Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2008, Article ID 940526, 11 pages
doi:10.1155/2008/940526

Research Article

Dynamical Models for Computer Viruses Propagation

José R. C. Piqueira and Felipe Barbosa Cesar

Escola Politécnica da Universidade de São Paulo, Avenida Prof. Luciano Gualberto, travessa 3 - 158, 05508-900 São Paulo, SP, Brazil

Correspondence should be addressed to José R. C. Piqueira, piqueira@lac.usp.br

Received 28 March 2008; Revised 9 May 2008; Accepted 30 May 2008

Recommended by Jose Balthazar

Nowadays, digital computer systems and networks are the main engineering tools, being used in planning, design, operation, and control of all sizes of building, transportation, machinery, business, and life maintaining devices. Consequently, computer viruses became one of the most important sources of uncertainty, contributing to decrease the reliability of vital activities. A lot of antivirus programs have been developed, but they are limited to detecting and removing infections, based on previous knowledge of the virus code. In spite of having good adaptation capability, these programs work just as vaccines against diseases and are not able to prevent new infections based on the network state. Here, a trial on modeling computer viruses propagation dynamics relates it to other notable events occurring in the network permitting to establish preventive policies in the network management. Data from three different viruses are collected in the Internet and two different identification techniques, autoregressive and Fourier analyses, are applied showing that it is possible to forecast the dynamics of a new virus propagation by using the data collected from other viruses that formerly infected the network.

1. Introduction

A few decades ago, computer viruses arose in the form of programs with simple code and able to undermine the smooth operation of a machine. Initially, in spite of the large number of viruses, they caused minor damages to machinery and their spread was very slow. Over the years, due to the rapid development of technology, such as software and hardware, the development and popularization of the Internet and the great variety of equipment using software and networks, viruses have become a major threat [1].

Currently, these virus programs have more complex codes, being able to produce mutations of themselves, and their detection and removal by antivirus programs became more difficult [2]. Their goals go much further than simply damaging a machine. They are capable of acquiring personal data of users of networks, such as a bank account, and cause severe damages to large corporations [3].

In view of these concerns, a better understanding of the computer viruses spreading dynamics is mandatory. To improve the safety and reliability in computer systems and networks, it is important to have the capacity of recognizing and combating the several types of infections faster and more effectively [4, 5].

Research actions started at the end of the 80s with the classical paper of Kephart et al. [6] proposing an ecosystem approach for computational systems. Then, the efforts were concentrated on the development of antivirus programs, responsible for the detection and removal of viruses, based on the previous recognition of the infection code based on the models shown in [2, 7, 8]. These programs have a great upgrading power, but act just as simple vaccines against diseases [2, 4]. They are not able to predict the behavior of networks when an infection is established in a machine and, consequently, cannot support preventive attitude against virus actions based on events of the network.

The first effort to produce models for the spreading of computer viruses based on their epidemiological counterparts is reported in [7] with the initial ideas for deriving long-term behaviors considering the graph representing the network connections. Then, with Markov chains representing the local behavior of infection action in a single node, susceptible-infected-removed (SIR) models were presented trying to fit the long-term behavior of the viruses propagation [9].

This kind of approach had some attention in the last five years and the relations between spreading viruses and topological parameters of the network were studied, being successful mainly when modeling the propagation by email networks [10]. Besides, SIR models were modified [5] and applied to guide infection prevention [11, 12], deriving expressions for epidemiological thresholds [11–13].

This work focuses on the achievement of models for the dynamics of the spread of certain viruses, mainly taking into account the correlation functions between the several viruses spreading data, during a certain period of time. Thus, the number of infections from a type of virus could be foreseen in the short term by comparison with other viruses or with notable events in the network, which would support preventive policies.

In order to provide simple algorithms to allow operational facility, simple autoregressive models are chosen [14, 15]. Considering the periodicity of the data collected, Fourier models are also tried, producing the same results of the autoregressive ones.

2. Methodology

The data to be collected for modeling computer infections propagation are the number of daily, weekly, and monthly infections for several computer viruses. These numbers are found in the Internet, for instance, in http://www.avira.com/, and support the development of linear identification models.

The next step is the choice of a specific virus to be analyzed, in the enormous range of possibilities. In this work, a premise was taken into consideration: in order to have an efficient identification, the several chosen viruses need to present similar propagation dynamics. Here, the high incidence of cases reported and the email spreading compose the chosen criterion.

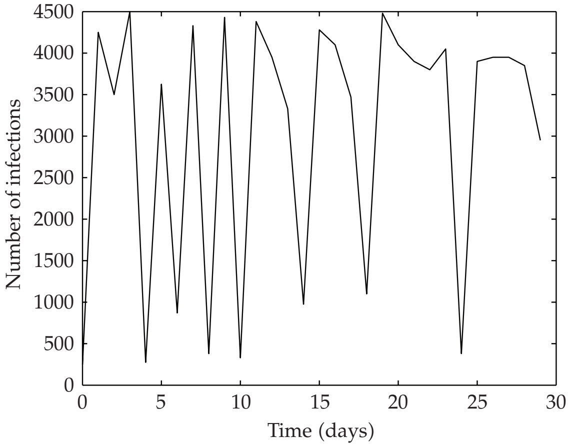

Figure 1: *Wormnetsky.p* temporal evolution.

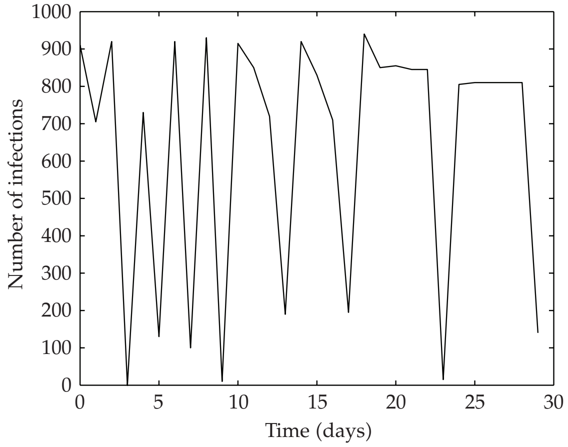

Figure 2: *Wormmytob.mr* temporal evolution.

$Wormnetsky.p$, $wormmytob.mr$, and $trdir.stration.ge$ were chosen, that is, two worms and a trojan. Figures 1, 2, and 3 show the dynamical evolution of the number of infections with $wormnetsky.p$, $wormmytob.mr$, and $trdir.stration.gen$, respectively.

First, in order to verify the relations among the viruses, cross-correlation coefficients are calculated. Considering two signals $x(t)$ and $y(t)$ simultaneously sampled in regular T intervals, and calling $x(nT)$ and $y(nT)$ their n samples, for a certain time interval containing N sample periods, the cross-correlation coefficient, ρ, between $x(t)$ and $y(t)$ measures how they are related with each other in this interval (see [16, page 206]). Table 1 summarizes the cross-correlation coefficients, calculated for the three pairs of infection signals, for the time interval of Figures 1, 2, and 3, sampling the data daily.

The results from Table 1 indicate acceptable correlation between the spread of the viruses chosen, corroborating the visual similarity between the temporal evolution of the three infections. Due to this, only *wormnetsky.p* is considered to identify the system parameters to be

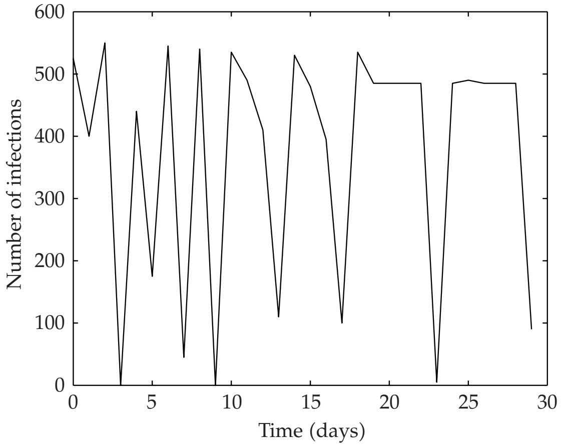

Figure 3: *Trdir.stration.gen* temporal evolution.

Table 1: Viruses cross-correlation coefficients.

Viruses	Cross-correlation coefficient
Wormnetsky.p, *wormmytob.mr*	0.39312
Wormnetsky.p, *trdir.stration.ge*	0.40541
Wormmytob.mr, *trdir.stration.ge*	0.99435

used to provide short-term forecasts for the three viruses. Following this identification strategy, model accuracy is checked.

3. System identification algorithms

In order to identify the parameters to model the temporal evolution of the infections by the three types of viruses selected here, two approaches were followed:

(i) using a linear autoregressive model, that is, consider that the current value of a variable depends only on the former values, up to a certain delay [14, 15];

(ii) identifying the main frequencies of the time series and treating them as Fourier series [14, 15].

3.1. Autoregressive model

Considering a regularly sampled signal $y(k)$, its estimated value at instant k is given by

$$y(k) = \sum_{i=1}^{d} p_i y(k-i), \tag{3.1}$$

where p_i are the model parameters to be estimated by using the minimum square method, and d is the maximum delay to be considered [14, 15], measured by the number of sampling intervals.

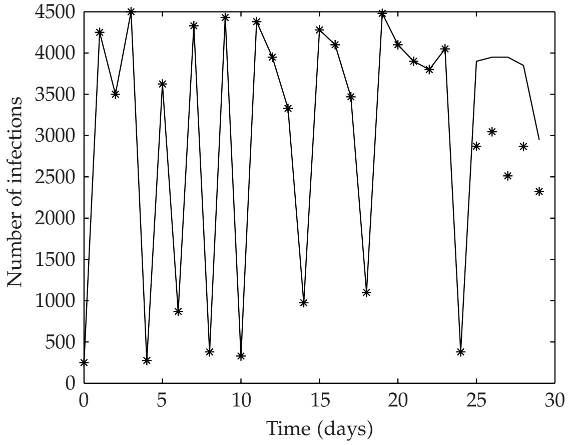

Figure 4: *Wormnetsky.p* temporal evolution simulation ($d = 10$).

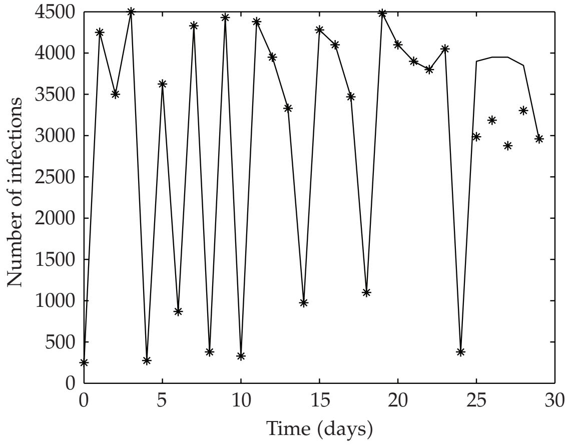

Figure 5: *Wormnetsky.p* temporal evolution simulation ($d = 15$).

By using a "free-prediction" strategy, the vector data are divided into two parts: one is used for the identification of the system parameters and the other for the simulation and validation of the model. In the case of the data described in Section 2, the 25 first samples are used for identification and the last 5 for simulation. Different values of d are considered and Figures 4, 5, and 6 show the results for d equal to 10, 15, and 24, respectively, with the continuous line representing the real data and the asterisks representing the simulation results.

In order to compare the several chosen delays, Table 2 shows the mean-square estimation error in each case. Considering these results, from now on, all models will use $d = 15$.

To have an idea of the efficiency of the adopted identification strategy, the estimated parameters for *wormnetsky.p* are used to model the dynamics of *wormmytob.mr* and *trdir.stration.gen*. The results are shown in Figures 7 and 8, respectively, with the continuous line representing the real data and the asterisks representing the simulation results. Table 3 summarizes the mean-square errors of these simulations.

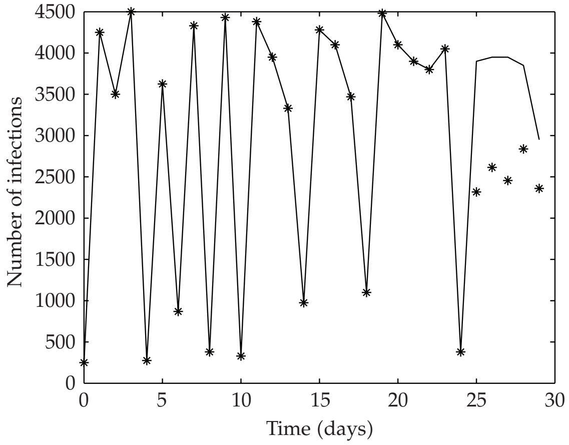

Figure 6: *Wormnetsky.p* temporal evolution simulation ($d = 24$).

Table 2: Estimation errors in autoregressive models for *wormnetsky.p*.

d	Mean-square estimation error(%)
10	5.5773
15	4.1751
24	6.8542

Table 3: Estimation errors in autoregressive models for *wormmytob.mr* and *trdir.stration.ge*.

Virus	Mean-square estimation error(%)
Wormmytob.mr	17.142
Trdir.stration.ge	18.32

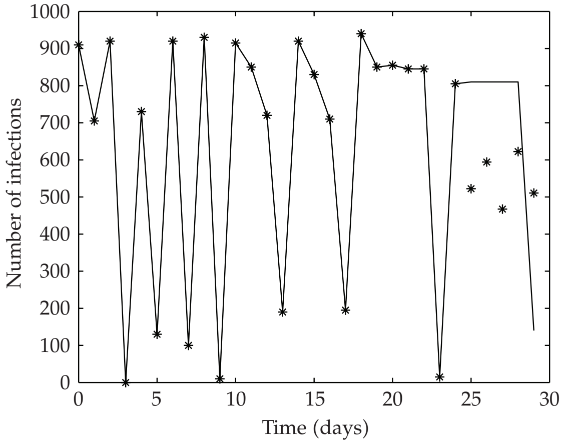

Figure 7: *Wormmytob.mr* temporal evolution simulation.

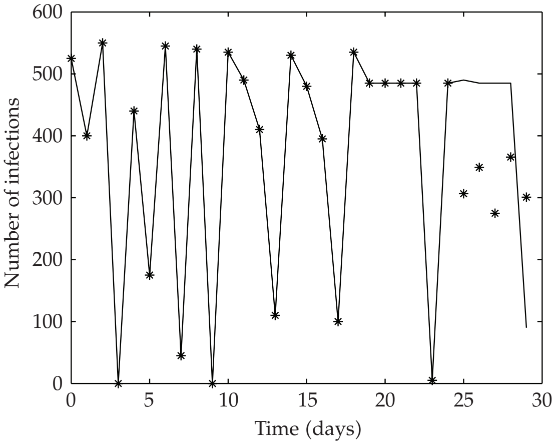

Figure 8: *Trdir.stration.gen* temporal evolution simulation.

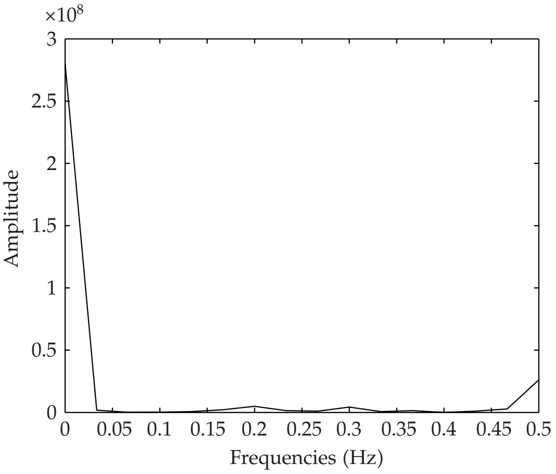

Figure 9: *Wormnetsky.p* frequency spectrum.

The simulations performed taking into account only the parameters calculated for the *wormnetsky.p* show that the short-term estimations of new infections are not precise for *wormmytob.mr* and *trdir.stration.ge*, as expected, because the same model is used for different viruses. Nevertheless, the model is able to predict with some accuracy the increasing and decreasing tendencies in their dynamics. This knowledge permits the implementation of preventive policies, considering only the *wormnetsky.p* propagation profile.

3.2. Fourier series model

Observing the strong oscillatory character of the three different viruses studied, a model considering the signals as a sum of cosines was developed. Figures 9, 10, and 11 present the frequency spectrum for the temporal evolution of the *wormnetsky.p*, *wormmytob.mr*, and

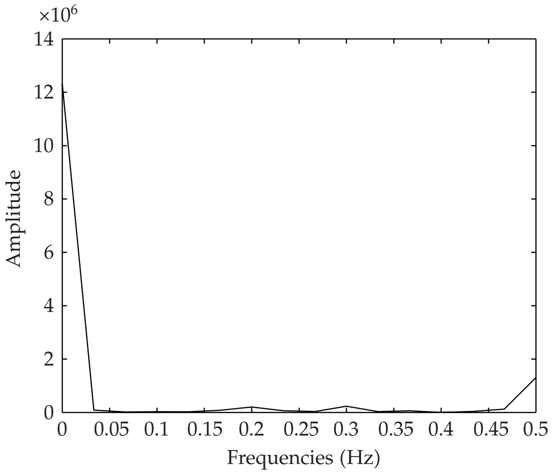

Figure 10: *Wormmytob.mr* frequency spectrum.

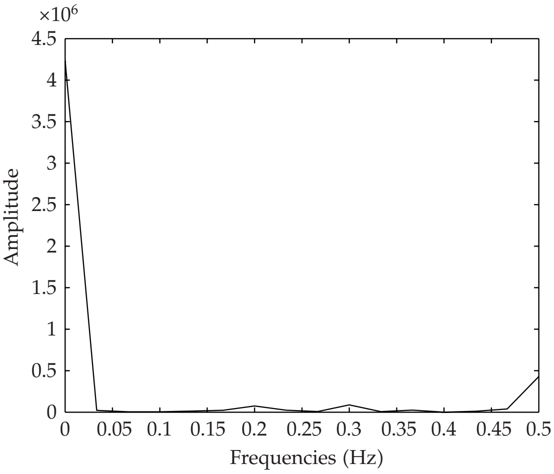

Figure 11: *Trdir.stration.gen* frequency spectrum.

trdir.stration.ge propagation. As one can see, the main frequencies of the three dynamic behaviors are the same.

Figures 9, 10, and 11 indicate that a good set of frequencies for developing the model is F = [0 0.2 0.3 0.5]. Following the same reasoning used in Section 3.1 for identification, the model parameters are calculated by using only the data from *wormnetsky.p* and the predictions of new infections for *wormmytob.mr* and *trdir.stration.ge* are obtained by using the same parameters.

To have an idea about the efficiency of the adopted identification strategy by using Fourier methods, Figures 12, 13, and 14 show the predicted dynamics of *wormnetsky.p*, *wormmytob.mr*, and *trdir.stration.gen*, respectively, with the continuous line representing the real data and the asterisks representing the simulation results. Table 4 summarizes the mean-square errors of these simulations.

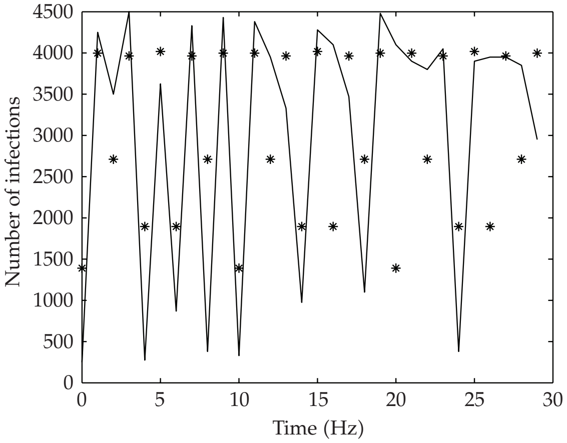

Figure 12: *Wormnetsky.p* Fourier model.

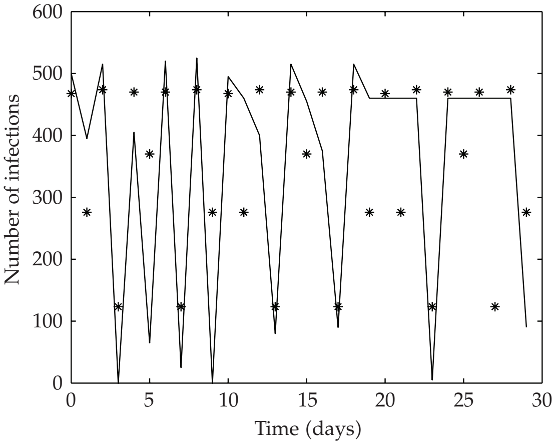

Figure 13: *Wormmytob.mr* Fourier model.

As in autoregressive models, simulations performed taking into account only the parameters calculated for the *wormnetsky.p* show that the short-term estimations of new infections are not precise for *wormmytob.mr* and *trdir.stration.ge*, as expected, because the same model is used for different viruses. But, again, the model is able to predict with some accuracy the increasing and decreasing tendencies in their dynamics, allowing to establish preventive policies by using only the data from *wormnetsky.p* propagation.

4. Conclusions

Two different models for the dynamics of computer viruses propagation were compared: autoregressive and Fourier analysis presenting similar results. They provide good predictions for three different types of infections by using the data collected for just one of them.

In spite of not being totally satisfactory, these models present the possibility of predicting increasing and decreasing tendencies in the propagation of a certain type of virus by using the

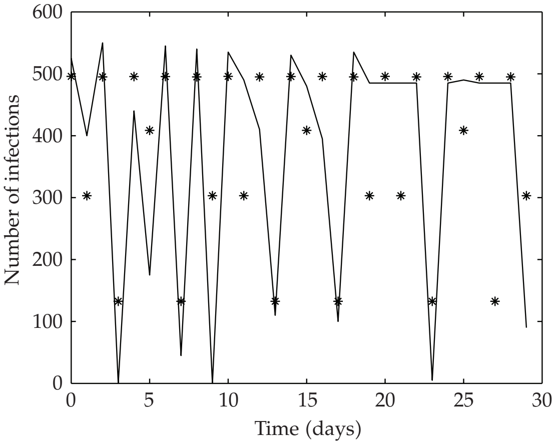

Figure 14: *Trdir.stration.gen* Fourier model.

Table 4: Estimation errors in Fourier models for *wormnetsky.p*, *wormmytob.mr*, and *trdir.stration.ge*.

Virus	Mean-square estimation error(%)
Wormnetsky.p	7.0354
Wormmytob.mr	16.778
Trdir.stration.ge	18.096

accumulated experience with another one. It seems that this point could be used to predict and control the infection levels in advance, providing preventive actions in order to increase safety and reliability.

Acknowledgments

The first author is supported by FAPESP and CNPq and the second author is supported by the Brazilian Oil Agency.

References

[1] P. J. Denning, *Computers under Attack*, Addison-Wesley, Reading, Mass, USA, 1990.

[2] P. S. Tippett, “The kinetics of computer virus replication: a theory and preliminary survey,” in *Safe Computing: Proceedings of the 4th Annual Computer Virus and Security Conference*, pp. 66–87, New York, NY, USA, March 1991.

[3] F. Cohen, “Models of practical defenses against computer viruses,” *Computers & Security*, vol. 8, no. 2, pp. 149–160, 1990.

[4] S. Forrest, S. A. Hofmayer, and A. Somayaji, “Computer immunology,” *Communications of the ACM*, vol. 40, no. 10, pp. 88–96, 1997.

[5] J. R. C. Piqueira, B. F. Navarro, and L. H. A. Monteiro, “Epidemiological models applied to viruses in computer networks,” *Journal of Computer Science*, vol. 1, no. 1, pp. 31–34, 2005.

[6] J. O. Kephart, T. Hogg, and B. A. Huberman, “Dynamics of computational ecosystems,” *Physical Review A*, vol. 40, no. 1, pp. 404–421, 1989.

[7] J. O. Kephart, S. R. White, and D. M. Chess, “Computers and epidemiology,” *IEEE Spectrum*, vol. 30, no. 5, pp. 20–26, 1993.

[8] J. O. Kephart, G. B. Sorkin, and M. Swimmer, "An immune system for cyberspace," in *Proceedings of the IEEE International Conference on Systems, Men, and Cybernetics (SMC '97)*, vol. 1, pp. 879–884, Orlando, Fla, USA, October 1997.

[9] L. Billings, W. M. Spears, and I. B. Schwartz, "A unified prediction of computer virus spread in connected networks," *Physics Letters A*, vol. 297, no. 3-4, pp. 261–266, 2002.

[10] M. E. J. Newman, S. Forrest, and J. Balthrop, "Email networks and the spread of computer viruses," *Physical Review E*, vol. 66, no. 3, Article ID 035101, 4 pages, 2002.

[11] B. K. Mishra and D. Saini, "Mathematical models on computer viruses," *Applied Mathematics and Computation*, vol. 187, no. 2, pp. 929–936, 2007.

[12] B. K. Mishra and N. Jha, "Fixed period of temporary immunity after run of anti-malicious software on computer nodes," *Applied Mathematics and Computation*, vol. 190, no. 2, pp. 1207–1212, 2007.

[13] M. Draief, A. Ganesh, and L. Massoulié, "Thresholds for virus spread on networks," *Annals of Applied Probability*, vol. 18, no. 2, pp. 359–378, 2008.

[14] L. Ljung, *System Identification*, Prentice-Hall, Upper Saddle River, NJ, USA, 1999.

[15] L. A. Aguirre, *Introdução à Identificação de Sistemas*, Editora UFMG, Belo Horizonte, MG, Brazil, 2004.

[16] P. Olofsson, *Probability, Statistics, and Stochastic Processes*, John Wiley & Sons, Hoboken, NJ, USA, 2005.