Conference on
Information Technology and
Data Science

On the Time Series of Antivirus Testing Procedures

László Bognár, Antal Joós, Bálint Nagy

University of Dunaújváros
`bognarl@uniduna.hu`
`joosa@uniduna.hu`
`nagyb@uniduna.hu`

Proceedings of the 1st Conference on Information Technology and Data Science
Debrecen, Hungary, November 6–8, 2020
published at http://ceur-ws.org

Abstract

In this work, a so-called "Time Evolution Model" is suggested to help categorize files of a sample set used in antivirus testing procedures. The basic time-dependent variable of this model is the *Ratio* of the *Infected* files within an investigated *Time Window*. To estimate the main characteristics of the time series describing the change of the *Ratio* values related to a specific file, a nonlinear, exponential curve fitting method is used. The free parameters of the model were determined by numerical searching algorithms. The effectiveness and the reliability of the model is also demonstrated by several real-word and numerically simulated examples.

Keywords: Vulnerability, probability, relative frequency

AMS Subject Classification: 60A99 62M10 62D05

1. Introduction

As business and people rely more and more on computer related devices (including smart devices and the IoT), they are increasingly vulnerable to cyber-attacks [3, 15]. These attacks include threats of social networks [6] data phishing, malicious programs [17], etc. The defense against malware is composed of malware detectors, systems that investigate malicious objects (mainly files and URLs). Several malware detection techniques and methods to investigate the vulnerability of systems are introduced in the literature [10].

Security solution testers use malicious files (sample set) coming from different sources to determine whether the defense is able to detect these files as malicious or not [2, 7, 9, 13, 15].

One of the most important parts of the testing procedure influencing the reliability of the procedure is the correct and relevant selection of the used sample set.

How to correctly classify samples of a sample set is one the major issues for security solution testers to ensure their service to be reliable and to be able to give relevant recommendations for their client about the capabilities of security solutions [5]. Evaluating the efficiency of different antiviruses (AV), different antivirus vendors or even testing the level of security in a corporation requires reliable information about the samples [1, 8].

Besides the main question whether a given object (file/URL) (abbreviated only file in what follows) in a sample set is "*Infected*" or "*Noninfected*" [4, 11], in case of the infected files the "freshness" of the infection is also an important issue. The starting time of operation of a malware is essential for categorizing the malware as "*New*" or "*Old*".

Sample selection can be broken down into three phases [16]:

- Collection
- Validation
- Classification

In this paper the classification phase is in focus, however some aspects of the validation phase are also incorporated. It is assumed that the collection was correct, and the sample consists of real-world, prevalent, fresh, diverse files collected independently.

The sample validation process essentially is series of tests to make sure that the sample is functional (has working malicious function). There are several methods trying to validate samples: reverse engineering, usage of automated tools or by using various specialized tools (e.g. sandboxes) to check file integrity or functionality.

Best practices show that validation is most valuable when it is based on sample functionality, but these methods are not applicable to all sample types and may need enormous efforts to pursue these kinds of activities on a daily basis with huge number of files.

In this paper a so-called "*Time Evolution Model*" is suggested to help categorize each file or even a whole sample set (also called as feed).

The basic time-dependent variable of this model is the *Ratio* of the "*Yes*" decisions to the question: Is this file infected? The answers come from the members of a set of antimalware where most of these members showed reliable operations in the past in malware detection.

After the appearance of a new malware it takes less or more time for the different antimalware to detect the fact of infection. Some antimalware is simply not able to recognize some specific infections. (Possibly due to some validation issue.) Hence the ratio of "*Yes*" decisions is gradually increasing in time and reaches the state when the increase and the variation of the *Ratio* value is small enough to establish this *Ratio* as the steady state value of the time evolution.

The main goal of this study is to establish the main characteristics of these time functions. A nonlinear curve fitting method is used to fit a smooth time function on the observed *Ratio* data to estimate the steady state value (called *Asymptote*), the *Start Time* (starting time of operation) and the *Slope* at the *Start Time* for each file in a feed. These parameters can be used later to classify a file belonging to a certain category ("*Old*", "*New*", "*Infected*", "*Noninfected*", etc.).

For this estimation past *Ratio* data within a *Time Window* are used. The *Time Window* ends at the moment of investigation ("Today") and goes back in time. Obviously, the length of time, how far the *Time Window* goes back, has influence on the estimation. It is also investigated.

The reliability of the "Yes" decisions of the antiviruses is crucial. In this study it is assumed that the antimalware set consists of properly selected members. The process of selection resulting in a reliable set is discussed elsewhere.

2. General Features of the Time History of Malware Detections

As an example, in Table 1 the results of classifications for a sample of files are summarized.

Table 1. Results of Files Classifications and the *Ratio* Values.

Date	File name	AV 1	AV 2	AV 3	...	AV 100	#AV	#Yes	*Ratio*
01.03	File 1	*	No	Yes		Yes	97	34	0.35
01.03	File 2	No	No	Yes		No	93	43	0.462
...									
01.03	File 1000	Yes	Yes	Yes		Yes	96	44	0.463
01.04	File 1	Yes	No	*		Yes	94	37	0.35
01.04	File 2	Yes	No	Yes		No	95	44	0.463
...									
01.04	File 1000	No	Yes	Yes		Yes	90	37	0.411
...									
01.19	File 1	Yes	Yes	*		No	89	42	0.472
01.19	File 2	Yes	*	Yes		No	98	50	0.510
...									
01.19	File 1000	Yes	Yes	*		Yes	99	51	0.515

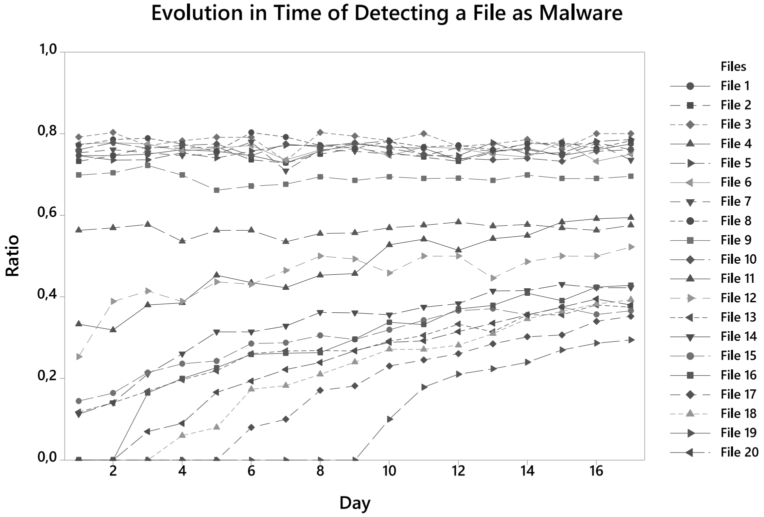

Figure 1. Time Series Graphs for the *Ratio* Values for Different Files in Different Phases of Antimalware Detection.

The cells of the table show the Yes/No results of classifications for $N_{files} = 1000$ different files by $N_{AV} = 100$ antiviruses for 17 days. Typically, not all files are checked by all AV on all days, so some cells contain no data. In the *Ratio* column the ratio of the number of Yes -es ($\#Yes$) and the number of nonempty cells ($\#AV$) in the given row is calculated.

The time evolution of the *Ratio* variables is better representable by time series graphs. In Figure 1 typical graphs for different files are shown where the different files are in different phases of antimalware detection. In Figure 2 those files are selected which can be considered as "*Old*" infected files.

Here the values of the *Ratio* variable show little fluctuation around different "imaginary", almost horizontal lines for the different files. This steady state feature forecasts not much change in the future, so these steady state *Ratio* values can be the basis for classifying the different files. Depending on the purpose of classification different threshold values can be decided in advance to categorize the files. If only two categories are used, above some predefined $Ratio_{Infected}$ value (e.g. $Ratio_{Infected} = 0.7$) a file can be taken as an "*Infected*" file, otherwise as "*Noninfected*". Sometimes three categories are better to use: "*Infected*", "*Noninfected*", "*Gray*". In this case two threshold values $Ratio_{Infected}$ and $Ratio_{Gray}$ (e.g. $Ratio_{Infected} = 0.7$ and $Ratio_{Gray} = 0.4$) can divide the zero-one interval into three different classification categories. The details for establishing these threshold values are not discussed here.

In Figure 3 and in Figure 4 the situations are different. Both figures suggest that the gradual increases of the *Ratio* values have not been finished, the infections are "*New*", they have been detected recently. These forecast additional increases in *Ratio* values, so the steady state *Ratio* values are in questions. In Figure 3 the graphs suggest the *Start Time*-s of the infection outside the *Time Window* while the graphs in Figure 4 suggest them inside.

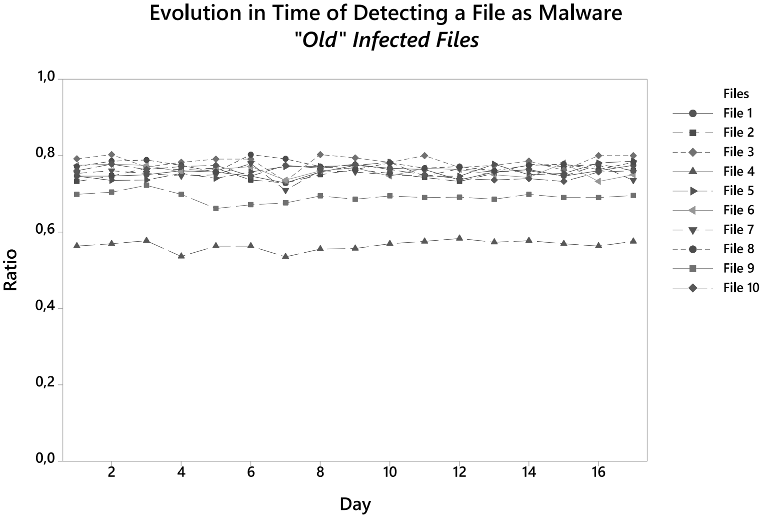

Figure 2. Time Series Graphs for the *Ratio* Values for "*Old*" Infected Files.

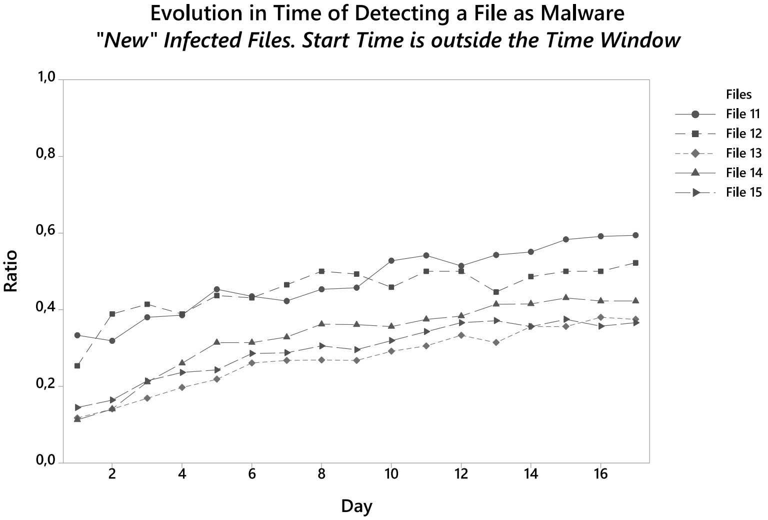

Figure 3. Time Series Graphs for the *Ratio* Values for "*New*" Infected Files. The *Start Time*-s are outside the *Time Window.*

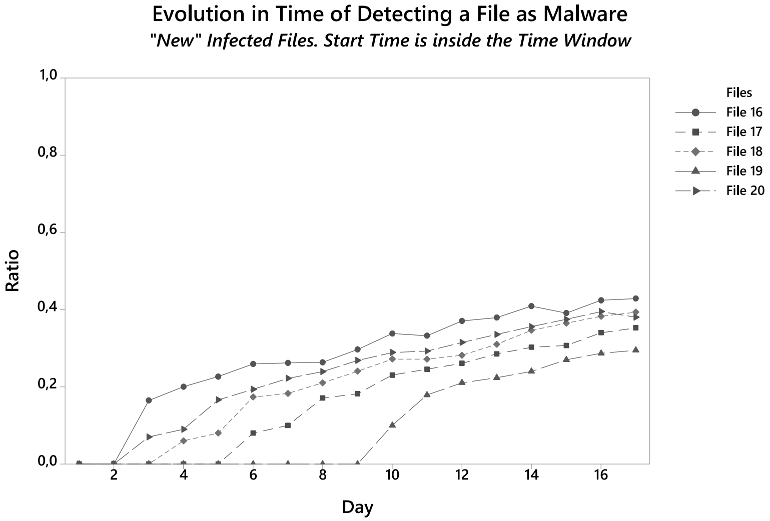

Figure 4. Time Series Graphs for the *Ratio* Values for "*New*" Infected Files. The *Start Time*-s are inside the *Time Window*.

3. The Time Evolution Model

To estimate the main characteristics of the time series describing the change of the *Ratio* values related to a specific file, nonlinear curve fitting method is used. For each file a theoretical time function is fitted to the observed *Ratio* values. In Figure 5 the notations are summarized.

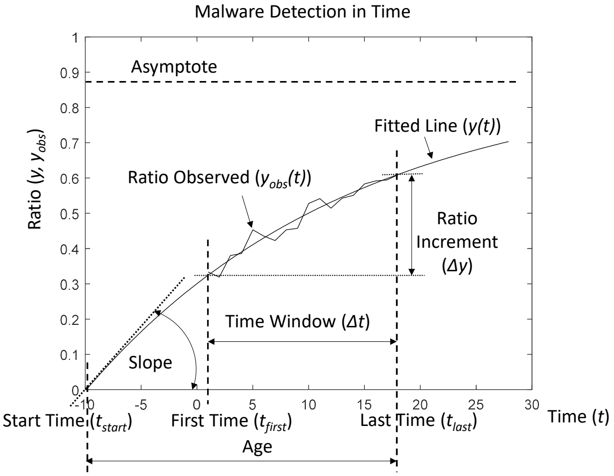

Figure 5. Notations in the Time Evolution Model.

In what follows $y_{obs}(t)$ or simply y_{obs} denotes the observed *Ratio* values in time. (*Sometimes the more precise $y_{obs}(t_i)$ is used since the observations are in discrete t_i time instants.*) The function $y(t)$ or simply y is the fitted function to be determined as the best fitted function to the observed y_{obs} values. Hence

$$y_{obs}(t) = y(t) + \epsilon(t) \tag{3.1}$$

where $\epsilon(t)$ is a random error term. The function $y(t)$ is searched in an exponential form widely used in different growth models [12, 14],

$$y(t) = \alpha_1(1 - e^{-\alpha_2(t-\alpha_3)})$$

where α_1, α_2 and α_3 are free parameters to be determined. The method of least squares can be used to determine the α_1, α_2, α_3 free parameters. It requires the minimization of the criterion

$$Q = \sum_{i=1}^{n} [y_{obs}(t_i) - y(t_i)]^2$$

where n is the number of observations within the *Time Window.*

Unlike linear curve fitting, it is not possible to find analytical solution for the least squares, instead numerical search procedures must be used. In the examples presented here Matlab `https://www.mathworks.com/products/matlab.html` software's built in optimization procedures have been used.

In the optimization the constrains

$$\alpha_1 \leq 1 \text{ and } \alpha_2 \geq 0$$

were applied.

It is worth seeing that concrete "physical" meaning can be attributed to the three free parameters in the model. The α_1 value corresponds to the *Asymptote* shown in Figure 5 Depending on the α_1 value, a specific file can directly be classified in to the "*Infected*" or "*Noninfected*" category.

The parameter α_3 is the *Start Time* when the infection begins. Depending on the α_3 value, a specific file can directly be classified in to the "*Old*" or "*New*" category.

The *Slope* $= \alpha_1\alpha_2$ product gives the tangent of the angle at *Start Time.* This *Slope* value may refer to the extent (the speed) of spread of a specific infection at the beginning.

4. Examples for Curve Fitting

Several samples, several real-world files' time evolution models have been set up in the present study. In Figure 6 some of them are presented where the final results of the numerical search procedures for the $\alpha_1, \alpha_2, \alpha_3$ parameters are also shown.

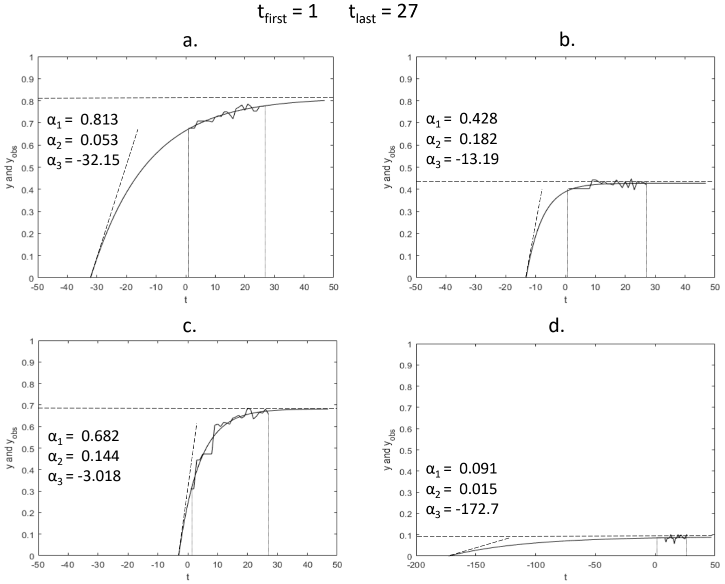

Figure 6. Examples for Curve Fitting in Different Detection Situations.

In all cases the present time ("Today") is denoted as $t_{last} = 27$. The *Time Window* goes back to $t_{first} = 1$ and the time scale goes beyond t_{first} and t_{last} to get some impression about the possible run of the curves in the past and in the future. The different panels of the figure depict different detection situations. In Figure 6b and in Figure 6d the graphs show more or less steady state situations where Figure 6d reveals a "*Noninfected*" case when the detection started long time before while the $\alpha_1 = 0.428$ value in Figure 6b refers to a rather uncertain case, probably better to classify it as a "*Gray*" situation.

In both, in Figure 6a and in Figure 6c the curves suggest further increases of y values in time in "*Infected*" situations. However, the observed y_{obs} values in Figure 6c suggest less uncertainty in the estimated $\alpha_1 = 0.682$ *Asymptote* and in the $\alpha_3 = -3.018$ *Start Time* values while the prediction for the $\alpha_1 = 0.813$ and the $\alpha_3 = -32.15$ values in Figure 6a are probably more uncertain.

5. Factors Influencing the Reliability of the Parameter Estimation

It is obvious that the extent of uncertainty in the predicted α_1, α_2 and α_3 parameters is dependent on the run of the observed y_{obs} values. More precisely they are influenced by the $\Delta(t)$ width, the location of the *Time Window* (relative to the *Time Start*), and the spread of the y_{obs} values (*this can be characterized by the*

Std(ϵ) *standard deviation in the model equation* (3.1)).

It needs further studies to establish confidence intervals for the α_1, α_2 and α_3 parameters. In this nonlinear curve fitting it is hopeless to get closed form analytical solutions for these intervals but numerical simulation studies covering the whole space of the influencing factors may help in determining them.

In practice the question arises in the form: What is the minimal size of the *Time Window* at a given file to get reliable estimates for the α parameters?

To get some impression about the effects of the influencing factors, in Figure 7, Figure 8 and Figure 9 the tendencies of the change in the α parameters are presented for two different files where the width of the *Time Window* is gradually widened.

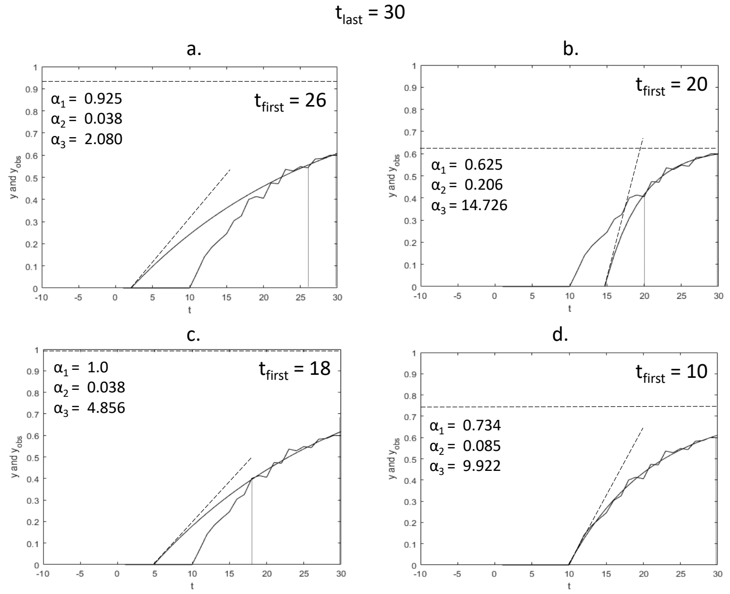

Figure 7. Malware Detection History in Time when the Operation of the Malware Commences Inside the *Time Window*.

In Figure 7 the y_{obs} curve depicts the detection history of such a file when the time of investigation ("Today") is at $t_{last} = 30$ and the operation of the malware commences at $t_{start} = 10$. In Figure 7d where the *Time Window* goes back in time to the malware's start of operation good fit can be seen within the whole *Time Window*. Hence the *Start Time* $= \alpha_3 = 9.922$ and the *Slope* $= \alpha_1\alpha_2 = 0.734 \cdot 0.085 = 0.062$ values can be regarded as good estimates. The value of *Asymptote* $= \alpha_1 = 0.731$ holds more uncertainty, however, this $\Delta(t) = 20$ wide *Time Window* seems to be wide enough to produce a reliable *Asymptote* value to classify the file as being "*Infected*".

In Figure 7a-c the fluctuations of the α parameter values can be seen as the t_{first} value approaches to the *Time Start* value.

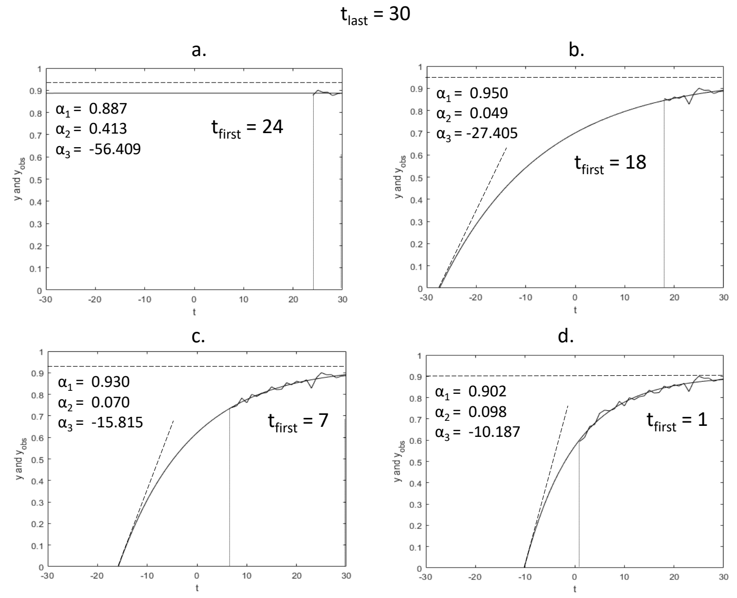

Figure 8. Simulated Malware Detection History in Time when the Operation of the Malware Commences Outside the *Time Window*.

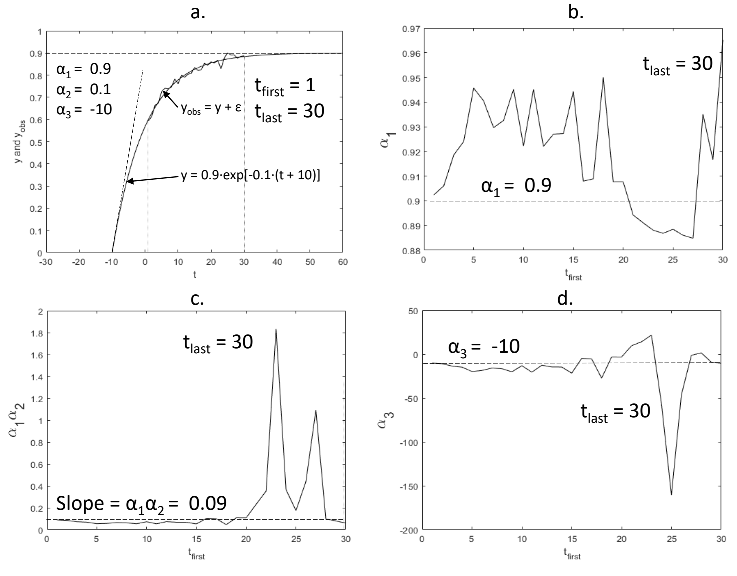

Figure 9. The Change of the α_1, α_2 and α_3 Parameters as the Width of the *Time Window* is Changing.

In Figure 8 and in Figure 9 not a real world but a numerically simulated y_{obs} curve is used to illustrate the situation when the *Start Time* is outside the available widest *Time Window*, so no chance to trace the α parameter values back to the *Start Time* value.

In Figure 9a the layout of the situation is summarized. The widest *Time Window* is $\Delta(t) = 30 - 1 = 29$ wide and within this window the y_{obs} values have been randomly simulated using the $y = 0.9e^{-0.1(t+10)}$ deterministic function and the normally distributed ϵ random errors with zero mean and $Std(\epsilon) = 0.015$ standard deviation.

In Figure 9 the y, y_{obs} curves and the numerically searched α parameter values are shown in case of four different *Time Windows*. It is worth seeing how far the three numerically searched α parameters are from the deterministic $\alpha_1 = 0.9, \alpha_2 = 0.1$ (*Slope* $= 0.9 \cdot 0.1 = 0.09$), $\alpha_3 = -10$ values.

In Figure 9 the change of the α parameters is shown as the width of the *Time Window* is gradually changing. It is worth following the tendencies of the change backwards along the t_{first} axis. In Figure 9c and in Figure 9d as the t_{first} value is decreasing from the $t_{last} = 30$ value (as the width of the *Time Window* is increasing) the estimated *Slope* and α_3 (*Start Time*) values are convincingly converging to their deterministic values. After some fluctuation when the width of the *Time Window* reaches the $\Delta(t) = 10$ values the fluctuation becomes almost negligible. In Figure 9b the situation is even better. The estimated α_1 parameter value (the *Asymptote*) is very close to the deterministic $\alpha_1 = 0.9$ value even in the very narrow *Time Window* cases. All these α_1 values would classify the given file as being "*Infected*".

6. Conclusion

In this paper a so called "Time Evolution Model" was suggested to help categorize files of a sample set in antimalware testing procedures. The basic time dependent variable of this model is the *Ratio* of the *Infected* files within an investigated *Time Window*. To estimate the main characteristics of the time series describing the change of the *Ratio* values related to a specific file, nonlinear, exponential curve fitting method was used. The free parameters of the model were determined by numerical searching algorithms, hence specific "physical" characteristics of the sample set were calculated.

The effectiveness and the reliability of the model was demonstrated by several real-word and numerically simulated examples.

Further studies are required to establish the extent of reliability of the model in the whole parameter space and to give general recommendations for the minimal size of the *Time Window* for reliable classifications.

References

[1] S. Brown, J. Gommers, O. Serrano: *From Cyber Security Information Sharing to Threat Management*, in: Proceedings of the 2nd ACM Workshop on Information Sharing and Collab-

orative Security, WISCS '15, Denver, Colorado, USA: Association for Computing Machinery, 2015, pp. 43–49, ISBN: 9781450338226,
DOI: 10.1145/2808128.2808133,
URL: https://doi.org/10.1145/2808128.2808133.

[2] I. Burguera, U. Zurutuza, S. Nadjm-Tehrani: *Crowdroid: Behavior-Based Malware Detection System for Android*, in: Oct. 2011, pp. 15–26, ISBN: 9781450310000,
DOI: 10.1145/2046614.2046619.

[3] K.-K. R. Choo: *The cyber threat landscape: Challenges and future research directions*, Computers and Security 30.8 (2011), pp. 719–731, ISSN: 0167-4048,
DOI: https://doi.org/10.1016/j.cose.2011.08.004,
URL: http://www.sciencedirect.com/science/article/pii/S0167404811001040.

[4] M. Christodorescu, S. Jha, S. Seshia, D. Song, R. Bryant: *Semantics-Aware Malware Detection*, in: June 2005, pp. 32–46, ISBN: 0-7695-2339-0,
DOI: 10.1109/SP.2005.20.

[5] Z. A. Collier, I. Linkov, J. H. Lambert: *Four domains of cybersecurity: a risk-based systems approach to cyber decisions*, English, Environment Systems and Decisions 33.4 (2013), pp. 469–470,
DOI: 10.1007/s10669-013-9484-z.

[6] W. Gharibi, M. Shaabi: *Cyber Threats In Social Networking Websites*, International Journal of Distributed and Parallel Systems 3 (Feb. 2012).

[7] K. Hadarics, F. Leitold: *Improving distributed vulnerability assessment model of cybersecurity*, Central and Eastern European eDem and eGov Days 331 (July 2018), pp. 385–393,
DOI: 10.24989/ocg.v331.32.

[8] W. Jansen: *Directions in Security Metrics Research* (Jan. 2010).

[9] F. Leitold: *Testing protections against web threats*, in: 2011 6th International Conference on Malicious and Unwanted Software, 2011, pp. 20–26,
DOI: 10.1109/MALWARE.2011.6112322.

[10] F. Leitold, K. Hadarics: *Measuring security risk in the cloud-enabled enterprise*, in: Oct. 2012, pp. 62–66, ISBN: 978-1-4673-4880-5,
DOI: 10.1109/MALWARE.2012.6461009.

[11] A. Moser, C. Kruegel, E. Kirda: *Limits of Static Analysis for Malware Detection*, in: Twenty-Third Annual Computer Security Applications Conference (ACSAC 2007), 2007, pp. 421–430,
DOI: 10.1109/ACSAC.2007.21.

[12] J. Murray: *Mathematical Biology I: An Introduction*, vol. 1, Jan. 2002.

[13] F. Osorio, F. Leitold, D. Mike, C. Pickard, S. Miladinov, A. Arrott: *Measuring the effectiveness of modern security products to detect and contain emerging threats — A consensus-based approach*, in: Oct. 2013, pp. 27–34, ISBN: 978-1-4799-2534-6,
DOI: 10.1109/MALWARE.2013.6703682.

[14] G. Serazzi, S. Zanero: *Computer Virus Propagation Models.* In: vol. 2965, Jan. 2003, pp. 26–50.

[15] Symantec: *Symantec Internet security threat report 2019*, https://www.symantec.com/content/dam/symantec/docs/reports/istr-24-2019-en.pdf/, [Online; accessed 2-Okt-2020], 2019.

[16] C. H. Tseng, S. Wang, S. Wang, T. Juang: *Proactive malware collection and classification system: How to collect and classify useful malware samples?*, in: 2014 International Conference on Information Science, Electronics and Electrical Engineering, vol. 3, 2014, pp. 1846–1849,
DOI: 10.1109/InfoSEEE.2014.6946241.

[17] H. Yin, D. Song, M. Egele, C. Kruegel, E. Kirda: *Panorama: Capturing system-wide information flow for malware detection and analysis*, in: Jan. 2007, pp. 116–127, doi: `10.1145/1315245.1315261`.