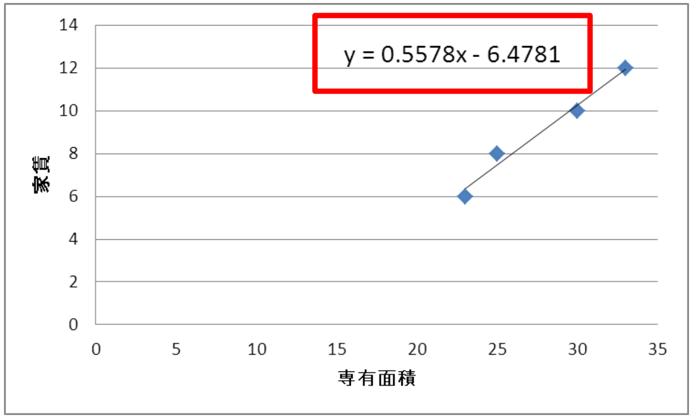
https://www.sans.org/reading-room/whitepapers/malicious/attributes-malicious-files-33979

Overview of regression analysis

- Statistically estimating relationships between a dependent value and independent values

- Ex.) estimating relationships between a rent and floor space according to following data and determining coefficients and a intercept in “y = ax + b”

  - data1:$600USD, 23㎡
  - data2:$800USD, 25㎡
  - data3:$1,000USD, 30㎡
  - data4:$1,200USD, 33㎡
    # in Japanese standard



- A method for multiple independent values and nonparametric estimation also exists

Overview of Logistic Regression(LR)

- One of the methods for nonparametric estimation

- Basically used when a dependent value is qualitative
  - Ex.)predicting if a man get cancer based on various tests
  - dependent value : become cancer(1) or not (0)
  - independent values : resuls of test-1, test-2, test-N

- By applying the same approach, we predict if files are malicious using values and rules introduced in the report
  - dependent value: malware(1) or not(0)
  - independent values: field values in PE header

Consideration for LR

- Preparation
  - selection of independent values
    - basically, selected based on knowledge of experts
    - following the report in this case
  - Data manipulation
    - the same as above
- Analysis
  - appropriate combinations of variables
  - interaction
    - an efffectiveness of X1 against Y is different depending on X2
    - just ignoring it for convenience this time
- Evaluation
  - Statistical significance
  - Odd ratio and its confidence interval
  - Goodness-of-fit
  - Model evaluation

Data manipulation

- Very important in regression analysis
  - Ex.) guessing a vector of age（11, 20, 25, 33, 60, 42)
    - Using as immediate (11, 20, 25, 33, 60, 42)
    - Round off by generations（10, 20, 20, 30, 60, 40)
    - If greater than 40 or not (0, 0, 0, 0, 1, 1)

- In general, nobody knows what conversion is appropriate
  - An accumulation of knowledge in a long range, never published in public
  - Appropriate method is different in each applied domain

- This time converting binary values(0 or 1) according to detection rule in the report (converting to 'dummy values')
  - Not matched: 0
  - Matched: 1

Selections and combinations for variables

- First of all, giving all variables and exploring a suitable combination using stepwise method (sequential variable selection)
  - using a step() function on R

- Indicator of goodness of models
  - AIC(Akaike’s Information Criterion)
  - It indicates goodness of models
  - An indicator of if a model is overfitting to target data
  - Less score means a better model
    - http://en.wikipedia.org/wiki/Akaike_information_criterion

Statistical significance / Odds ratio and its Confidence Interval

- Statistical significance(p-value)
  - The probability that the result is accidental
  - In general, under 5% means it is significant

- Odds ratio(OR)
  - An indicator which represents strength of relationship between a dependent value and independent values
  - In general it can be considered it is significant if "> 1.0"
  - OR contains errors and is dealt along with Confidence Interval(CI)
    - “95% CI” means that a value resides in an expressed range of score with 95% confidence

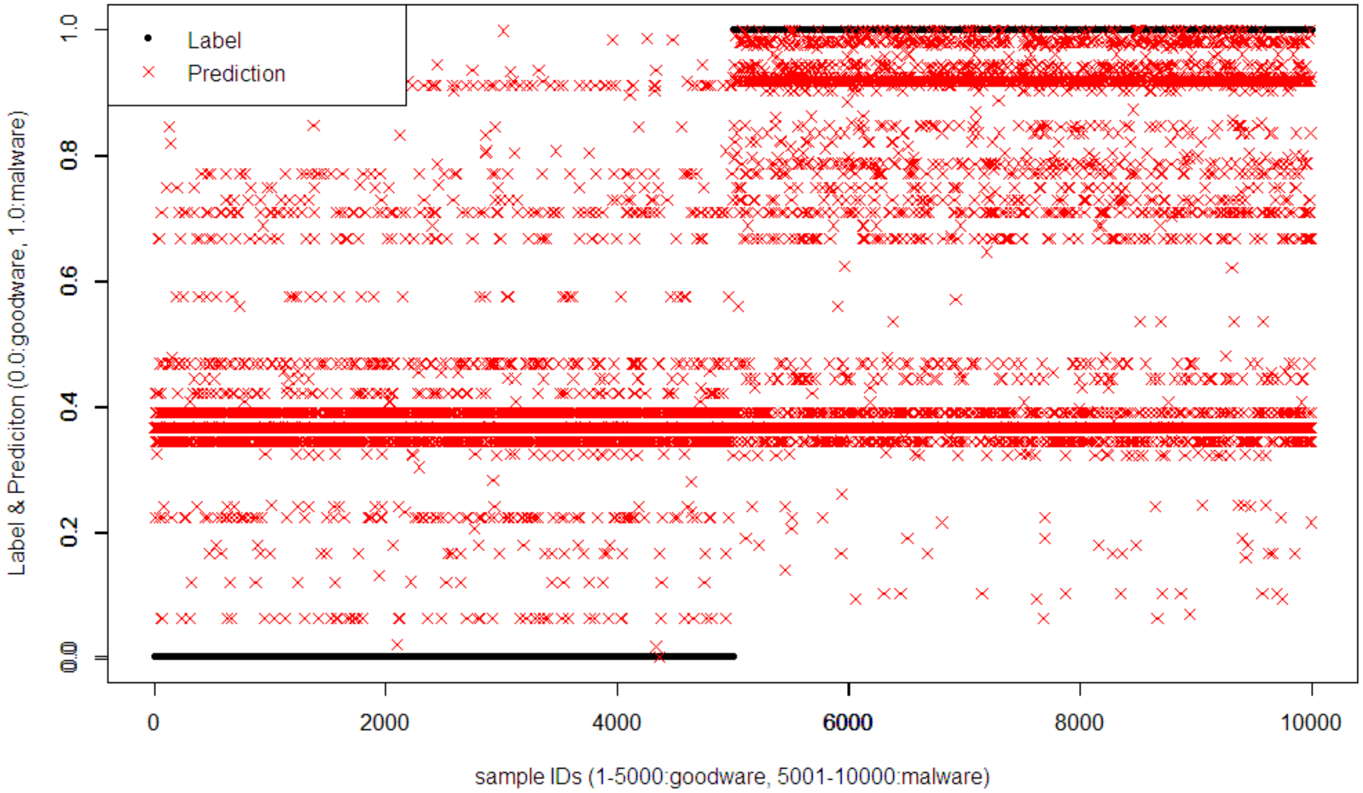
The result(1/2)

- Extracting top 3 variables in terms of p-value
  - Focusing on matched p-values in comparison with "not-matched" values
  - Rules of TimeDateStamp and SECTION_entropy are significant
  - Rule of ImageVersion is not significant since p-value is under 1.0

independent value / assigned value		OR（95% CI)	p-value
TimeDateStamp	0	(Reference)	-
	1	19.5 (16.1 - 23.9)	<2E-16
SECTION_entropy	0	(Reference)	-
	1	4.18(3.48 - 5.05)	<2E-16
ImageVersion	0	(Reference)	-
	1	0.174(0.123 – 0.241)	<2E-16

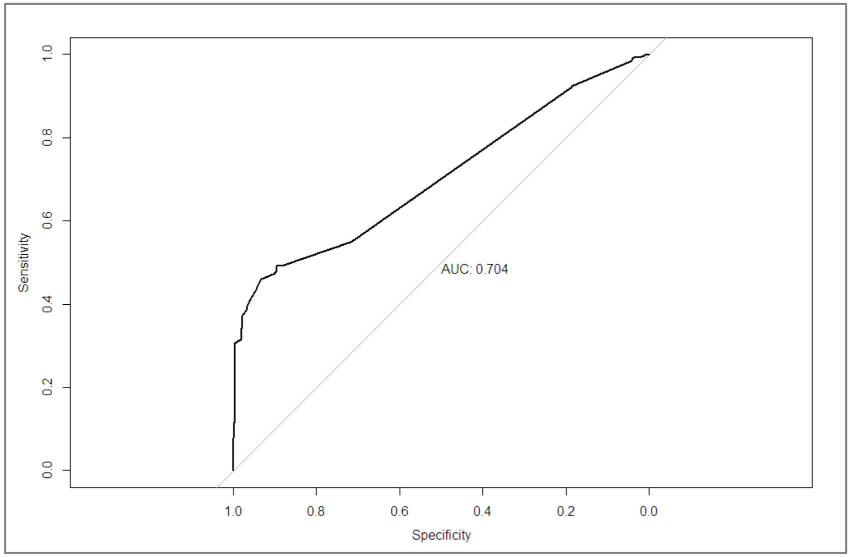
The result(2/2)

- Comparing correct labels and predictions
  - x-axis : sample IDs(1-5,000:goodware, 5,001-10,000:malware)
  - y-axis : goodware and malware likeness (0.0:goodware, 1.0:malware)



Evaluation and Consideration (1/2)

- We can understand significance of the others whose p-values are under 5% by checking those OR
- Variables whose p-value under 5% and OR is under 1.0
  have to be considered to be removed or changed the manipulation rules
- Goodness-of-fit
  - The Indicator that how well a model fit to target data
  - AUC(Area Under the Curve)
    - represented by 0.0 – 1.0
    - complete match:1.0
    - classified randomly: 0.5
    - the result: 0.704



ROC curve and AUC

Evaluation and Consideration (2/2)

- Model evaluation
  - Evaluating a validity of model using target data(Internal validity)
  - Using non-target data(External validity)

- In this case, we evaluate only internal validity using K-fold cross validation
  - Dividing all data into 13 chunk sets
  - Using 12 chunks for building a model and the rest is used for evaluation
  - Carrying out all of 13 combinations in this matter
  - Calculating prediction error of a model
  - The result : 19.8%(prediction error)

- During tuning a model, it is important to check if indicators like goodness-of-fit and prediction error are improved

Conclusions

- To aim to R&D of static heuristic detection, we focus on static file information between goodware and malware
- By using various information in PE header as variables of logistic regression, we can understand which variables and what combinations of them is how much effective quantitatively
- We can improve detection logic iteratively based on those indicators

Contact Information

- E-Mail: research-feedback@ffri.jp
- twitter: @FFRI_Research