mathematics

MDPI

Article

Analyzing Malware Propagation on Wireless Sensor Networks: A New Approach Using Queueing Theory and HJ-Biplot with a SIRS Model

Elisa Frutos-Bernal [1,*], Miguel Rodríguez-Rosa [1], María Anciones-Polo [1] and Ángel Martín-del Rey [2]

[1] Department of Statistics, Universidad de Salamanca, 37007 Salamanca, Spain; miguel_rosa90@usal.es (M.R.-R.); mariaanciones@usal.es (M.A.-P.)
[2] Department of Applied Mathematics, Universidad de Salamanca, 37008 Salamanca, Spain; delrey@usal.es
* Correspondence: efb@usal.es

Abstract: Most research on malware focuses mainly on its detection, without paying attention to its propagation trends. However, modeling the spread of malware is an important research problem because it allows us to predict how malware will evolve and to take steps to prevent its propagation, hence the interest in analyzing this spread from a statistical point of view. This work proposes a malware propagation prediction methodology based on multivariate statistical techniques such as HJ-Biplot in combination with closed queuing networks. Datasets generated using individual-based SIRS models are used to validate the proposed methodology, although any other model could have been chosen to test its validity. Experimental results show that the proposed model can effectively predict and classify malware and discover the influence of different model parameters on the malware propagation situation.

Keywords: malware propagation; closed queuing networks; HJ-Biplot; individual-based models; wireless sensor networks; SIRS models

MSC: 60K30; 62P30; 62H10; 68M25

Citation: Frutos-Bernal, E.; Rodríguez-Rosa, M.; Anciones-Polo, M; Martín-del Rey, A. Analyzing Malware Propagation on Wireless Sensor Networks: A New Approach Using Queueing Theory and HJ-Biplot with a SIRS Model. *Mathematics* **2024**, *12*, 135. https://doi.org/10.3390/math12010135

Academic Editor: Alexander Zeifman

Received: 31 October 2023
Revised: 21 December 2023
Accepted: 29 December 2023
Published: 31 December 2023

1. Introduction

The advent and widespread adoption of new paradigms derived from the Internet of Things (IoT), such as Industry 4.0, Smart Cities, Smart Agriculture, etc., have brought wireless connectivity to the forefront. This ubiquitous connectivity is instrumental in managing vast datasets and facilitating intelligent process automation, which promises substantial societal benefits [1]. Among the key enablers of these technological advancements, wireless sensor networks (WSNs) play a fundamental role [2].

The proliferation of connected devices (of all types) within WSNs and the massive implementation of this technology in the most diverse environments also brings forth several cybersecurity threats [3,4]. Many of these exploit inherent limitations in WSN devices, particularly concerning their computing resources and energy consumption. Consequently, effectively managing cybersecurity in WSNs necessitates a keen focus on devising and implementing strategies for controlling one of the basic tools in cyber attacks: malware.

In this context, an in-depth understanding of the propagation and infection processes of the different specimens of malicious code used in cyber attacks against WSNs is paramount for designing effective detection and control techniques. As a consequence, the development and analysis of mathematical models that simulate malware spread become indispensable in mitigating its impact.

Most of the mathematical models that have appeared in the scientific literature to simulate malware propagation on WSNs are compartmental (the population of devices is divided into different classes based on their epidemiological state: susceptible, infected,

recovered, etc.) and of a global nature. That is, the sole intention is to study the temporal evolution of the size (number of devices) of each of these classes. Consequently, the dynamics of malware are usually described by systems of ordinary differential equations [5]. While these models offer theoretical insights, they suffer from significant limitations by overlooking both complex network topology and node-particular characteristics. Fortunately, these limitations can be addressed by adopting the individual-based paradigm [6]. In this sense, individual-based models consider the unique and particular characteristics of each network node/device, including computing resources, energy consumption, location, activity level, communication attributes, and the explicit contact topology. As a result, they provide a detailed representation of the epidemiological state of each device over time, rather than only the global evolution of each compartment. Regrettably, only a handful of such individual-based models have been proposed [7–9].

Additionally, mathematical models can be classified into deterministic and stochastic. These types of models represent two fundamental approaches to understanding malware propagation in wireless sensor networks. Deterministic models are based on fixed and predictable coefficients and relations to simulate the spread of malware. They assume that the behavior of malware and its spreading on network devices follow strict rules, making them suitable for analyzing scenarios with well-defined characteristics. On the other hand, stochastic models deal with randomness and uncertainty, recognizing that real-world malware propagation often involves probabilistic characteristics. In this sense, stochastic models incorporate randomness in both epidemiological coefficients (for example, infection or recovery rates) and structural relations (node interactions), making them more suitable for capturing the unpredictability inherent in the dynamics of malware. These two modeling approaches provide valuable insights into different aspects of malware propagation, enabling researchers and cybersecurity experts to develop effective strategies for prediction, prevention, and mitigation.

These mathematical models can be studied from different perspectives, mainly the following four: qualitative study, numerical study, deep learning-based study, and statistical study. Through qualitative study (for example, the study of ergodicity in the case of stochastic models whose dynamics are described by means of Markov chains), we can determine the nature of steady states and consequently understand the dynamics of the system subject to some conditions, in which the basic reproductive number plays a fundamental role. Once initial conditions are established, the numerical study allows us to obtain simulations and also to explore the possibility of developing ad hoc numerical methods. Very recently, different architectures have been proposed in the field of deep learning to solve problems modeled by differential equations. In this sense, malware propagation models can be analyzed using both differential neural networks and physics-informed neural networks. Finally, to complete the analysis of malware propagation models, we understand that a statistical study of the results derived from simulations is also necessary (and very important). This type of study can provide a detailed view of the variability and uncertainty associated with malware behavior in dynamic environments, especially wireless sensor networks. By considering different network topologies, and taking into account various initial conditions and numerical values of epidemiological coefficients, statistical approaches allow for the identification of hidden patterns and the assessment of the robustness of the proposed models. Furthermore, they facilitate the extraction of valuable information about the effectiveness of containment strategies and the vulnerability of the network.

WSN malware propagation models, whether global or individual, deterministic or stochastic, serve a dual purpose: predicting malware behavior and aiding in the development of control and mitigation strategies. Specifically, individual-based models are valuable not only for their ability to predict both global and individual malware evolution with greater detail but also for generating datasets suitable for machine learning malware detection techniques. Consequently, in addition to qualitative and numerical analyses on individual-based or global models, different types of statistical analyses also offer an

intriguing avenue for extracting hidden correlations and meaningful insights from data obtained through simulations [10].

Most statistical malware research focuses mainly on malware detection and identification, without predicting malware propagation trends and predicting subsequent infected nodes. For example, in [11], Recurrent Neural Networks (RNNs) are used to analyze the execution operation codes (OpCodes) of ARM-based IoT applications, and in [12], RNNs are also used to predict whether an executable is malicious or benign within the first five seconds of execution with 94% accuracy. AMalNet, a robust deep learning framework for Android malware detection, is proposed in [13]. This framework uses graph convolutional networks (GCNs) to capture graphical semantics and independent recurrent neural networks (IndRNNs) to decode deep semantic information. In [14], a graph attention network-based (GAN) framework is proposed to detect malware attacks targeting intelligent transportation systems (ITS). In [15], GCNs are also used to detect Android malware. Few studies have attempted to model the dynamics of malware propagation and predict the transient behavior of these dynamics. In this sense, [10] uses a Bayesian structural time series (BSTS) model to study the data-driven dynamics of malware propagation in cyberspace. In [16], a malware propagation prediction model based on representation learning and GCNs is proposed, such that the proposed model predicts whether potential nodes will be infected by malware. As far as we know, the use of statistical techniques as a complement to the qualitative and numerical studies mentioned above has not been thoroughly explored. This is the motivation that has driven us to initiate this line of research.

The main goal of this work is to explore the use of two statistical techniques—queuing theory and multivariate analysis or, more specifically, the HJ-Biplot—to characterize the behavior of malware specimens as they propagate on wireless sensor networks (that can be mathematically described by complex networks). As mentioned earlier, and to the best of our knowledge, this approach has not been employed before, making the research relevant from an academic perspective. This statistical analysis is performed using a synthetic dataset generated from several simulations obtained from an individual-based stochastic SIRS compartmental model. In this regard, the results and conclusions obtained are limited to the scope of such examples, leaving the extension to other compartmental models for future developments.

Queuing theory has been applied to various fields. For example, it has been used in telecommunications to analyze Ethernet traffic [17], in logistics and supply chains to analyze online shopping stores [18], and to analyze healthcare systems [19]. It has also been used in malware analysis to propose probabilistic models that aim to describe the overall behavior of the system when attacked by malicious nodes. In [20], a probabilistic model based on the susceptible–infected–susceptible (SIS) paradigm and on the theory of closed queuing networks is proposed to analyze the spread of malicious software in ad hoc networks. Queuing theory enhanced with stochastic analysis is employed to overview the dynamics of epidemic cases in [21] and to assess cyber disruptions and attacks to the national airspace system [22].

As far as biplots are concerned, they have also been used to analyze data from different fields of knowledge, such as hydrology [23], transport networks [24], sustainability [25,26], healthcare [27], and the chemosphere [28], but never to analyze malware propagation in wireless sensor networks. The main advantage of biplot methods is that they do not take into account parametric assumptions and have the advantage of being a specific statistical tool for the representation of multivariate data.

Our analytical approach, combining queuing theory and HJ-Biplot with an SIRS model, represents a novel perspective on understanding malware propagation in wireless sensor networks. In connecting our results with previous studies, we emphasize the departure from traditional methods and highlight the potential benefits of our unique combination of analytical tools. By referencing established literature on malware dynamics, we underscore the innovative aspects of our work, positioning it within the evolving landscape of network

security research. This integration serves not only to validate our methodology but also to contribute fresh insights to the existing body of knowledge.

Beyond the immediate scope of malware propagation, our results hold broader implications for epidemic modeling and network security. We explore how the principles derived from our study can be applied to understand the dynamics of other infectious phenomena in diverse networked environments. By considering the broader context, we underscore the versatility of our approach and its potential relevance to fields beyond malware research.

Our current work lays the theoretical groundwork for analyzing malware propagation using queuing theory and HJ-Biplot with an SIRS model. Future research should focus on empirical validation using real-world datasets to assess the model's effectiveness in practical scenarios. Another avenue for further exploration involves a detailed sensitivity analysis of the model parameters, since investigating the impact of variations in key parameters on the model outcomes will provide a more nuanced understanding of the dynamics of malware propagation. Additionally, the generalizability of our approach to diverse topologies is an exciting prospect for future research, as exploring how the model performs in various network structures will enhance its applicability across different wireless sensor network configurations.

The rest of the paper is organized as follows: Section 2 briefly describes the SIRS models (deterministic and stochastic) and the simulations performed to generate the datasets used in the analysis. Section 3 presents the methodology used: closed queuing networks and HJ-Biplot. Section 4 presents the results of applying the above methodology to the datasets obtained from the simulations. Finally, Section 5 discusses the main conclusions.

2. SIRS Model and Data Description

2.1. SIRS Models

Two mathematical models, one deterministic and the other stochastic, for malware propagation on WSNs are used to generate the datasets employed in the statistical analysis. They are compartmental and individual models, in which the devices are classified into three different compartments: susceptible S (non-infected devices), infectious I (devices reached by malware), and recovered R (devices during a temporal immunity period). As a recovered device is endowed with temporal immunity, the dynamics of the models correspond to the SIRS type.

The SIRS models employed in this study are particularly well-suited to address the current problem due to the fact that the SIRS model closely mimics the dynamics of infectious diseases. In this context, it effectively represents the propagation and recovery dynamics of malware. The deterministic model allows for a precise mathematical description, while the stochastic model introduces randomness, acknowledging the inherent uncertainties in real-world scenarios. Therefore, the selection of the SIRS models for this study was motivated by a combination of their relevance to real-world malware dynamics and the practical constraints of working with a "simplified" dataset.

The communication topology of the wireless sensor networks determines the contact topology of nodes/devices; that is, the set of devices for which there exists a communication link with the i-th device is called its neighborhood and is denoted by $\mathcal{N}_i$. The state of the i-th device at timestep t is denoted by $\texttt{state}[i,t] \in \mathcal{S} = \{S, I, R\}$; that is:

$$\texttt{state}[i,t] = \begin{cases} S, & \text{if the } i\text{-th node is susceptible at } t \\ I, & \text{if } i\text{-th is infectious at } t \\ R, & \text{if } i\text{-th is recovered at } t \end{cases}. \tag{1}$$

The deterministic local transition function that describes the dynamics of the deterministic individual-based model is the following:

- If $\texttt{state}[i,t] = S$, then

$$\texttt{state}[i,t+1] = \begin{cases} I, & \text{if } \frac{|\mathcal{I}_i(t)|}{k_i} > u_i(t) \\ S, & \text{otherwise} \end{cases} \tag{2}$$

- If $\texttt{state}[i,t] = I$, then

$$\texttt{state}[i,t+1] = \begin{cases} R, & \text{if } T_{inf}(i,t) = p_{inf}(i) \\ I, & \text{if } T_{inf}(i,t) < p_{inf}(i) \end{cases} \tag{3}$$

- If $\texttt{state}[i,t] = R$, then

$$\texttt{state}[i,t+1] = \begin{cases} S, & \text{if } T_{inf}(i,t) = p_{inm}(i) \\ R, & \text{if } T_{inf}(i,t) < p_{inm}(i) \end{cases}, \tag{4}$$

where $\mathcal{I}_i(t) \subseteq \mathcal{N}_i$ is the set of infectious neighbors of the i-th node at timestep t, k_i is the degree of the i-th node, $0 < u_i(t) \leq 1$ is a threshold parameter that depends on the particular characteristics of the i-th node at the step of time t, $p_{inf}(i)$ (*resp.* $p_{inm}(i)$) is the length of the infectious (*resp.* immunity) period associated with the i-th node, and $T_{inf}(i,t)$ (*resp.* $T_{inm}(i,t)$) is the number of timesteps that the i-th node has been in the infectious (*resp.* immunity) period at timestep t.

On the other hand, the stochastic local transition function employed is defined as follows:

- If $\texttt{state}[i,t] = S$, then

$$\texttt{state}[i,t+1] = \begin{cases} I, & \text{with probability } a_i(t) \\ S, & \text{otherwise} \end{cases} \tag{5}$$

- If $\texttt{state}[i,t] = I$, then

$$\texttt{state}[i,t+1] = \begin{cases} R, & \text{with probability } b_i(t) \\ I, & \text{otherwise} \end{cases} \tag{6}$$

- If $\texttt{state}[i,t] = R$, then

$$\texttt{state}[i,t+1] = \begin{cases} S, & \text{with probability } c_i(t) \\ R, & \text{otherwise} \end{cases}. \tag{7}$$

2.2. Data Description

The SIRS model is used to perform the simulations. These simulations provide the datasets used in the proposed statistical analysis. In all of them, the number of devices is $n = 100$, and the simulations are computed during a period of 100 units of time: $0 \leq t \leq 100$. Furthermore, it is assumed that there is only one infectious device ("patient zero") at $t = 0$, which in all cases will be the device with the highest betweenness centrality. Different types of complex networks—taking into account their degree distributions—are considered to carry out the simulations, namely, full networks, random networks, scale-free networks, and small-world networks. For random networks, ER connecting probabilities of 0.1 and 0.2 are considered. The Barabási–Albert (BA) algorithm with values of 3 and 5 for the m parameter is used to generate the scale-free network and, in small-world networks, probabilities of 0.1 and 0.2 for the occurrence of a new edge are considered. The choice of patient zero was made on the basis that this assumption would favor a higher initial rate of spread of the epidemic.

Furthermore, deterministic and stochastic local transition functions are employed. When stochastic transition functions are used, the values of the parameters considered are $u_i(t) = 0.1$ and 0.01, $p_{inf}(i)$ as a random integer between 2 and 7 and between 5 and 9, and $p_{inm}(i)$ as a random integer between 3 and 5 and between 4 and 6. If the transition

function used is of the stochastic type, the parameter values considered are the probability $a_i(t) = 0.1$ and 0.5, probability $b_i(t) = 0.025$ and 0.005, and probability $c_i(t) = 0.2$ and 0.4. It is important to note that the selection of parameter values listed above is for illustrative purposes only.

By modifying all the above parameters (type of complex network, transaction probabilities, length of periods), 56 stochastic and 56 deterministic simulations are carried out. The result of each simulation is a data matrix that stores the state of node i at time unit t (state$[i, t]$). Therefore, the methodology based on closed queuing networks is applied to a sequence of data matrices, $\mathbf{X}_k$, with $k = 1, \ldots, K$, with $K = 56$, generated by a stochastic SIRS model, and another sequence generated by a deterministic SIRS model, each of them obtained by varying the parameters in the simulations. From now on, we will label these matrices as $Model_1, \ldots, Model_K$, and we will represent our dataset made of K matrices as $\underline{\mathbf{X}} = (\mathbf{X}_1, \ldots \mathbf{X}_K)$

3. Methodology

3.1. Closed Queuing Networks

Let us consider a complex network that can be represented by an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a finite set of nodes, with $|\mathcal{V}| = N$, and $\mathcal{E}$ is a finite set of edges, with information about which pairs of nodes are related. Let $\mathcal{S}$ be the finite set of states that can be assumed by each node at each timestep, where state$[i, t] \in \mathcal{S}$ represents the state of the i-th node at step t, with $|\mathcal{S}| = M$.

The same network can be understood as a closed queuing network [29] by means of another graph, a directed one, $\mathcal{G}' = (\mathcal{V}' = \mathcal{S}, \mathcal{E}')$; that is, its vertices are now the set of states, so for each time step, the graph represents the amount of nodes from $\mathcal{V}$ that belong to each state: given a state $s \in \mathcal{S}$, $|s_t| = |\{v_i \in \mathcal{V} \mid \text{state}[i, t] = s\}|$. Additionally, the new edges $\mathcal{E}'$ represent which states are related and how states can vary over time. Therefore, the network $\mathcal{G}$ can be conceptualized as a closed queuing network $\mathcal{G}'$, where that concept can be thought of as follows: $\mathcal{G}'$ keeps the information about the epidemiological process that is happening in $\mathcal{G}$, as well as how many nodes belong to each state or how they evolve, but $\mathcal{G}'$ cannot be used to regain $\mathcal{G}$.

3.1.1. Traffic Equations

The arrival rates to each state (λ_s) need to be obtained, and that can be done by using the following traffic equations:

$$\lambda_s = \sum_{r \in \mathcal{S}} \lambda_r \cdot p_{rs} \quad \forall s \in \mathcal{S},$$

where $p_{rs} \geq 0$ is the probability of a node changing its state from r to s.

However, the preceding M equations form an indeterminate system of linear equations with one degree of freedom. Therefore, since the arrival rates cannot be directly obtained, it is necessary to set an arbitrary positive value for one of the unknowns, such as $\lambda^*_{s_1} = 1$, and then the solution to this system will be the *relative* arrival rates for each state (λ^*_s).

$$\lambda^*_s = \sum_{r \in \mathcal{S}} \lambda^*_r \cdot p_{rs} \quad \forall s \in \mathcal{S}$$

3.1.2. Performance Measures

The following performance measures can be calculated for a network with N nodes:

- $L_s(N)$: This is the average number of nodes in state s.
- $W_s(N)$: This is the actual arrival rate at the s-th state.
- $\lambda_s(N)$: This is the *real* arrival rate at the s-th state.

In order to compute the three performance measures, the following iterative Algorithm 1 is used, which computes these measures for increasing values of n starting from $n = 1$:

Algorithm 1 Algorithm to compute the performance measurements: L_s, W_s, and λ_s

for $s \in \mathcal{S}$ **do**
$L_s(0) \leftarrow 0$
end for

for $n = 1, \ldots, N$ **do**

for $s \in \mathcal{S}$ **do**

$$W_s(n) \leftarrow \frac{1}{\mu_s} + \frac{L_s(n-1)}{c_s \cdot \mu_s} \tag{8}$$

end for

for $s \in \mathcal{S}$ **do**

$$L_s(n) \leftarrow n \cdot \frac{\lambda_s^* \cdot W_s(n)}{\sum_{r \in \mathcal{S}} \lambda_r^* \cdot W_r(n)} \tag{9}$$

end for

for $s \in \mathcal{S}$ **do**

$$\lambda_s(n) \leftarrow \frac{L_s(n)}{W_s(n)} \tag{10}$$

end for

end for

where μ_s are the departure ratios, that is, the average number of nodes that abandon the s-th state per unit of time, and c_s is the number of nodes that can abandon the s-th state at the same time.

Actually, a fourth performance measure can be useful: the occupation ratio $U_s(N)$ defined as the probability of a state having all nodes in it,

$$U_s(N) = P(\{\text{state}[i,t] = s, \quad \forall v_i \in \mathcal{V}\}), \tag{11}$$

which can be computed within the same algorithm by means of the following step:

Addition of U_s to the Algorithm

for $s \in \mathcal{S}$ **do**

$$U_s(n) \leftarrow \frac{\lambda_s(n)}{c_s \cdot \mu_s} \tag{12}$$

end for

The analysis of the sequence of data matrices obtained from the simulations is therefore carried out in three distinct steps:

- In the first step, for each model, the departure ratios μ_s and the number of nodes that can leave the s-th state at the same time, c_s, are computed from the data collected in the sequence matrices $\underline{\mathbf{X}}$. This process allows for the computation of the performance measures given by Equations (8) and (12).
- Next, the closed queueing network approach, specifically the above algorithm (one for each simulation), is used to fit the data (see Figure 1). In this way, the performance

measures of each of the models are obtained in terms of $L_s(N), W_s(N), \lambda_s(N), U_s(N)$, with $s \in \mathcal{S}$:

$$L_{s_1}(N), \ldots, L_{s_M}(N),$$
$$W_{s_1}(N), \ldots, W_{s_M}(N),$$
$$\lambda_{s_1}(N), \ldots, \lambda_{s_M}(N),$$
$$U_{s_1}(N), \ldots, U_{s_M}(N)$$

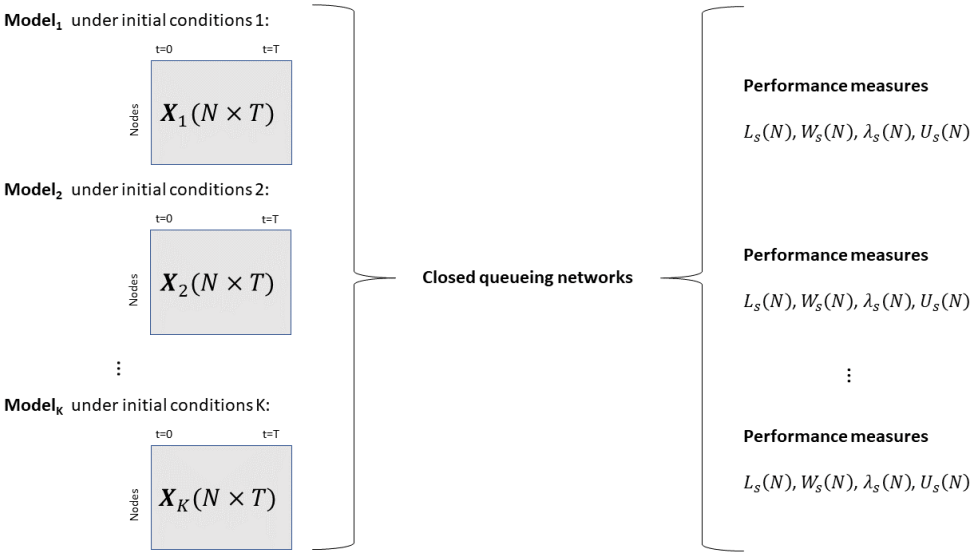

Figure 1. Second step: Computation of the performance measures from the closed queuing networks approach.

- Finally, a matrix $\mathbf{Z}(K \times 4M)$ is built from the performance measures previously obtained for each of the models. This matrix is analyzed using multivariate techniques, more specifically using the HJ-Biplot (see Figure 2). The main advantage of using HJ-Biplot is that we can simultaneously interpret the position of the variables (performance measures) as represented by vectors and the observations (models) as represented by points, as well as the relationships between them.

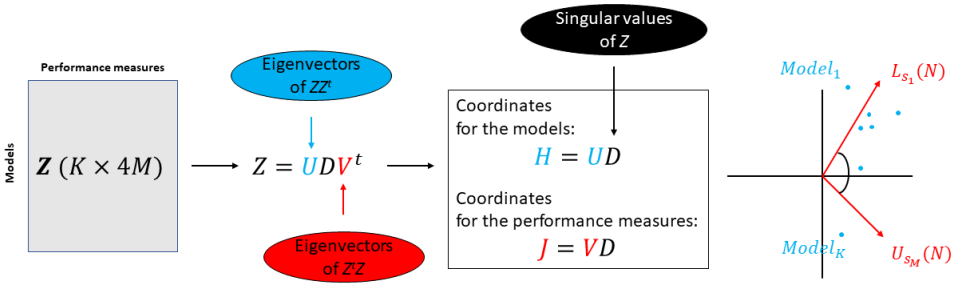

Figure 2. Third step: HJ-Biplot from the results of the closed queuing networks approach.

3.2. HJ-Biplot

In HJ-Biplot [30], two principal component analysis (PCA) techniques are performed simultaneously on a data matrix consisting of several rows or observations—the models in our case—and columns or variables, which are the performance measures in our case. This data matrix is then transformed into a standardized matrix where, for each column, the mean is 0 and the standard deviation is 1.

Given the data set $\mathbf{Z}$ with K models and $4M$ performance measures (the four measures L, W, λ, U for each state), HJ-Biplot involves computing the singular value decomposition (SVD) of the standardized data matrix $\mathbf{Z}$:

$$\mathbf{Z} = UDV^t,$$

where U is a $K \times K$ matrix containing the eigenvectors of $\mathbf{Z}\mathbf{Z}^t$ or, equivalently, the left singular vectors of $\mathbf{Z}$; V is a $4M \times 4M$ matrix containing the eigenvectors of $\mathbf{Z}^t\mathbf{Z}$ or, equivalently, the right singular vectors of $\mathbf{Z}$; and D is a $K \times 4M$ diagonal matrix containing the singular values of $\mathbf{Z}$.

HJ-Biplot displays the scores of the observations on the first few principal components—usually two—and the loadings of the variables. The rows are represented as points, while the columns are represented as vectors, where the lengths of the vectors represent the variability of the variables.

The coordinates of the observations on the two-dimensional HJ-Biplot graph are given by the product UD, where only the first two columns in U are considered and only the first two rows and two columns are considered in D. Analogously, the coordinates of the variables on the same graph are given by VD, where only the first two columns in V are considered.

HJ-Biplot is a powerful tool for visualizing relationships among variables. Actually, the biplot graph displays the relationships among the observations and the variables, can be used to identify patterns and groupings among them, and can be used to explore the underlying structure of the data as follows [31]:

- One important aspect of interpretation is understanding the distance between the points. The closer two points are to each other, the more similar the observations are in terms of the variables represented on the graph. This information can be used to identify clusters or groups of observations that share similar characteristics.
- Another important aspect of interpretation is understanding the angle between the vectors. The angle between two vectors on the graph represents the correlation or association between the variables they represent. The closer the angle is to 0°, the stronger the positive correlation between the variables; the closer the angle is to 180°, the stronger the negative correlation between the variables.
- Additionally, the length of the vectors represents the magnitude or importance of the variables in explaining the variability in the dataset. Longer vectors indicate variables that have a greater influence on the data, while shorter vectors indicate variables that have a smaller influence.
- The relationships between the points and the vectors in the HJ-Biplot can be interpreted as well. The proximity of a point to a vector can indicate the degree of association between the corresponding observation and variable. For example, if a point representing a certain observation is located close to a vector representing a particular variable, it may suggest a strong relationship between that observation and the corresponding variable. On the other hand, if a point is located far from all the vectors, it may suggest that the corresponding observation does not have a clear relationship with any of the variables represented on the graph.

Once HJ-Biplot has been used to analyze the structure of a data matrix, it can also serve as a valuable tool for making predictions about new observations. HJ-Biplot provides a visual representation of the relationships between observations and variables in the dataset, making it easier to understand the underlying structure and identify patterns and trends. By using HJ-Biplot for prediction, one can extend one's understanding of the data to include new observations and examine how these new data points relate to the existing patterns within the dataset.

Let us suppose an HJ-Biplot has been constructed using a set of models and their respective performance measures. It started with a sequence of matrices called $\underline{\mathbf{X}}$, derived from a sequence of models used to create the HJ-Biplot. Now there is another data ma-

trix, $\mathbf{X}new(N \times T)$, but there is no knowledge of the underlying model, initial conditions, network type, epidemiological coefficients, or whether deterministic or stochastic local transition functions were used. The aim is to predict the underlying model based on the evolution of the nodes in $\mathbf{X}new$ by calculating the coordinates of the new observation within the same HJ-Biplot graph. This can be achieved using the following methodology:

- First, μ_s and c_s are computed for each state $s \in \mathcal{S}$ as done for the known models.
- Next, the previous algorithm is utilized to calculate the $4M$ performance measures for this matrix $\mathbf{X}new$ (see Figure 3), resulting in a new observation—a new row—denoted as $u_{new} = (u_1, u_2, \ldots, u_{4M})$.
- Finally, the coordinates of this new observation are computed on the same HJ-Biplot graph. Since we have the matrices U, D, and V from the SVD of $\mathbf{Z}$ containing information from all our original models, we can project the new observation onto the same graph as $u_{new} \cdot V$. In this projection, we only take into account the first two columns in V (see Figure 4).

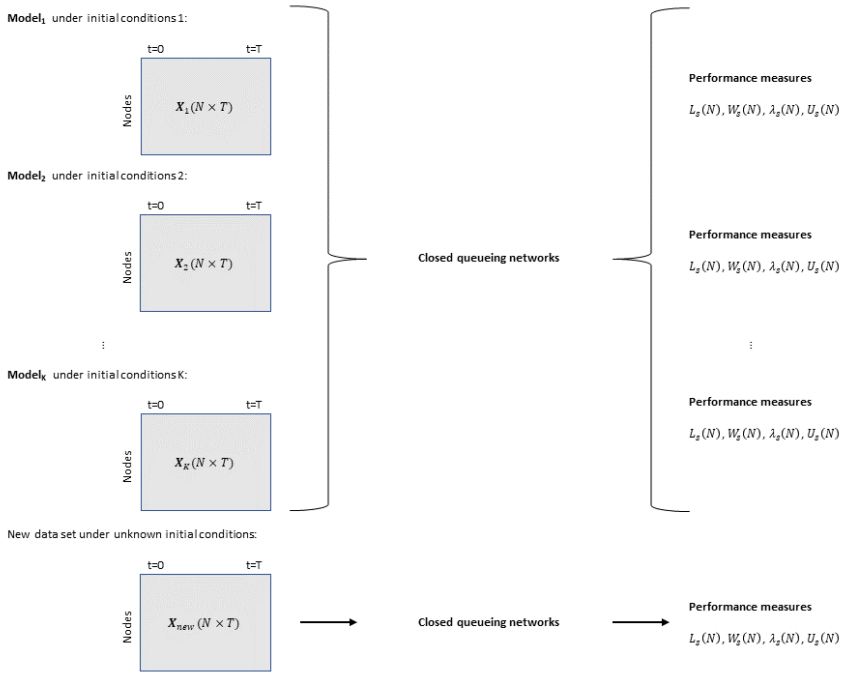

Figure 3. Computation of the performance measures for the new dataset.

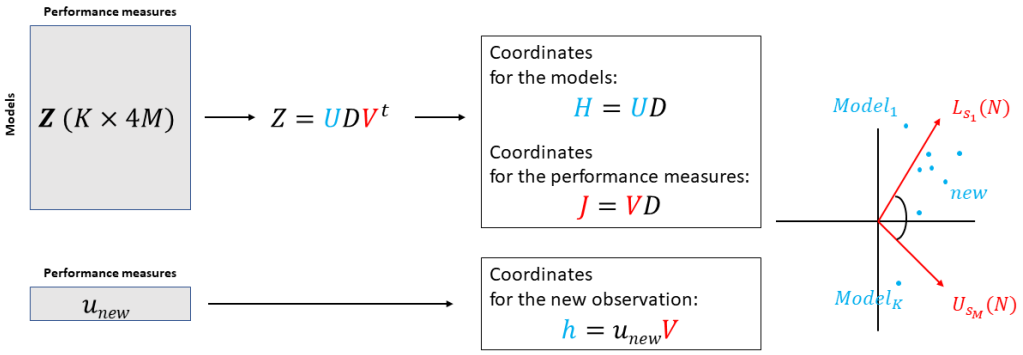

Figure 4. Projection of the new observation on the HJ-Biplot.

Once the coordinates of the new observation have been calculated and plotted on the existing HJ-Biplot, its position relative to the original models and variables becomes

apparent. This positioning helps to identify and predict the potential model from which the new dataset may have originated.

4. Results

Let us now present a precise description of the results, specifically, the generated plots obtained through the application of HJ-Biplot on the two datasets. These databases contain information regarding the propagation and infection dynamics of various simulated malware types. Notably, this malware follows either a deterministic or a stochastic SIRS model for its propagation. Furthermore, in this section, we interpret these theoretical findings from an epidemiological perspective and draw meaningful conclusions.

Figures 5 and 6 show three groups of deterministic SIRS models according to their propagation dynamics. The first group, located on the left side, comprises malware propagation models characterized by a significant number of nodes being in the susceptible state (high L_0), a prolonged period of susceptibility (high W_0), and a higher probability of all nodes being susceptible (high U_0). These models may aptly be labeled as "passive" since they do not induce nodes to transition to infectious or recovered states; rather, they predominantly keep the nodes in a susceptible state. In the second group, situated in the bottom right corner, we find a different category of models. These are characterized by a substantial number of nodes being infectious (high L_1) and a higher probability that all of them are in the infectious state (high U_1). The third group, positioned in the top right corner, includes models featuring a significant proportion of nodes in the recovered state (high L_2). These malware models also exhibit a higher probability that all nodes have recovered (high U_2). Furthermore, they are distinguished by rapid transitions from recovered to susceptible and from infectious to recovered states, as indicated by high values of λ_0 and λ_2, respectively. In summary, models within the second and third groups infect nodes swiftly, with the distinguishing factor being that those in the second group maintain nodes in an infectious state, while those in the third group enable nodes to recover quickly.

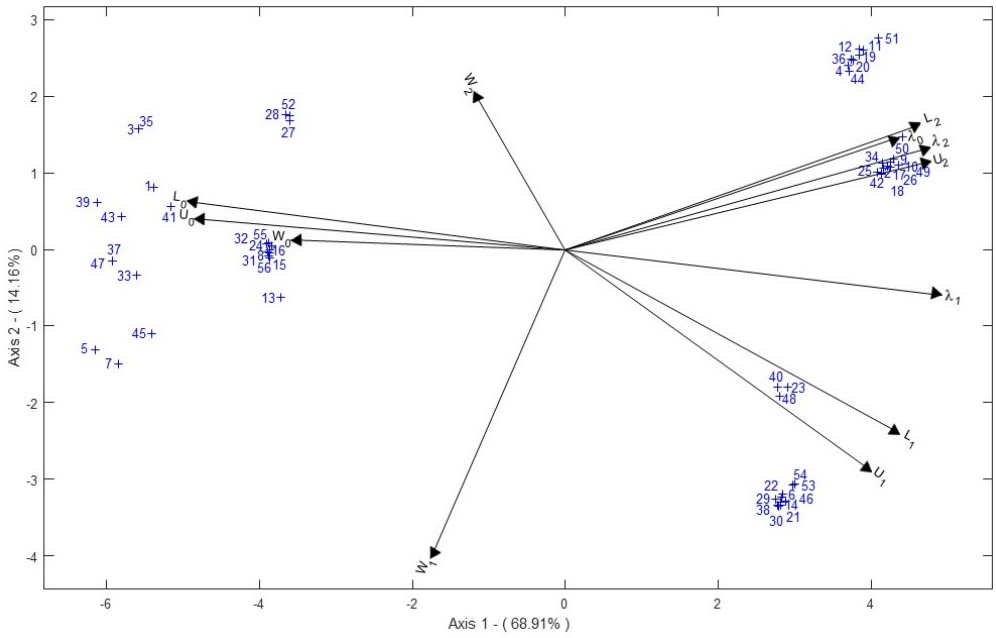

Figure 5. Resulting HJ-Biplot for simulated malware models that propagate according to a deterministic SIRS model.

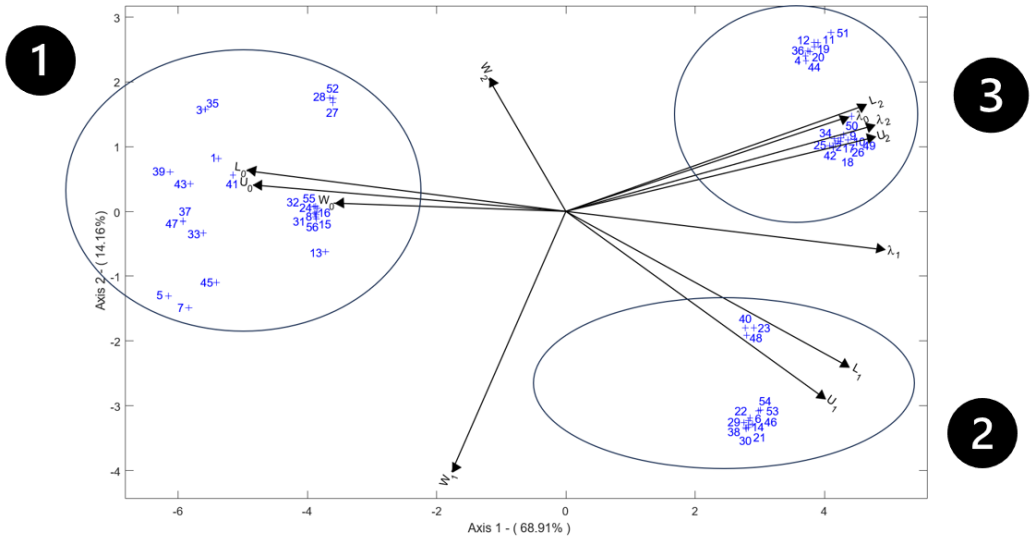

Figure 6. HJ-Biplot with different groups of deterministic SIRS model malware.

Performance measures W_1 and W_2 do not exhibit significant correlations with any malware propagation model. This lack of correlation is evident, as the angles they form with respect to the other measures or groups are approximately 90°. In other words, the behavior of nodes is not notably influenced by the duration of time they spend in the infectious state (W_1) or the recovered state (W_2).

Regarding the relationship between these groups, we can observe that malware models in the first group, the "passive" category, exhibit strong negative correlations with the other two groups, namely, those that keep nodes in an "infectious" state and those that facilitate a "quick recovery" of nodes (forming angles close to 180° with these groups). This observation can be interpreted as follows: given that "passive" models are characterized by high values of L_0, W_0, and U_0, and models that keep nodes "infectious" or enable them to "recover quickly" are characterized by a high value of λ_1, it follows that a "passive" model will have a low value of λ_1 (resulting in a slower transition of nodes from susceptible to infectious). Conversely, if a model is designed to keep nodes "infectious" or promote "quick recovery", it will have low values of L_0, W_0, and U_0 (with almost no nodes in the susceptible state; if any are, they spend minimal time in the susceptible state, making it highly unlikely that all nodes are susceptible).

Furthermore, groups 2 and 3 exhibit a strong positive correlation with each other and with λ_1, highlighting their common characteristic of facilitating rapid transitions of nodes from the susceptible to the infectious state.

Again, we can see how the different malware propagation models are also sorted into three large groups in the stochastic SIRS model plots (Figures 7 and 8). These graphs are interpreted by talking about the groups the models form, the angles from which performance is measured, and the relationships between the two characteristics [31].

A detailed description of the three groups to which the models belong is given. The first group, situated in the bottom left corner, consists of models characterized by their ability to render a significant number of nodes in a susceptible state (L_0). These models also have a higher probability of keeping all nodes in a susceptible state (high U_0) and facilitate swift transitions from the recovered state to the susceptible state (λ_0). This group bears similarities to the first "passive" group observed in the deterministic SIRS model plot, with the key distinction that there is a high rate of transition to the susceptible state. Thus, we can still classify these models as "passive", but with the added feature of a high rate of transition to susceptibility. In the second group, positioned in the top left corner, we find a different category of models. These models are characterized by their influence on a significant number of nodes transitioning to the recovered state (L_2), a higher probability of all nodes being in the recovered state (high U_2), and swift transitions from the susceptible

to infectious state and from the infectious to recovered state (high λ_1 and λ_2). This group shares similarities with the third "quickly recovered" group from the deterministic SIRS model plot, with the exception that, in this case, the transition speed from recovered to susceptible is not high (λ_0). Instead, the rate of recovery is high after being infectious, meaning that nodes quickly transition from the infectious to the recovered state. Lastly, on the right side, the third group is located. These malware models are characterized by the fact that nodes are predominantly in an infectious state (high L_1), and there is a high probability that all nodes are in the infectious state (high U_1). This is analogous to what was observed in the second "stay infectious" group from the deterministic SIRS model plot. However, in this scenario, the nodes spend a significant amount of time in the infectious state (high W_1), and the speed of transition from susceptible to infectious (λ_1) does not significantly affect their behavior.

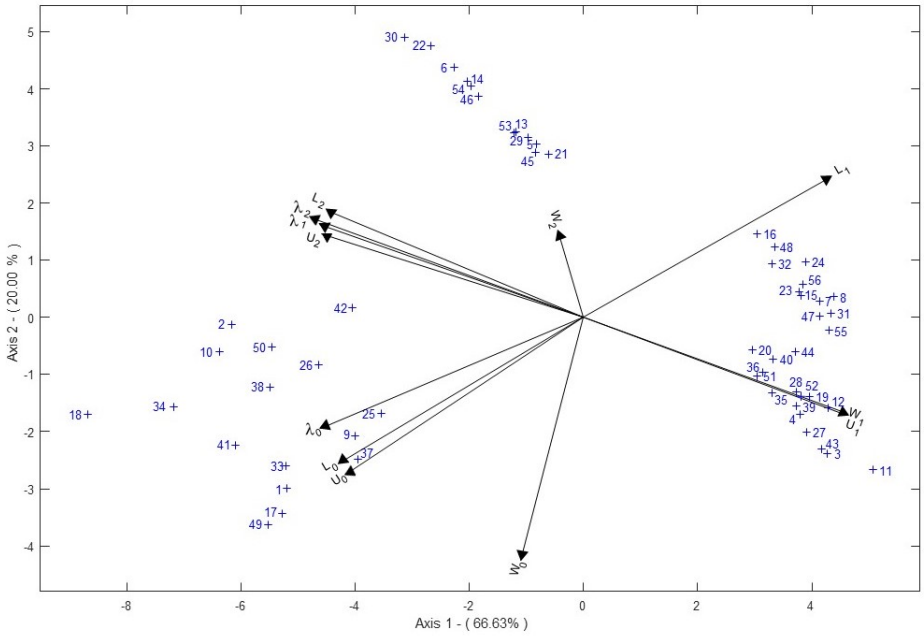

Figure 7. Resulting HJ-Biplot for simulated malware models that propagate according to a stochastic SIRS model.

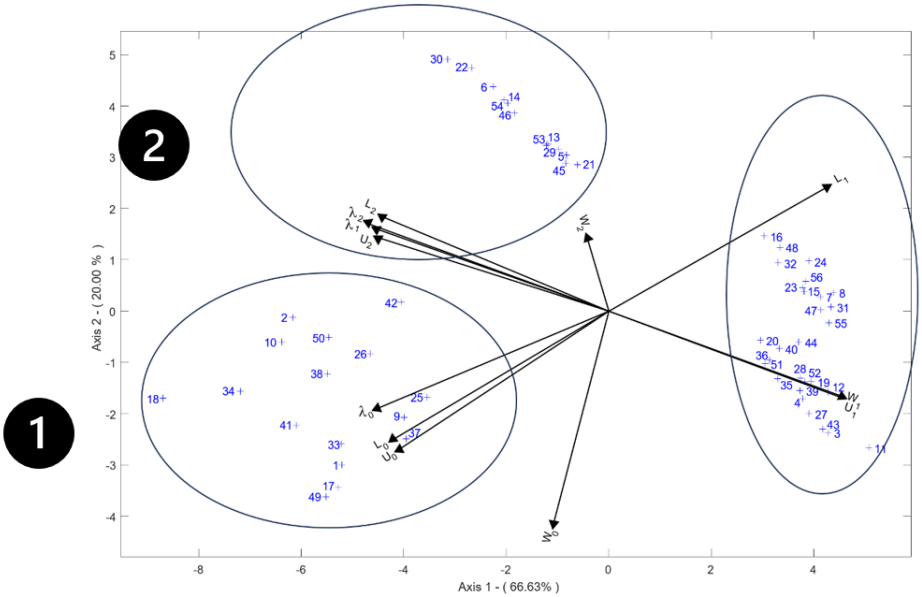

Figure 8. HJ-Biplot with different groups of stochastic SIRS model malware.

In the context of the stochastic SIRS model, performance measures W_0 and W_2 are not significantly correlated with any model. This observation suggests that the behavior of nodes is not substantially influenced by the duration of time they spend in the susceptible state (W_0) or the recovered state (W_2).

In contrast to the deterministic SIRS model plot, in the stochastic SIRS model plot shows a different pattern of relationships between the groups of models. Specifically, models in the first group, the "passive" category, are highly negatively correlated with only the third group, which comprises models that keep the nodes "infectious". Groups 2 and 3 also exhibit strong negative correlations with each other. Interestingly, "passive" models and models that promote "quick recovery" are positively correlated.

Once the HJ-Biplots have been interpreted to characterize the existing models, they can indeed be valuable for making predictions about new, unknown models. Suppose there is another data matrix containing information about an epidemiological process attributed to an unknown model, and the goal is to identify which model is responsible for it. In this scenario, one can plot this new matrix on both graphs associated with the deterministic SIRS model and the stochastic ones. By doing so, it becomes possible to assess which of the original models bear the closest resemblance to the new, unknown model. This approach enables making two predictions regarding the model that initiated the observed process: one based on the most similar deterministic SIRS model malware and another based on the most similar stochastic model malware.

5. Conclusions

The results of our study hold significant implications for the field of computer security and computational epidemiology. The understanding of malware propagation dynamics in different environments is crucial for developing robust cybersecurity measures. These findings could aid in the creation of more effective strategies to counteract malware propagation, enhancing the security of computer networks and systems.

The generated HJ-Biplots serve as powerful tools for the classification and prediction of new, unknown malware models. By plotting data from an unidentified model onto two distinct epidemiological contexts—deterministic and stochastic SIRS models—one can assess which of the original models closely resembles the new, unknown model. This dual-approach prediction methodology contributes to the potential identification of the responsible malware model in a given epidemiological scenario, thereby enhancing the preparedness of cybersecurity professionals.

In this comprehensive study, we have undertaken an in-depth analysis of the propagation and infection dynamics of simulated malware across both deterministic and stochastic SIRS models. Through the utilization of HJ-Biplots, we have generated insights into the behavior and characteristics of these malware types, yielding valuable knowledge with relevance to cybersecurity and computational epidemiology.

Our exploration began with a close examination of malware propagation in the context of a deterministic SIRS model. This analysis unveiled a division of the malware models into three discernible groups, each possessing unique attributes that profoundly impact propagation dynamics. The first group, often labeled as "passive" models, significantly maintains nodes in a susceptible state while providing extended susceptibility periods. The second group is characterized by the rapid transition of nodes to the infectious state, where they remain. The third group exhibits swift transitions from a recovered state to a susceptible state. The relationship between these groups unveils a complex interplay that sheds light on the diverse mechanisms behind malware propagation.

Moving on to the stochastic SIRS model, we conducted a similar analysis, once again discovering three predominant groups within the malware models. These groups are distinguished by the number of nodes in specific states and the speed of state transitions. The first group is similar to the "passive" group in deterministic models but involves a high rate of transition to susceptibility. The second group enables quick transitions from the infectious state to recovery, while the third group primarily maintains nodes in an

infectious state, with little sensitivity to state transition speeds. The interplay among these groups provides further insights into the stochastic malware propagation landscape.

One of the most intriguing findings pertains to the relationships between these groups. In deterministic models, "passive" models exhibit a negative correlation with groups that keep nodes "infectious" or facilitate "quick recovery". This suggests that the speed of transitioning nodes from the susceptible state is inversely related to the ability to maintain nodes in an infectious or quickly recovered state. Interestingly, positive correlations exist between the groups that keep nodes "infectious" and those that facilitate "quick recovery", signifying a shared characteristic in terms of facilitating rapid transitions.

It is important to recognize certain inherent limitations of this study that must be taken into account when interpreting its results. Firstly, the validity and generalizability of our conclusions may be affected by the size of the dataset used to build the malware evolution prediction model. In addition, it should be noted that this study relied exclusively on two individual-based SIRS propagation models. Future research could address these limitations by incorporating larger datasets obtained by manipulating a greater number of parameters and exploring different propagation models to improve the robustness and comprehensiveness of the results obtained. As mentioned in the introduction, this is an initial exploration of the uses of statistical techniques to better understand the underlying mechanisms of malware propagation dynamics. Two techniques have been employed (queuing networks and HJ-Biplot), but it is clear that we should also consider, as another future research line, extending the study using additional statistical methods.

Author Contributions: Conceptualization, E.F.-B., M.R.-R. and Á.M.-d.R.; methodology, E.F.-B. and M.R.-R.; formal analysis, M.R.-R. and M.A.-P.; draft manuscript preparation: E.F.-B., M.R.-R., M.A.-P. and Á.M.-d.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by Fundación Memoria D. Samuel Solórzano Barruso (Universidad de Salamanca, Spain) under research grant FS/2-2022.

Data Availability Statement: We have described the data used and the acquisition methods in detail in Section 2.2

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Furstenau, L.B.; Rodrigues, Y.P.R.; Sott, M.K.; Leivas, P.; Dohan, M.S.; Lopez-Robles, J.R.; Cobo, M.J.; Bragazzi, N.L.; Choo, K.K.R. Internet of things: Conceptual network structure, main challenges and future directions. *Digit. Commun. Netw.* **2023**, *9*, 677–687. [CrossRef]
2. Kandris, D.; Nakas, C.; Vomvas, D.; Koulouras, G. Applications of Wireless Sensor Networks: An Up-to-Date Survey. *Appl. Syst. Innov.* **2020**, *3*, 14. [CrossRef]
3. Faisal, M.; Ali, I.; Khan, M.; Kim, J.; Kim, S. Cyber Security and Key Management Issues for Internet of Things: Techniques, Requirements, and Challenges. *Complexity* **2020**, *2020*, 6619498. [CrossRef]
4. Chen, N.; Qiu, T.; Daneshmand, M.; Wu, D. Robust Networking: Dynamic Topology Evolution Learning for Internet of Things. *ACM Trans. Sens. Netw.* **2021**, *17*, 28. [CrossRef]
5. Nwokoye, C.H.; Madhusudanan, V. Epidemic Models of Malicious-Code Propagation and Control in Wireless Sensor Networks: An Indepth Review. *Wirel. Pers. Commun.* **2022**, *125*, 1827–1856. [CrossRef]
6. Keesen, F.; Castro e Silva, A.; Pinheiro, C.; Arashiro, E.; Ligeiro, Y.; de Grelle, C. New applications of an old individual-based model for biological dynamics. *Ecol. Model.* **2023**, *476*, 110234. [CrossRef]
7. Nepomuceno, E.; Barbosa, A.; Silva, M.; Perc, M. Individual-based modelling and control of bovine brucellosis. *R. Soc. Open Sci.* **2018**, *5*, 180200. [CrossRef]
8. Batista, F.; Martin del Rey, A.; Queiruga-Dios, A. A New Individual-Based Model to Simulate Malware Propagation in Wireless Sensor Networks. *Mathematics* **2020**, *8*, 410. [CrossRef]
9. Martin del Rey, A.; Hernandez, G.; Bustos Tabernero, A.; Queiruga Dios, A. Advanced malware propagation on random complex networks. *Neurocomputing* **2021**, *423*, 689–696. [CrossRef]
10. Fang, Z.; Zhao, P.; Xu, M.; Xu, S.; Hu, T.; Fang, X. Statistical modeling of computer malware propagation dynamics in cyberspace. *J. Appl. Stat.* **2022**, *49*, 858–883. [CrossRef]
11. HaddadPajouh, H.; Dehghantanha, A.; Khayami, R.; Choo, K.K.R. A deep Recurrent Neural Network based approach for Internet of Things malware threat hunting. *Future Gener. Comput. Syst.* **2018**, *85*, 88–96. [CrossRef]

12. Rhode, M.; Burnap, P.; Jones, K. Early-stage malware prediction using recurrent neural networks. *Comput. Secur.* **2018**, *77*, 578–594. [CrossRef]
13. Xinjun, P.; Long, Y.; Shengwei, T. AMalNet: A deep learning framework based on graph convolutional networks for malware detection. *Comput. Secur.* **2020**, *93*, 101792. [CrossRef]
14. Catal, C.; Gunduz, H.; Ozcan, A. Malware Detection Based on Graph Attention Networks for Intelligent Transportation Systems. *Electronics* **2021**, *10*, 2534. [CrossRef]
15. Wu, H.; Luktarhan, N.; Tian, G.; Song, Y. An Android Malware Detection Approach to Enhance Node Feature Differences in a Function Call Graph Based on GCNs. *Sensors* **2023**, *23*, 4729. [CrossRef] [PubMed]
16. Li, T.; Liu, Y.; Liu, Q.; Xu, W.; Xiao, Y.; Liu, H. A malware propagation prediction model based on representation learning and graph convolutional networks. *Digit. Commun. Netw.* **2022**, *9*, 1090–1100. [CrossRef]
17. Leland, W.; Taqqu, M.; Willinger, W.; Wilson, D. On the self-similar nature of Ethernet traffic, IEEE/ACM. *Networking* **1994**, *2*, 1–15. [CrossRef]
18. Ahmadi-Javid, A.; Fathi, M. Design of multi-service systems with facilities functioning as open Jackson queueing networks: Application to online shopping stores. *OR Spectr.* **2022**, *44*, 1255–1286. [CrossRef]
19. Peter, P.; Sivasamy, R. Queueing theory techniques and its real applications to health care systems-Outpatient visits. *Int. J. Healthc. Manag.* **2019**, *14*, 114–122. [CrossRef]
20. Karyotis, V.; Kakalis, A.; Papavassiliou, S. Malware-propagative mobile ad hoc networks: Asymptotic behavior analysis. *J. Comput. Sci. Technol.* **2008**, *23*, 389–399. [CrossRef]
21. Kondakci, S.; Kondakci, D. Building epidemic models for living populations and computer networks. *Sci. Prog.* **2021**, *104*, 00368504211017800. [CrossRef] [PubMed]
22. Roy, S.; Tamimi, A.; Hahn, A.; Xue, M.; Das, S.; Vosughi, A.; Warnick, S. A modeling framework for assessing cyber disruptions and attacks to the national airspace system. In Proceedings of the 2018 AIAA Modeling and Simulation Technologies Conference, Kissimmee, FL, USA, 8–12 January 2018. [CrossRef]
23. Carrasco, G.; Molina, J.L.; Patino-Alonso, M.C.; Castillo, M.D.C.; Vicente-Galindo, M.P.; Galindo-Villardon, M.P. Water quality evaluation through a multivariate statistical HJ-Biplot approach. *J. Hydrol.* **2019**, *577*, 123993. [CrossRef]
24. Frutos Bernal, E.; Martin del Rey, A.; Galindo Villardon, P. Analysis of Madrid Metro Network: From Structural to HJ-Biplot Perspective. *Appl. Sci.* **2020**, *10*, 5689. [CrossRef]
25. Gallego-Álvarez, I.; Galindo-Villardón, M.; Rodríguez-Rosa, M. Analysis of the Sustainable Society Index Worldwide: A Study from the Biplot Perspective. *Soc. Indic. Res.* **2015**, *120*, 29–65. [CrossRef]
26. Martinez-Regalado, J.A.; Murillo-Avalos, C.L.; Vicente-Galindo, P.; Jimenez-Hernandez, M.; Vicente-Villardon, J.L. Using HJ-Biplot and External Logistic Biplot as Machine Learning Methods for Corporate Social Responsibility Practices for Sustainable Development. *Mathematics* **2021**, *9*, 2572. [CrossRef]
27. Escobar, K.M.; Vicente-Villardon, J.L.; de la Hoz-M, J.; Useche-Castro, L.M.; Alarcon Cano, D.F.; Siteneski, A. Frequency of Neuroendocrine Tumor Studies: Using Latent Dirichlet Allocation and HJ-Biplot Statistical Methods. *Mathematics* **2021**, *9*, 2281. [CrossRef]
28. Nieto-Librero, A.B.; Sierra, C.; Vicente-Galindo, M.P.; Ruiz-Barzol, O.; Galindo-Villardon, M.P. Clustering Disjoint HJ-Biplot: A new tool for identifying pollution patterns in geochemical studies. *Chemosphere* **2017**, *176*, 389–396. [CrossRef]
29. Jackson, J. Networks of waiting lines. *Oper. Res.* **1957**, *5*, 518–521. [CrossRef]
30. Galindo, M. Una alternativa de representación simultánea: HJ-Biplot. *Questioó* **1986**, *10*, 13–23.
31. Gabriel, K. The Biplot Graphic Display of Matrices with Applications to Principal Component Analysis. *Biometrika* **1971**, *58*, 453–467. [CrossRef]