Time Evolution Model for Analysing Malicious Samples

Ferenc Leitold

Óbuda University, Bécsi út 96/b, H-1034 Budapest, Hungary
leitold.ferenc@nik.uni-obuda.hu

Abstract: In this paper, the results of the practical examination of the Time Evolution Model ([1] [2] [3]) used to categorize malicious samples are summarized. This method provides effective assistance in anti-malware testing procedures as well as cyberattack detection. With its help, the typical properties of malicious codes can be determined more easily and quickly with automatic tools. The Time Evolution Model can help security experts better understand the behavior of malicious attacks and malware families. The Time Evolution Model works based on variables describing changes in the detection capabilities of different protection systems related to a specific malicious file. An exponential curve fitting method is used to estimate the main characteristics of the spread of the malicious code. During the curve fitting, three parameters are determined, with the help of which the properties of the spread of a malware or a malware family can be identified more precisely. In the case of malicious program families, the aggregation of these parameters can be an effective solution for estimating cyberthreat trends. The Time Evolution Model was tested on a large number (more than 1000) of malicious samples, based on which different groups can be distinguished according to when the investigation starts after the first appearance of the malicious code.

Keywords: malware; threat intelligence; vulnerability metric; time evolution model

1 Introduction

The unique number of different malware (including malicious programs, computer viruses, phishing threats, etc.) increasing very rapidly today. It means that it results in a very difficult task for everybody who is dealing with the technical part of information security:

- Information security teams in different sizes of organizations can meet many and many different types of threats.
- Vendors of different protection systems have to work with this huge number of malware to provide efficient protection.

- Independent protection testers as well as QA testers inside the protection vendor organizations have to maintain the older and newer malicious samples too.
- Among the different types of actors in the security industry (vendors, researchers, testers, users) there is information sharing including malicious samples. They are usually regular sharing providing a continuous malware sample feed related to the newest threats.

In this paper, the practice of the Time Evolution Model ([1] [2] [3]) is demonstrated, which is an efficient tool to help categorize each file or a sample set (it can be, e.g. a malware family or a sample feed). The Time Evolution Model can help provide information using the detection capability results of threat protections. It means that using this technique the results of reverse engineering processes already executed by security vendors are indirectly used to extract malware-related information. The Time Evolution Model is based on regular iterations of the protection capability testing and results are concluded from the changes in the detection capabilities.

2 Related Works

Malware detection methods are introduced in the literature [4] [5] [14] [15]. The most effective way of malware protection is based on the “known malware”. It means that if the code of a particular malware is known by a protection vendor, then after the analysis of the code of the particular malware, the detection and remediation algorithm can be added to the protection database. There are plenty of tools for automatic analysis (static or dynamic) as well as for helping the manual reverse engineering process. However, reverse engineering is the most expensive way to extract detailed information from the code.

On the other hand, security protection testers use malicious files (sample set) coming from different sources to determine whether the protection can detect them as malicious or not and can remediate them resulting in the original environment [7] [8] [10] [16]. One of the most important parts of the testing procedure influencing its reliability is the correct and relevant selection of the sample set to be used [16]. On the other hand, security experts receive plenty of information about newer and newer malicious threats [9] [11] (e.g., from threat intelligence services). They have to prioritize their effort and they have to focus on the most dangerous threats and malware families to calculate the most valuable metrics [6] [12] [13].

How to correctly classify samples of a sample set or an incoming feed is one of the major issues for security testers and security experts. Besides the main question of whether a given object (file/URL) (abbreviated only file in what follows) is "Infected" or "Noninfected" in the 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" and it can help to determine how quickly can different malware protections updated against the particular malware (time-to-detect metric).

In the security field, it is especially important to understand the behavior of malicious attacks in the wild. The Time Evolution Model ([1] [2] [3]) described and demonstrated in this paper can help security experts understand much better the spreading mechanism of programmed attacks and malware families as well. This methodology can be used to provide more effective protection techniques against cyberthreats. Using the mentioned time evolution model can provide insights into understanding what are the most relevant threats in the wild that danger mostly the IT infrastructure. On the other hand, quality assurance and testing different protection solutions is one of the critical parts of the fight against cybersecurity threats. The methodology can provide valuable information for example in the following:

- Prioritizing the required protection tasks of an IT infrastructure.
- Selecting the relevant malicious sample set for protection testing.
- Calculating valuable behavioral information of a set of malicious samples (e.g.: malware family, threat intelligence feed).

Using the Time Evolution Model real malware samples were tested and analyzed extracting some basic information. The results of this analysis are summarized in this paper as well as validating the methodology of the Time Evolution Model.

3 Threat Intel Services used during the Analysis

During the analysis live malicious samples were used. Samples were downloaded from the AMTSO RTTL database and they were uploaded daily to the Virustotal service to determine the detection possibility of different protections.

3.1 AMTSO RTTL

The AMTSO RTTL (Real Time Threat List) is a tool managed by the Anti-Malware Testing Standards Organization (AMTSO), designed to provide a comprehensive, up-to-date list of current malware threats. AMTSO is an international organization that sets standards for testing anti-malware solutions, aiming to improve the objectivity, quality, and relevance of security testing.

The RTTL serves as a dynamic resource for the cybersecurity industry, offering real-time data on emerging threats. This includes information on malware samples, URLs, and other malicious indicators that are actively being used in attacks.

The goal of the RTTL is to aid in the development and testing of anti-malware products by providing access to a live feed of threat intelligence. This helps security vendors and researchers to better understand the threat landscape and to test their products against current and emerging threats effectively.

The RTTL is a collaborative effort, relying on contributions from AMTSO members, security vendors, researchers, and other stakeholders in the cybersecurity ecosystem. This collaborative approach ensures that the RTTL remains a valuable and relevant resource for those involved in malware detection and analysis, enabling them to stay ahead of the latest cybersecurity threats.

The samples in the AMTSO RTTL (Real Time Threat List) come from a variety of sources, reflecting the collaborative nature of the cybersecurity industry and the effort to maintain a comprehensive and up-to-date list of threats. These sources include:

- **Security Companies:** Major antivirus and security software vendors contribute samples to the RTTL. These companies have extensive detection and research teams dedicated to finding and analyzing new threats. Their contributions are a cornerstone of the RTTL, providing a vast array of data on malware, phishing sites, and other security threats.
- **Independent Researchers:** The cybersecurity community is vast and includes independent researchers and smaller security firms that often discover and share new threats. These contributors play a crucial role in diversifying the RTTL's samples, adding unique finds that might not be captured through larger networks.
- **Automated Systems:** Many organizations employ automated systems to collect and analyze malware samples. These systems can include honeypots, sandboxes, and other types of cybersecurity detection tools that automatically identify and categorize new threats as they emerge.
- **Threat Intelligence Platforms:** The RTTL also integrates data from threat intelligence platforms and feeds that aggregate information on known threats. These platforms compile data from various sources, including law enforcement takedowns, cybersecurity research papers, and user submissions.
- **Community Submissions:** An often-overlooked source is the community at large, which can include anyone from IT professionals to end-users who encounter and submit samples of malware or other malicious content they come across in their daily activities.

The AMTSO RTTL's effectiveness lies in its ability to aggregate data from these diverse sources, ensuring that it covers as broad a spectrum of threats as possible. This inclusivity is crucial for developing and testing security solutions that can protect users from the latest malware and cyberattacks.

3.2 Virustotal

VirusTotal is a free online service that allows users to analyze suspicious files and URLs to detect types of malware and automatically share them with the security community. It aggregates multiple antivirus engines, website scanners, file and URL analysis tools, and user contributions to provide a comprehensive view of the security level of the file or webpage in question.

VirusTotal scans submitted files and URLs using over 70 different antivirus scanners and URL/domain blacklisting services, providing a wide spectrum of detection capabilities. Users can submit files up to a certain size limit and URLs for scanning. The platform then provides detailed reports on the detected threats. On the other hand, detected malware samples are shared with the security community, contributing to the overall improvement of global security defenses. VirusTotal offers an API that allows developers to integrate its features into their own applications, enabling automated file and URL scanning within their services. Users can comment on and rate URLs and files, offering insights or warnings about potential false positives or overlooked malicious content.

Researchers and cybersecurity professionals use VirusTotal for quick assessments of potential threats. Regular users can check the safety of files or links they receive before opening or visiting them, reducing the risk of infection. Security experts and researchers analyze malware samples and campaigns by examining detailed reports and behavior analysis provided by VirusTotal.

Due to the broad range of antivirus engines used, there can be false positive detections. Users are advised to consider the context and other evidence before determining the maliciousness of a file or URL. VirusTotal complements antivirus solutions, but should not replace dedicated endpoint protection, as it does not offer real-time scanning or removal capabilities.

VirusTotal is widely respected and used both by security professionals and the general public for its comprehensive analysis capabilities and its role in facilitating a collaborative approach to cyberdefense.

4 Model Analysis

1007 randomly selected freshly uploaded samples were used from the AMTSO RTTL database [16]. The used samples were expected to be malicious, but it is not proven; however, it is unknown as well when and how fast these samples started to breed. During 54 days the VirusTotal service was used to determine how many and which antimalware protections were able to detect the particular malicious samples. All of the data were collected during the test, but according to some internet access or service availability problem results were not collected on all of the days.

However, the number of successful query processes was in the range from 46 to 50 days for each sample. These data were analyzed after the testing procedure. The Table 1 and Figure 1 show the histogram of the successful queries.

Table 1
Successful daily collections

Successful daily collections	Number of samples
46	4
47	6
48	172
49	615
50	210

Figure 1
Successful daily collections

4.1 Different Curve Types

During the analysis, different types of fitted curves provide the possibility to visually classify the malicious samples.

4.1.1 Examination Started a Few Days after the First Appearance

In these cases, the graphs initially show a large increase until it reaches a near-steady state. (Figures 2, 3, 4)

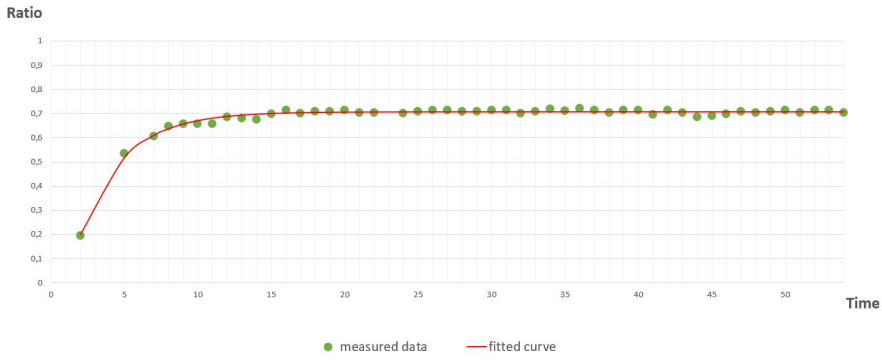

Figure 2
id = 10447; α1 (asymptote) = 0,71; α2 = 0,33; α3 (start time) = 1,00

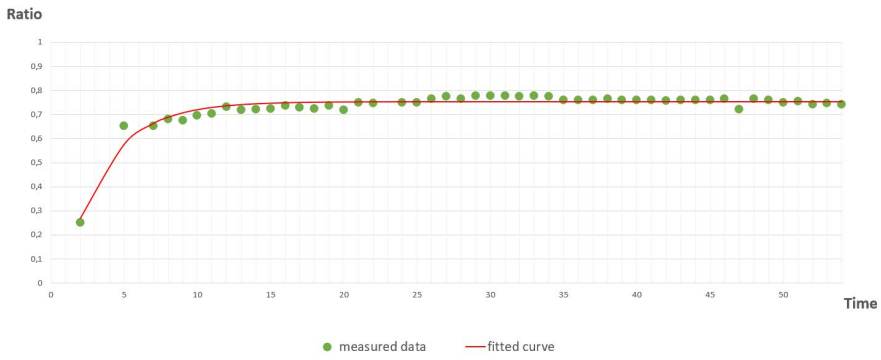

Figure 3
id = 10039; α1 (asymptote) = 0,75; α2 = 0,34; α3 (start time) = 0,69

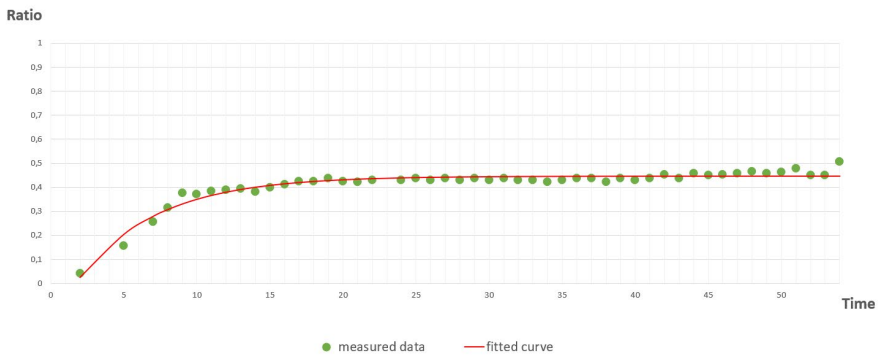

Figure 4
id = 10099; α1 (asymptote) = 0,45; α2 = 0,18; α3 (start time) = 1,68

4.1.2 Examination Started about 5-15 Days after the First Appearance

In these cases, the graphs initially show a small increase until it reaches a near-steady state. (Figures 5, 6, 7)

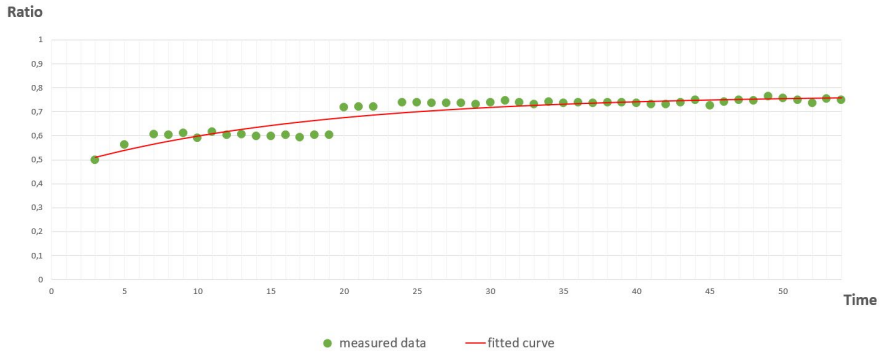

Figure 5
id = 10923; α1 (asymptote) = 0,77; α2 = 0,06; α3 (start time) = -15,28

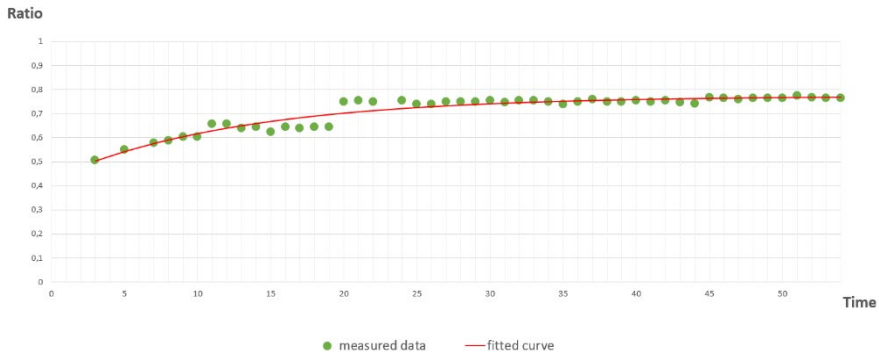

Figure 6
id = 11106; α1 (asymptote) = 0,77; α2 = 0,08; α3 (start time) = -10,37

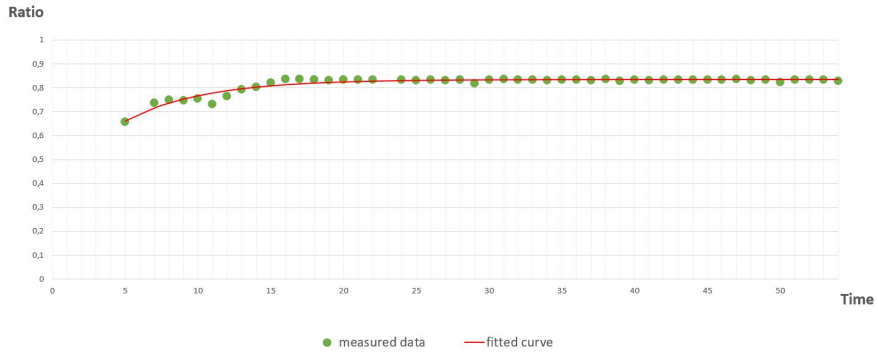

Figure 7
id = 10386; α1 (asymptote) = 0,83; α2 = 0,19; α3 (start time) = -3,37

4.1.3 Examination Started at Least 15 Days after the First Appearance

In these cases, we can initially see a very small increase on the graphs, while in the following tests, it already reaches a near-steady state. (Figures 8, 9, 10)

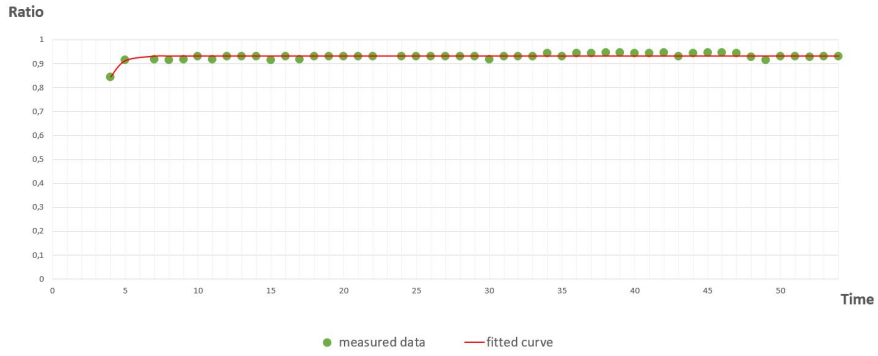

Figure 8
id = 10350; α1 (asymptote) = 0,93; α2 = 1,50; α3 (start time) = 2,41

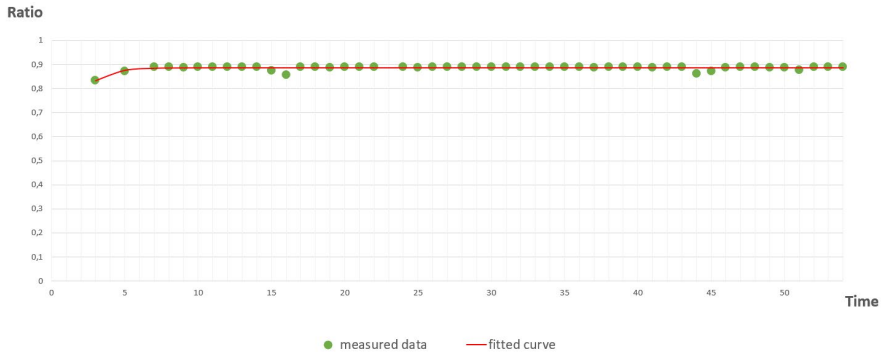

Figure 9
id = 11538; α1 (asymptote) = 0,89; α2 = 0,81; α3 (start time) = -0,46

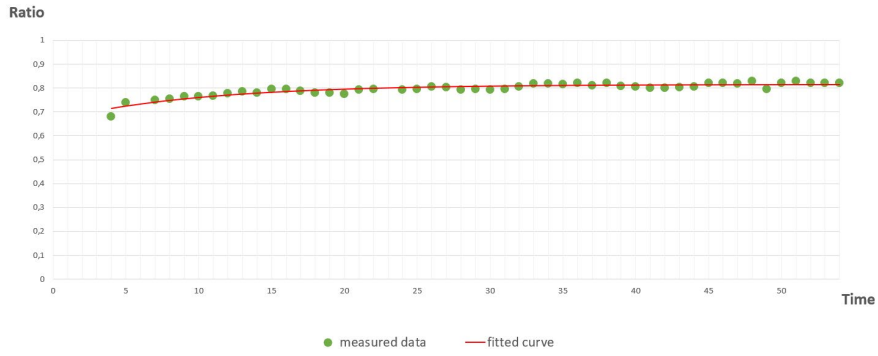

Figure 10
id = 10117; α1 (asymptote) = 0,81; α2 = 0,10; α3 (start time) = -16,71

4.2 Analysis of the Age Value

As can be seen in Chapter 4.1, we expect that as we start the investigation sooner after the appearance of malicious codes, the results will be more accurate. This mainly applies to the age and slope values since the asymptotes can also be evaluated in later stages. During this analysis, we compared the age value with information from the F-Secure antivirus company. We analyzed when F-Secure first detected the given malicious sample and how the calculation according to the fitted curve compares to this. For the cases described in Chapter 4.1.1 we expect the highest accuracy, while in the cases of Chapter 4.1.3 the least accurate.

To be able to perform the analysis on the groups mentioned in Chapter 4.1, the samples are divided into 10 decades according to the ratio of the minimum and

maximum of the measured values. The *k*th decade includes the samples, if

$$\mathbf{0,1(k-1)} \leq \frac{\min_i x_i}{\max_i x_i} < \mathbf{0,1\,k} \qquad \text{if } \mathbf{1 \leq k \leq 9} \tag{2}$$

$$\mathbf{0,1(k-1)} \leq \frac{\min_i x_i}{\max_i x_i} \leq \mathbf{0,1\,k} \qquad \text{if } \mathbf{k = 10} \tag{3}$$

where $\boldsymbol{x_i}$ is the *i*th measured value of the particular sample.

During the analysis, the difference between the time given by F-Secure (when they saw the particular sample first) and the time calculated according to the fitted curve for each malicious sample was calculated. In the case of a negative value, F-Secure encountered the sample later than the one given by the curve fitting, while in the case of a positive value, the curve fitting gave a later time. In each decade, the average, minimum, and maximum of these differences were determined for the samples belonging to that decade (Table 2, Figure 11).

Table 2
Average, minimum, and maximum differences in decades

decade (k)	average	minimum	maximum
1	0,20	-0,44	0,70
2	0,38	-0,70	1,64
3	0,71	-1,60	7,31
4	2,82	-0,71	7,01
5	2,08	-1,45	21,99
6	5,14	-0,95	40,88
7	6,30	-1,02	19,13
8	8,96	0,10	41,65
9	5,61	-1,60	56,36
10	9,16	-1,56	59,97

Based on the data, it can be concluded that in the case of $\boldsymbol{k \leq 2}$, curve fitting can determine the first appearance of the malicious code with an accuracy of $\boldsymbol{\pm 2}$ days, but if the tests start much later, this accuracy is radically reduced. In the case of $\boldsymbol{k = 5}$, the difference is already 3 weeks, which is increased in the case of $\boldsymbol{k \geq 9}$ where it can be 50-60 days.

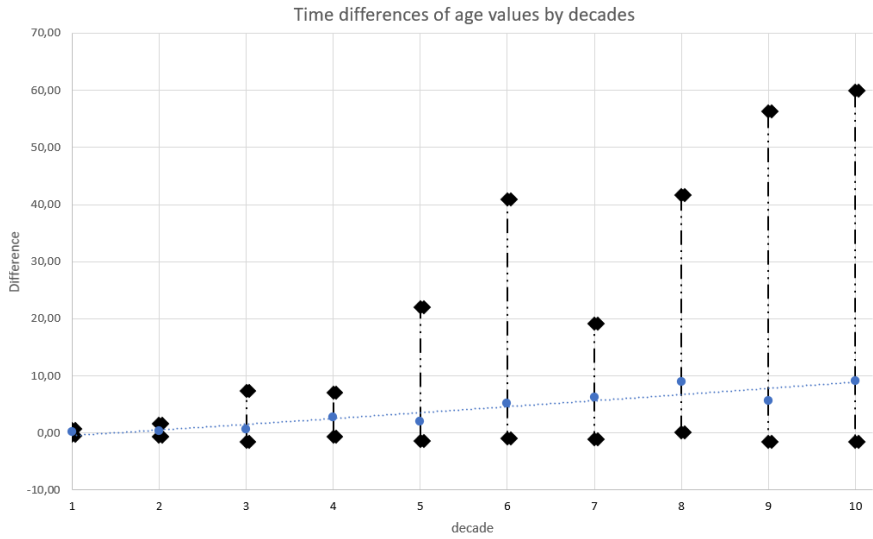

Figure 11
Time differences of age values by decades

Conclusions

In this paper, the Time Evolution Model was proposed to mathematically characterize the behavior of malicious objects and possibly their families. Using this method can answer some open questions:

- How suitable is the mathematical model for describing the real situation of malware/malware families? It must be demonstrated that a sufficient number of discrete measurement points are available to provide valuable information.
- What is the correlation between the measurement time window required to provide the minimum valuable information? Also, how can we calculate the margin of error for this model?
- How can we aggregate the parameters provided by the model for a single malware to describe the behavior of a set of malware?

If we can answer these questions and provide the pros and cons associated with them, then this methodology can be widely used to provide much better and more usable threat intelligence information about current malware and malware families.

We tested the Time Evolution Model on a large number (more than 1000) of malicious samples, based on which different groups can be distinguished, according to when the investigation begins after the first appearance of the malicious code. This article clearly shows that in the analysis of the results, it is useful to divide the

samples into decades, depending on the ratio of the measured minimum and maximum values. We will also show based on this classification, when can be the age value almost realistic (maximum error of 2 days).

References

[1] LEITOLD F.; Holló K.; Király Z.: Quantitative metrics characterizing malicious samples 2021 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), Dublin, Írország, 2021, pp. 82-83

[2] BOGNÁR L.; Joós A.; Nagy B.: Time Evolution Model for Classifying Files in Antivirus Testing Procedures, The $1st$ Conference on Information Technology and Data Science, Debrecen, Hungary, 2020, https://arato.inf.unideb.hu/biro.piroska/CITDS2020/abstract_citds.pdf

[3] BOGNÁR L.; Joós A.; Nagy B.: On the Time Series of Antivirus Testing Procedures, In: Fazekas I.; Hajdu A.; Tómács T. Proceedings of the 1st Conference on Information Technology and Data Science CITDS 2020) Debrecen, Hungary: CEUR Workshop Proceedings (2021) pp. 77-89, 13 p

[4] CHOO K.-K. R. , The cyber threat landscape: Challenges and future research directions, Computers & Security (2011) 30,8:719-731, https://doi.org/10.1016/j.cose.2011.08.004

[5] IDIKA N., Mathur A. P., A survey of malware detection technique, Tech. Rep., Purdue, Univ., February 2007, https://profsandhu.com/cs5323_s17/im_2007.pdf

[6] JANSEN W, Directions in security metrics research, The National Institute of Standards and Technology. NISTIR 7564 (2009) https://csrc.nist.gov/publications/detail/nistir/7564/final

[7] LEITOLD F., Independent AV testing 11th Annual EICAR Conference, Berlin, Germany, 2002, https://pdfs.semanticscholar.org/ce7d/eb66b0f76976a0e3b4dca299c71679cdbcc0.pdf

[8] LEITOLD F., Testing protections against web threats Malicious and Unwanted Software (MALWARE), 6th International Conference on Malicious and Unwanted Software, (2011) pp: 20-26, ISBN: 978-1-4673-0031-5

[9] LEITOLD, F., Arrott A., Hadarics K., Automating visibility into user behavior vulnerabilities to malware attack, Proceedings of the 26th Virus Bulletin International Conference (VB2016) pp. 16-24, Denver, USA, 2016

[10] LEITOLD F., Arrott A., Osorio F. C., Mike D., Pickard C., Miladinov S., Measuring the effectiveness of modern security products to detect and contain emerging threats: a consensus-based approach Proceedings of the 8th IEEE International Conference on Malicious and Unwanted Software Fajardo, Puerto Rico: IEEE (2013)

[11] LEITOLD F., Arrott A., Hadarics K., Quantifying cyber-threat vulnerability by combining threat intelligence, IT infrastructure weakness, and user susceptibility 24th Annual EICAR Conference, Nuremberg, Germany, 2016

[12] LEITOLD F., Hadarics K., Measuring security risk in the cloud-enabled enterprise, In: Dr Fernando C Colon Osorio, 7th International Conference on Malicious and Unwanted Software (MALWARE). Fajardo, Puerto Rico, 2012.10.16-2012.10.18. Piscataway (NJ): IEEE, 2012, pp. 62-66 (ISBN:978-1-4673-4880-5)

[13] LEITOLD F., Yu K., Arrott A., Component Protection Metrics for Security Product Development: CheckVir Endpoint Test Battery Malicious and Unwanted Software (MALWARE), 7th International Conference on Malicious and Unwanted Software (2012) ISBN: 978-1-4673-4880-5

[14] MOSER A., Kruegel C., Kirda E., Limits of Static Analysis for Malware Detection, Proceedings - Annual Computer Security Applications Conference, ACSAC (2008) 421-430, 10.1109/ACSAC.2007.21

[15] YIN H., Song D., Egele M., Kruegel C., Kirda E., Panorama: Capturing System-wide Information Flow for Malware Detection and Analysis, In Proceedings of ACM Computer and Communications Security (2007)

[16] Anti-Malware Testing Standards Organization, https://www.amtso.org/