A Survey on Epidemiological Propagation Models of Botnets

Ioana APOSTOL

Abstract—**Botnets pose great challenges in defending against cyber threats, as they consist of networks of many infected machines that can be controlled by other entities in order to conduct different types of attacks. Studying botnets evolution and behavior may help in analyzing the effectiveness of the mechanisms used to combat these threats. To this end, many researchers approached the method of modeling and simulating botnets. Concerning the propagation process, current models can be classified in theoretical control models and epidemiological models. This paper comprises a review of five malware epidemiological propagation models, along with a brief analysis of each model, using actual data provided by reports and studies on certain botnets that may be studied using the respective model. The contribution of this work consists in providing a guided tour through some of the previously proposed epidemilogical propagation models, analyzing their applicability and limitations, in order to identify parameters that should be considered in improving propagation modeling.**

Index Terms—**botnets, cyber security, epidemiological models, malware, modeling, propagation**

I. Introduction

Botnets represent networks of infected machines that are controlled by other entities, causing a large number of security threats to the Internet. Bots consist in malware programs developed with the intention of taking control of a large number of hosts that become part of the botnet. Transmitting the bot malware to victim machines in order to increase the botnet determines the propagation phenomenon. Malware programs could be installed in systems in different ways, using multiple means. Therefore, knowing the propagation techniques and its characteristics is imperative when it comes to determining the effective countermeasures to prevent the expansion of the bots.

Most of the binaries through which the infection with a bot is performed include mechanisms that facilitate the propagation to other machines. Depending on the necessity of human intervention, the propagation mechanisms in a botnet can be active or passive. Active propagation does not involve user’s intervention in order to spread the bots, while passive propagation implies, at a certain level, actions made by users.

Modeling enables researchers to predict the damages that can be made by new threats, to understand the behavior of a malware, including the way it is being spread along with other factors involved in this process, but also to evaluate the effectiveness of some countermeasures.

The existing propagation models are focused more on estimating the number of infected hosts, since the exact size of a botnet can only be estimated, and not precisely known. In time, researchers used different methods to measure botnets (infiltration, DNS redirecting, external information), but their results were expressed only through variations. In order to explain these variations, the necessity of theoretical propagation models appeared. Current propagation models can be divided into two main categories: epidemiological models and theoretical control models. The epidemiological propagation models were derived from epidemiological deterministic models and are focused on the number of compromised hosts and their distribution, whereas the control models are focused on detecting and embedding the spread of the malware involved by the botnet.

This paper reviews several malware propagation models inspired by epidemiological models. Most of the models presented in this paper have been proposed in the literature not recently, but they can still be used in studying botnets and can represent a foundation in developing other improved models of propagation.

Each model in this paper is detailed and then discussed, referring to at least one particular botnet that can be studied based on it. Five epidemiological propagation models have been selected: *diurnal propagation model, controlled propagation model, RCS model, the probability model* and *WT-SIR model*. For the first model Conficker botnet was chosen as the case study. The second one is used in trying to explain the Storm Worm activity. RCS model can be reference for Code Red I, whilst the probability model of propagation is claimed to be suitable for Code Red II, although I find this difficult to demonstrate. The last model in this paper is discussed referring to Koobface analysis, but also to Torpig botnet.

The aim of this work is to provide a well-guided tour through some of the epidemiological propagation models proposed by security researchers, to analyze their applicability, along with their limitations in order to highlight which parameters should be taken into

I. APOSTOL is with the Doctoral School for Defense and Security Systems Engineering, Military Technical Academy “Ferdinand I”, Bucharest, Romania (e-mail: ioana.apostol@mta.ro).

Digital Object Identifier 10.32754/JMT.2020.1.05

consideration for proposing improvements in propagation modeling.

II. Epidemiological Models of Malware Propagation in Botnets

Epidemiological malware propagation models are inspired from the epidemiological deterministic models of disease spreading, such as SI, SIS, SIR, SISR models. SI model (*Susceptible-Infected*) implies dividing subjects in only two groups: susceptible subjects and infected subjects. SIS model (*Susceptible-Infected-Susceptible*) is similar with SI model, but it takes into account also the fact that after the subjects get healed they may be again susceptible. SIR model implies three states for subjects: susceptible, infected, recovered. This model presumes that once a subject gets healed, it acquires immunity and cannot be considered susceptible anymore. The following figures describe these three epidemiological models, in the way they were illustrated to study the infectious disease dynamics.

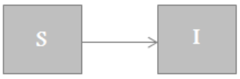

Figure 1. SI epidemiological model states

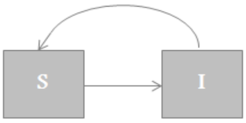

Figure 2. SIS epidemiological model states

Figure 3. SIR epidemiological model states

Regarding the spread of malware in cyberspace, the incipient papers that resorted to epidemiological models presented mostly SI model approaches, in order to determine the total number of infected hosts over time. Further then, SIR and SIS models approaches also appeared in papers addressing malware propagation.

A. Diurnal propagation model

Dagon et al.[1] proposed a propagation model based on SIR epidemiological model, starting from the idea that powered off or offline hosts could not be infected. Therefore, the susceptible hosts are only chosen from online hosts and, as users' activity is more pronounced in daytime than during the night, a behavioral diurnal propagation model was established. The bots in the botnet are grouped based on time zones and the propagation rate varies depending on the time zones' day time.

This propagation model, that takes into account bots diurnal activity, defines a modeling function, $\alpha(t)$, as the number of hosts that are online at the time *t* and can be exploited by the botnet in one region belonging to a specific time zone. $\alpha(t)$ is a periodic function, with a 24 hours period, reaching the highest values during day time and the lowest values during nighttime.

Firstly, Dagon et al. [1] describe the characteristics of propagation for one single region belonging to a specific time zone and then they express these characteristics in multiple regions.

In this model, considering that all the hosts in a network from one region have the same diurnal behavior, the following are defined:

- $I(t)$ – number of infected hosts at the time t;
- $N(t)$ – the total number of hosts that the malware recognizes at the time *t*;
- $S(t)$ – number of vulnerable/susceptible hosts at the time t.

Taking into consideration that a host may also be in one of the states *online* or *offline*, the following relations are described in the diurnal model:

- $I'(t)=\alpha(t)I(t)$ – number of online infected hosts at the time *t*
- $N'(t)=\alpha(t)N(t)$ – number of online hosts from the total amount of hosts, $N(t)$
- $S'(t)=\alpha(t)S(t)$ – number of online susceptible hosts

As the SIR model considers that some subjects, after infection, can be recovered, $R(t)$ is defined as the number of infected hosts that are recovered after infection at the time *t*, and γ represents the parameter of recovery from the classic SIR model.

$$\frac{dR(t)}{dt}=\gamma I'(t) \tag{1}$$

Therefore, this model defines the propagation characteristics in one region as follows:

$$\frac{dI(t)}{dt}=\beta I'(t)S'(t)-\frac{dR(t)}{dt} \tag{2}$$

- $S(t)=N(t)-I(t)-R(t)$
- β – infection ratio defined in epidemiological studies

The main relation that defines diurnal propagation model is:

$$\frac{dI(t)}{dt}=\beta\alpha^2(t)I(t)\left[N(t)-I(t)-R(t)\right]-\gamma\alpha(t)I(t) \tag{3}$$

If multiple regions are considered, the hosts are grouped between the regions, and we can have 24 groups of hosts for the 24 existing time zones. $N_i(t),S_i(t),I_i(t),R_i(t),\alpha_i(t),\gamma_i$ are thus considered the parameters described above for the region *i*. β_{ji} represents the infection ratio from region *j* to region *i*. The propagation characteristics of a malware for region *i* can be defined by the following expressions:

$$\frac{dI_i(t)}{dt}=\sum_{j=1}^{K}\beta_{ji}I_j{}'(t)S_i'(t)-\frac{dR_i(t)}{dt},K\le 24 \tag{4}$$

$$\frac{dI_i(t)}{dt} = \alpha_i(t)\left[N_i(t) - I_i(t) - R_i(t)\right] \cdot \\ \cdot \sum_{j=1}^{K} \beta_{ji}\alpha_j(t) I_j(t) - \gamma_i \alpha_i(t) I_i(t) \tag{5}$$

The limitations of the model proposed by Dagon et al. [1] consist in the fact that this model can be applied only on a malware that uses scanning as the propagation mechanism and it depends on the scanning ratio used for each region. This means that a malware like Storm Worm that was used in the Storm Botnet expansion cannot be the subject of a propagation study using this model, since it used email spam for spreading. Instead, Conficker worm, which relied on scanning a critical vulnerability in the Windows Server, may represent a study case to test the practicability of this diurnal propagation model.

UCSD Network Telescope [11] monitored traffic during the period when the first version of Conficker took off and the analysis performed on the obtained data [12] shows that this malware presents a consistent diurnal pattern, having the peak during 2pm–3pm UTC.

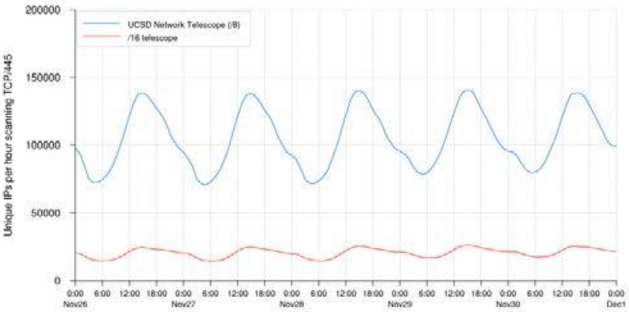

Figure 1. Diurnal pattern of Conficker scanning, as recalled by [12]

Taking into account the data provided in [12], where a significant increase of TCP/445 scanning from different IPs was registered after November 21 midnight, reaching in that same day to a number of over 110000 different IPs that scanned the port for the vulnerability that Conficker was based, it is assumed that 21st November 2008 is the day that Conficker took off and that the scanning hosts were actually infected hosts trying to search for susceptible hosts in order to propagate further the malware. After November 21, the scanning had a relatively stable pattern, a daily periodic one, until another version of the malware appeared. Considering the model proposed by Dagon et al. [1], the stability of the scanning pattern may be given by the recovery ratio, meaning that while other hosts get infected and start to scan for vulnerabilities, almost the same amount of hosts are being recovered. This assumption is questionable since the chances of having the infection ratio almost equal to the rate of recovery daily are extremely small. Beside this, the total number of IP addresses in one region is limited and if it is considered that the recovered hosts cannot be infected anymore, even if the rate of infection remains daily almost equal with the rate of recovery, after a period of time a higher number of hosts from the total amount of hosts get recovered, and the number of susceptible hosts gets smaller and smaller and the number of infected hosts should also get smaller and smaller. A SIS epidemiological model could be considered in this case, or another assumption that can apply to the model proposed in [DAGON]: on November 21 the malware reached to allmost all the susceptible hosts during day time, having a very high rate of infection, while the recovery ratio tends toward 0, and the almost stable scanning activity is mostly diurnal because the infected hosts are in their largest number during the day. This second assumptions is also questionable since it is known that prior to the date Conficker took off, Microsoft announced a security update that resolved the critical vulnerability in Windows Server service, so the possibility of recovering infected hosts should have been increasing or the number of susceptible hosts would start to fall.

It should be emphasized that the *online/offline* state of hosts and the higher activity in malware propagation during day time due to the number of online active hosts is a good consideration, since the report from UCSD Network Telescope data [12] highlights a diurnal pattern in Conficker scanning activity. In addition, grouping hosts according to regions is also a good point for the propagation model, and it should take into consideration not only the time zone, but also the fact that around the globe the IP addresses are unevenly distributed.

B. *Controlled propagation model*

W. Xin-liang et al. [2] proposed another propagation model based on SIR epidemiological model. Considering a controlled propagation in the botnet, the authors analyzed the effects of different propagation rates on the size and stability of the botnet. The control feature of the botnet can easily adapt to the decentralization tendency of this type of network's structure.

Controlled propagation model [2] classifies the infections into three categories:

- Spreaders – infectors responsible with propagating the bots;
- Command and control servers – hosts that do not infect other hosts, they only control them;
- Hidden nodes – bots used only for launching an attack from the botnet.

Thereby, the specific parameters of the SIR model can be adapted in the following way:

N – the total number of hosts in the network;

$I(t)$ – the number of infected hosts at the time t;

$S(t)$ – the number of susceptible hosts at the time t;

$R(t)$ – the number of recovered hosts after infection;

$\alpha_p(t)$ – the infection ratio of the hosts that are meant to become spreaders;

$\alpha_s(t)$ – the infection ratio of the hosts that are meant to become control servers;

Digital Object Identifier 10.32754/JMT.2020.1.05

$\alpha_a(t)$ – the infection ratio of the hosts that are meant to become hidden nodes;

$\beta(t)$ – the infection ratio of network hosts at the time *t*;

$\gamma_p(t)$ – recovery ratio of the spreaders (propagating hosts);

$\gamma_s(t)$ – recovery ratio of the control servers infected hosts (second type of infection);

$\gamma_a(t)$ – recovery ratio of the hidden nodes.

Considering the processes involved in SIR epidemiological model, this controlled propagation model [2] can be outlined by Fig. 5.

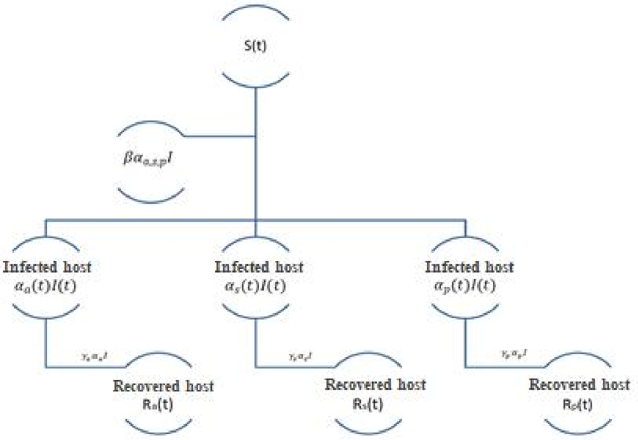

Figure 2. State diagram for the controlled propagation model

The following relations describe the model outlined in Fig. 5:

$$\begin{aligned}
&\frac{dI(t)}{dt}=\beta(t)\alpha_p(t)I(t)S(t)-\frac{dR(t)}{dt}\\
&\frac{dS(t)}{dt}=-\beta(t)\alpha_p(t)I(t)S(t)\\
&\frac{dR(t)}{dt}=\frac{dR_p(t)}{dt}+\frac{dR_s(t)}{dt}+\frac{dR_a(t)}{dt}\\
&\frac{dR_p(t)}{dt}=\gamma_p(t)\alpha_p(t)I(t)\\
&\frac{dR_s(t)}{dt}=\gamma_s(t)\alpha_s(t)I(t)\\
&\frac{dR_a(t)}{dt}=\gamma_a(t)\alpha_a(t)I(t)\\
&\alpha_p(t)+\alpha_s(t)+\alpha_a(t)=1\\
&N=S(t)+I(t)+R(t)
\end{aligned}\tag{6}$$

For studying the impact of the propagation ratio on the botnet's dimension and stability, W. Xin-liang et al. [2] considered $\gamma_s=\gamma_a,\gamma_p>\gamma_a$, since the spreaders have a greater influence in botnet expansion. Thus, the equations defining the model become:

$$\begin{cases}
\frac{dI(t)}{dt}=\beta(t)\alpha_p(t)I(t)S(t)-\frac{dR(t)}{dt}\\
\frac{dS(t)}{dt}=-\beta(t)\alpha_p(t)I(t)S(t)\\
\frac{dR(t)}{dt}=\gamma_p(t)\alpha_p(t)I(t)+\gamma_a(t)\left(1-\alpha_p(t)\right)I(t)\\
N=S(t)+I(t)+R(t)
\end{cases}\tag{7}$$

Authors describing the controlled propagation model [2] conclude that if the propagation ratio is too big, botnet's stability decreases and the infected hosts may be detected more rapidly. If the propagation ratio si neither too high, nor too small, the number of infected hosts may increase monotonically with propagation ratio.

Controlled propagation model was destined to address above all the decentralization trend of botnets. So, a suitable example of a known botnet that could be studied using this model would be a P2P botnet, such as the Storm worm.

Storm worm uses the e-mail as the propagation vector. E-mails containing various messages to trick the recipient into opening an attachment or clicking on a link were sent in different spam campaigns used by Storm. [3] describes Storm as a botnet that utilizes only a part of the bots to propagate the malware and other less bots are used for the command and control function, while some other bots wait to receive commands in order to be executed. These characteristics are the same ones that were taken into account in the model proposed by W. Xin-liang et al. [2].

Storm's behavior is characterized by "patience". After a worm attack is conducted, it shuts off for a while, hiding more easily. This type of behavior, along with the propagation vector that implies the user's intervention for infection, gives Storm worm a lower propagation ratio, since not all the users that receive spams launched in Storm campaignes get tricked. On the other hand, the messages sent in Storm campaignes contained news or events of public interest, the social engineering theme being changed quite often, making it more possible for users to get tricked into accessing the attachment or the link embedded inside the email and keeping the propagation ratio not too low. Thus, the conclusions formulated around the controlled propagation model in [2] are well founded when it comes to the Storm botnet: keeping a propagation ratio neither too high, nor too small, a botnet using a controlled propagation mechanism cannot be easily detected and gets more stability. But this is an isolated case, since not all P2P botnets operate in this manner concerning their propagation. Bots are not generally grouped as presented in controlled propagation model. For example, in Mirai botnet, as described in [4], all bots may participate in the scanning process, trying to infect vulnerable devices, and all the same bots may execute commands received from the C&C Servers.

Digital Object Identifier 10.32754/JMT.2020.1.05

C. RCS Model

The first large scale botnet that successfully targeted enterprise networks was Code Red worm. The first version of this worm attacked around 360 000 computers running Microsoft IIS web server. Staniford [5] studied Code Red by developing a propagation model called RCS model (Random Constant Spread).

RCS model assumes that a host cannot be compromised several times and includes the fact that Code Red worm uses a random number generator for the infection process. This model may be considered inspired by SI epidemiological model since it considers the total number of vulnerable servers (susceptible parts) which can be compromised, ignoring the patching of systems during the worm spread, turning on/off the systems or the removal of the worm from the systems. The variables taken into account for RCS model are:

- *K* – constant average compromise rate, that depends on the bot's processing speed, the bandwidth of the network and the location of the infected host;
- *a*(*t*) – the proportion of vulnerable hosts that have been compromised by the time *t*;
- $N \cdot a(t)$ – the number of infected hosts, each of them being able to scan other vulnerable hosts with a *K* rate (*N* represents a constant large number of IP addresses);

Since a portion of the vulnerable hosts $(a(t))$ have already been infected, there are maximum $K\ (1-a(t))$ new infections that an already infected host might generate. The number of hosts which can be compromised during the time interval *dt* can be given by the following relation:

$$n = (Na)K(1-a)dt$$

$$Nda = (Na)K(1-a)dt$$

$$\frac{da}{dt} = Ka(1-a) \quad (8)$$

$$a = \frac{e^{K(t-T)}}{1+e^{K(t-T)}}$$

This model [5] may predict the number of infected hosts at the time *t*, if the compromise rate (*K*) is known. The study of Staniford [5] concludes that the higher the rate of compromise (*K*), the faster the malware will reach the saturation phase. This is a valid theory concerning Code Red I, since it is known that the worm spread very rapidly and after compromising almost all vulnerable IIS servers on the Internet, the spreading stopped.

D. The probability model of propagation

Another model that can be used in studying Code Red is the one proposed by Wang et al. [6]. In classical epidemiological models only the states of the hosts are taken into consideration, and these features may not be so efficient when it comes to malware propagation in P2P communications, where the recovery strategy of the hosts must be optimized.

The model proposed by Wang et al. [6] is based by propagation probabilities and is aimed to find a method of counteracting P2P botnets. This model takes into account three main parameters: the infected state, the vulnerability distribution and the patch (recovery) strategy.

In P2P communications each two nodes are considered absolutely connected. Thus, a node is able to propagate malware at any time with a certain probability. P2P network representations are often made through graphs. The probability model of propagation is based on a square matrix $P_{n,n}$, called topology propagation matrix (TPM), with elements like t_{ij} representing the probability of propagating the malware from the node *i* to the nod *j*, when the node *i* is infected. Each line of this matrix indicates the probabilities of propagating the malware from an infected node to the other nodes, $t_{i,j}$ equaling 0, since an infected node cannot infect itself again.

$$P = \begin{bmatrix} t_{11} & \cdots & t_{1n} \\ \vdots & \ddots & \vdots \\ t_{n1} & \cdots & t_{nn} \end{bmatrix}, t_{ij} = p\left(N_j \middle| N_i\right), t_{ij} = 0 \leftrightarrow i = j \quad (9)$$

In fact, malware propagation from *i* node to *j* node is achieved through several intermediate nodes (*k*), so the propagation probability $t_{ij}^{(k)}$ is defined as follows:

$$t_{ij}^{(k)} = p\left(N_j^{(k)} \middle| N_i \cup \overline{N}_i^{(k-1)}\right) = \sum_{m=1}^{m=n, m\neq i} t_{im}^{(k-1)} t_{mj} \,,$$

$$k \in [1, n-2], i, j = 1, \ldots, n \quad (10)$$

$\overline{N}_i^{(k-1)}$ – means all nodes except N_i

In the model [6] γ function is defined: $\gamma^k(P) = \overbrace{P \;*\; P \;*\; \cdots \;*\; P}^{k+1}, \gamma^0(P) = P, \gamma^1(P) = P * P$, where „*" represents the multiplying operator between two matrices.

Given that the propagation from one node to another is made through *k* intermediary nodes, TPM becomes:

$$P^{(k)} = \begin{bmatrix} t_{11}^{(k)} & \cdots & t_{1n}^{(k)} \\ \vdots & \ddots & \vdots \\ t_{n1}^{(k)} & \cdots & t_{nn}^{(k)} \end{bmatrix} = \gamma^k(P) \quad (11)$$

For the infection state, the model defines an infection vector containing 1 and 0 values. When the value is 1, it means that the node is infected and it can propagate malware with maximum probability. When the value is 0, the node is not infected and cannot propagate malware.

$$S = [s_1, s_2, \ldots, s_i, \ldots s_n], s_i = 0 \text{ or } 1 \quad (12)$$

At first, it is considered that all TPM nodes are vulnerable, so in the propagation process each intermediary node can be infected. After reaching k nodes through propagation, S vector can be defined as follows:

$$S^{(k)} = \left[\left[S^{(k-1)}\right] * \gamma\left(S^{(k-1)} \&_L P^{(k-1)}\right)\right], \tag{13}$$

where $\&_L$ is *Left Logic AND* operation between a vector and a matrix.

The infection probability is:

$$P_S^{(k)} = \gamma\left(S^{(k-1)} \&_L P_S^{(k-1)}\right) \tag{14}$$

A node without vulnerabilities cannot be infected. So, in TPM, another vector is considered. This vector represents the vulnerability distribution and contains also 0 and 1 values. 1 is for a vulnerable node, while 0 is for a node without vulnerabilities that cannot be infected.

$$V = \left[\upsilon_1, \upsilon_2, \ldots, \upsilon_i, \ldots, \upsilon_n\right], \upsilon_i = 0 \text{ or } 1 \tag{15}$$

The propagation probability through vulnerable nodes is:

$$P_{s\upsilon}^{(k)} = \gamma\left(P_s^{(k-1)} \&_R V\right) \tag{16}$$

where $\&_R$ is *Right Logic AND* operation between a vector and a matrix.

As the model also takes into consideration the possibility of recovering nodes from the infected state, a patching vector is also defined. For this vector, the value 1 indicates a recovered node, while 0 represents the vulnerable nodes.

$$Q = \left[q_1, q_2, \ldots, q_i, \ldots, q_n\right] \tag{17}$$

TPM is finally represented as follows:

$P_{s\upsilon q}^{(k)} = \gamma\left(P_{s\upsilon}^{(k-1)} \&_R \left(V \underline{\&} Q\right)\right)$, where $\underline{\&}$ is defined in the table below.

TABLE I. BINARY OPERATOR $\underline{\&}$

V	Q	$V \underline{\&} Q$
1	1	0
0	1	0
1	0	1
0	0	0

An advantage of the probability model of propagation is that, along with the propagation capacity, it also defines the quarantine capacity. The propagation capacity is expressed by the number of nodes through which a malware can be propagated with the highest probability. The quarantine capacity is represented by the number of nodes that have the lowest probability of infection.

The model presented by Wang et al. [6] is a theoretical mathematical model. In developing this model no real data were used and no other parameters involved in the propagation were considered, such as the influence of the number of nodes in P2P topology or the possibility of eliminating the malware from infected hosts but keeping them still vulnerable. The authors of this model [6] claim that it can be applied for studying Code Red II botnet (the second edition of Code Red worm). Applying Code Red II patterns into this model can be challenging, since although the propagating worm probes neighbor hosts with a higher probability, it is known that it was programmed to spread more aggressively in China than anywhere else. [7] If the infected system is identified having the language set to Chinese, the scanning is more aggressive, so the propagation probability should increase considerably and is a variable parameter in time.

E. WT-SIR Model

Yang et al. [8] proposed a propagation and evolution model of botnets (WT-SIR), taking into consideration a characteristic that is often found in the Internet: the possibility of the infected or recovered bots to pass back into the susceptible state. While traditional models assume achieving propagation only through active mechanisms (scanning), this model can also be addressed in the case of malware propagation through web services (e.g. Web Trojan).

If malicious code is inserted into a well-known website, the probability of the infected web page being visited is quite high. Hence, WT-SIR model [8] includes the following cases:

- A node denoted *A* accesses the compromised page and gets infected, but without acknowledging what happened. If *A* visits other infected sources, his state changes from “infected” to “vulnerable”;
- Node *B* visits a web page and gets infected, but the compromising source is discovered and the bot is eliminated, without getting immunity, thus changing the state from “recovered” to “vulnerable”;
- Node *C* is infected and permanently restores the bot on the victim, waiting for other hosts to connect, and following, through this, to transmit the bot.

For all these three cases it is assumed [8] that the probability of reinfecting a node is μ. Following the SIR model, α is the probability of initial infection and β represents the probability of recovering an infected node. In Fig. 6 the state diagram of this model is depicted.

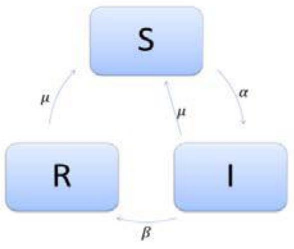

Figure 6. WT-SIR state diagram

The coefficient μ lowers the speed of extending a botnet, since reinfecting a previously infected node does not increase the number of infected hosts. As stated in [8], μ also lowers the capacity of recovering infected nodes, since it gives the ability to reinfect a host.

In a real case scenario, it is possible that, due to the fact that the vulnerability gains notoriety, special instruments for detecting and eliminating are produced and, therefore, some nodes become immune. The model proposed in [8] also takes into account an *immunization rate*, denoted ρ. The efficiency of reducing the number of infected hosts is proportional to this coefficient, ρ. Hence, the number of completely recovered hosts from the botnet is $\left(\mu+\beta\right)\rho I\left(t\right)$.

WT-SIR model is defined by the following differential equation:

$$\begin{cases} \dfrac{dS(t)}{dt} = -\alpha I(t)S(t) + \mu\left[I(t)+R(t)\right] \\ \dfrac{dI(t)}{dt} = -\alpha I(t)S(t) - (\beta+\mu)(1-\rho)I(t) \\ \dfrac{dR(t)}{dt} = \beta I(t) - \mu R(t) \end{cases} \tag{18}$$

Yang et al. [8] also performed experimental studies concerning their model, observing the evolution of infection in time for different values of the parameters taken into consideration in this model. Authors claim that this model is suitable for passive propagation botnets, but they do not associate it with a specific named botnet. Their conclusion is that following this model of propagation, a botnet does not get too extended and, although at the beginning the number of infected hosts increases rapidly, the botnet gains some stability after the number of nodes controlled by the botnet drops due to isolation of discovered infections.

Koobface is an example of a botnet that uses passive propagation mechanisms. It is based on a computer worm that takes advantage of users through social network messages in Facebook, Twitter, MySpace and others. Koobface needs user's intervention for infecting hosts. In fact, most of the passive propagation botnets are based on users' intervention to expand making the infection ratio inversely proportional to the users' awareness about threats in cyberspace.

K. Thomas and D.M Nicol [9] explored and analyzed Koobface's activity during a certain period of time. Their analysis shows an evolution somewhat similar to the one depicted by WT-SIR model [8], although by the time of this analysis, the botnet had already expanded and gained stability. The authors [9] also highlight the importance of developing good countermeasure techniques for diminishing this threat. Koobface had a real impact on social networking platforms as it demonstrated a high capacity to evade shutdown and continue propagating. K.B. Tanner et al. [10] present the modification made by Koobface in the Command and Control infrastructure to avoid detection. These evading techniques often characterizing cyber threats have an influence on the rate of recovering infected hosts, but also on the capacity of reinfecting hosts. Consequently, I believe that WT-SIR model [8] could be improved by adding another coefficient that expresses the rate of the obfuscation techniques' efficiency.

Another passive propagation botnet example is Torpig. This botnet infected hosts through drive-by-download attacks. Vulnerable web pages belonging to legitimate sites were modified so that the web browser sends JavaScript code requests from another website controlled by the attackers. The work presented in [13] consists in taking over Torpig botnet for ten days in order to study its behavior along with its magnitude. Compared to the results obtained from the study based on WT-SIR model [8], when analyzing Torpig new infections, B. Stone-Gross et al. [13] observed a consistent diurnal pattern of infections, feature that was not taken into account in WT-SIR model [8]. However, the analysis performed inside Torpig [13] mentions an initial spike, when the bots started to contact the infiltrated server, behavior that is also present in the charts resulting from WT-SIR model [8].

III. Conclusions and Future Work

Identifying propagation characteristics may represent a method of bot detection and botnet countering. During the propagation process, network and hosts characteristics may vary so much that a valid general model of propagation has not yet been formulated.

Most of the propagation models proposed by researches focus more on estimating the number of infected machines and justifying the evolution of botnets, most of these models being based on epidemiological theories.

In this paper five epidemiological propagation models were studied and also discussed concerning the extent to which they can be used in characterizing certain known botnets. The contribution of this work consists in making several epidemiological propagation models available all together and conducting an analysis of each model regarding specific botnets that would fall into the category for which the model in question was proposed. Some limitations of the studied models have also been observed and this leads us to consider it the foundation of our future work, consisting in improving botnet modeling by proposing a new propagation model that would take into consideration other parameters determined by the behavior of some more recent botnets.

As also stated in this paper, the propagation models proposed so far are focused on characterizing a specific type of botnet, since these threats differ and evolve very much and their features cannot be generalized. Besides, all the traditional models studied so far can only approach what is happening in the real environment. The current cyber landscape continues to bring a lot of challenges.

References

[1] D. Dagon, C. Zou, W. Lee., "Modeling botnet propagation using time zones", in *Proc. of the 13th Network and Distributed System Security Symposium NDSS*, 2006.

[2] W. Xin-Iiang, C. Lu-Ying, L. Fang, and L. Zhen-Ming, "Analysis and Modeling of Botnet Propagation Characteristics", *Wireless Communications Networking and Mobile Computing (WiCOM), 2010 6th International Conference*, doi:10.1109/WICOM.2010.5601301

[3] B. Smith, "A Storm (Worm) Is Brewing" [J]. IEEE Computer Society Press, 2008, 41(2): 20-22. doi:10.1109/MC.2008.38

[4] Z. Ling, K. Liu, Y. Xu, C. Gao, Y. Jin, C. Zou, X. Fu, W. Zhao. "IoT Security: An End-toEnd View and Case Study". CoRR, 2018. arXiv:1805.05853.

[5] S. Staniford, V. Paxson, N. Weaver, "How to 0wn the Internet in Your Spare Time," in *Proc. of the 11th USENIX Security Symposium,* 2002.

[6] Y. Wang, S. Wen, W. Zhou, W. Zhou, Y. Xiang. "The probability model of Peer-to-Peer botnet propagation," in *Proc. of the 11th International Conference, ICA3PP*, Part I, pp. 470-480, 2011. doi:10.1007/978-3-642-24650-0_41

[7] D. Moore, C. Shannon, and K. Claffy, "Code-Red: a case study on the spread and victims of an Internet worm." pp. 273-284, 2002. doi:10.1145/637201.637244.

[8] Y. Yang, W. Niu, G. Hu, H. Li, "A botnet passive propagation and evolution model," *Second International Conference on Instrumental & Measurement, Computer, Communication and Control*, 2012. doi:10.1109/IMCCC.2012.23

[9] K. Thomas and D. M. Nicol, "The Koobface botnet and the rise of social malware," *2010 5th International Conference on Malicious and Unwanted Software*, Nancy, Lorraine, 2010, pp. 63-70, doi:10.1109/MALWARE.2010.5665793.

[10] B. K. Tanner, G. Warner, H. Stern and S. Olechowski. "*Koobface: The evolution of the social botnet*". 2010 eCrime Researchers Summit. doi:10.1109/ecrime.2010.5706694

[11] UCSD Network Telescope https://www.caida.org/projects/network_telescope/

[12] https://www.caida.org/research/security/ms08-067/conficker.xml

[13] B. Stone-Gross, M. Cova, L. Cavallaro, B. Gilbert, M. Szydlowski, R. Kemmerer, C. Kruegel, and G. Vigna, "Your botnet is my botnet: Analysis of a botnet takeover" in *CCS '09:Proc. of the 16th ACM Conference on Computer and Communications Security*. ACM, 2009, pp. 635–647. doi:10.1145/1653662.1653738