ARTICLE IN PRESS

Available online at www.sciencedirect.com
ScienceDirect

Applied Mathematics and Computation xxx (2006) xxx–xxx

APPLIED MATHEMATICS AND COMPUTATION

www.elsevier.com/locate/amc

Mathematical models on computer viruses

Bimal Kumar Mishra [a,*], Dinesh Saini [b]

[a] Birla Institute of Technology and Science, Mathematics Group, Pilani 333031, India
[b] Birla Institute of Technology and Science, Computer Science & Information System Group, Pilani 333031, India

Abstract

An attempt has been made to develop mathematical models on computer viruses infecting the system under different conditions. Mathematical model 1 discusses the situation to find the probability that at any time t how many software components are infected by virus, assuming the recovery rate and proportion of un-infected population receiving infection per unit time does not change with time. Mathematical model 2 is to estimate the proportion of software component population infected at any time and at any indefinite time under different cases. The third model is to find out the rate of change of proportion of total population with exactly j viruses ($1 \leqslant j < \infty$) and proportion of total population with zero virus, assuming that the total population is distributed into different groups based on the number of viruses present in a particular module. The fourth model is to find out what is the probability that at any time t, z number of software components are infected, assuming that initially (i.e. at $t = 0$), $\boldsymbol{a}$ number of components are infected and also there is a change from infected to uninfected or vice versa.

Keywords: Computer virus; Vaccination; Malicious agents; Software; Mathematical model; Super-infection; Virus breeding

1. Introduction

These are days of networked computers. Lot of efforts has been devoted to the development of virtual vaccines each time a new virus appears. Given the widespread use of sharing in current computer systems, the threat of a virus causing widespread integrity corruption is significant [3]. In a certain sense, the propagation of virtual viruses in a system of interacting computers could be compared with a disease transmitted by vectors when dealing with public health. Concerning diseases transmitted by vectors, one has to take into account that the parasites spend part of its lifetime inhabiting the vector, so that the infection switches back and forth between host and vector [9].

Predicting virus outbreaks is extremely difficult due to human nature of the attacks but more importantly, detecting outbreaks early with a low probability of false alarms seems quiet difficult [10]. By developing models it is possible to characterize essential properties of the attacks. In the present paper various mathematical models have been developed taking into account the different cases of probabilistic virus attacks.

* Corresponding author.
E-mail address: bimal@bits-pilani.ac.in (B.K. Mishra).

0096-3003/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.amc.2006.09.062

Please cite this article in press as: B.K. Mishra, D. Saini, Mathematical models on computer viruses, Appl. Math. Comput. (2006), doi:10.1016/j.amc.2006.09.062

Nomenclature

x_t	proportion of software component population infected at time t
r	recovery rate
h	proportion of unaffected population receiving infections per unit time
p_t	population of infected software component detected at time t using some diagnosis procedure of testing
n_i	number of new software component observed
a_i	number of software component found to be infected at time t_i where $(1 \leqslant i \leqslant I)$
$f_j(t)$	proportion of total population with exactly j viruses, $(1 \leqslant j < \infty)$
N	total population
X_t	number of infected computers at time t

2. Modeling the dynamics of transmission

Presence of computer viruses and quality factors of software development environment in the cyberspace, effect the functionality of the others components. Any software in cyberspace, which could be a running on server, workstation or a network router, exhibits its presence in various layers owing to the various applications running on it and its hardware configurations. Hardware layers could be taken to be network or removable media and such. Where as software layers would primarily be based on applications running on the host, like emailing connectivity and so on [8]. Software and Hardware layers are interdependent. Hence a host has to have at least one incarnation in the software layers and one in the hardware layers. Multiple incarnations in various layers contribute towards increasing connections between host's software and the number of peer's host software could communicate with. For example software A, could communicate with software B over h2 using s2 and s3 (Fig. 1). However B would not be able to talk directly to C since they do not share the same hardware layer. A hardware layer would be analogous to the medium of transfer of information where as the software layer can be associated with the format of information.

Computer viruses would need the transfer of the infected component to the various hosts' software. In Fig. 1, if s3 were associated with the file layer, which can be infected by a particular virus, the virus infecting B would be able to infect all three hosts software component. Then the infected file would need to be copied over the network (h2) onto A and then transferred over a floppy disk (h1) to C. Viruses are traditionally medium sensitive and hence a virus infecting B cannot infect C, since there is no connectivity between them (assumed to be h2). Operating system exploits could either be file based where in they need user intervention for transfer or self-aware component based. These would pose a serious threat as they have the combined power of viruses or malfunction or under performance of the resource and software component [7].

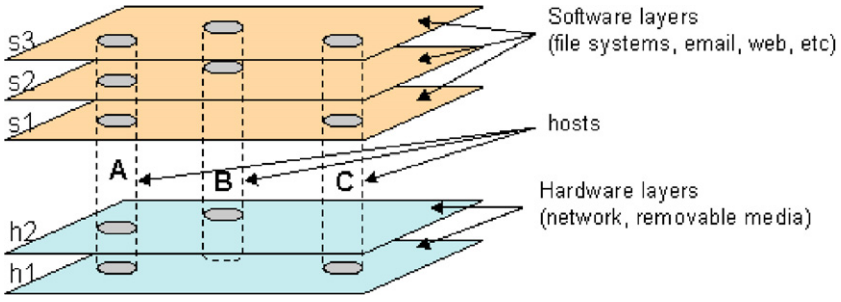

Fig. 1. Incarnations of hosts over various software and hardware layers.

Please cite this article in press as: B.K. Mishra, D. Saini, Mathematical models on computer viruses, Appl. Math. Comput. (2006), doi:10.1016/j.amc.2006.09.062

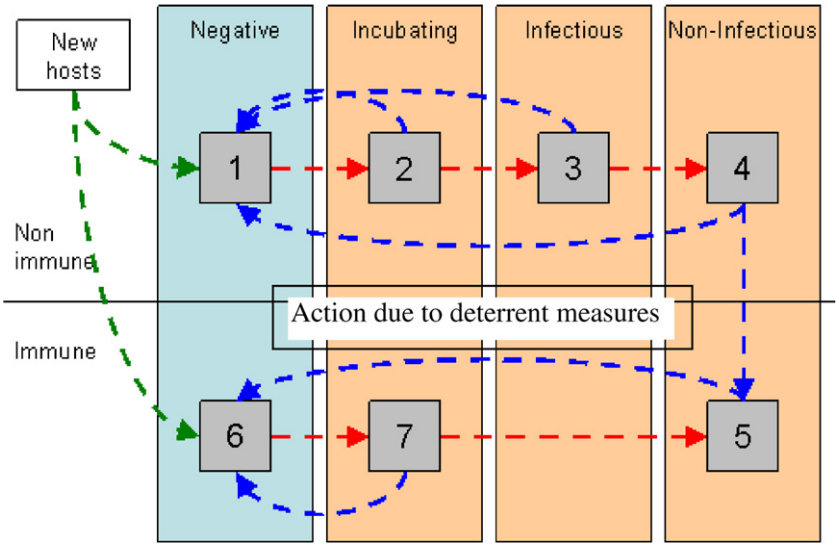

Fig. 2. Host software infection stages.

The spread of various malicious agents and their rate of infections could be effectively modeled based on their behaviors on individual layers, linked with the relationships between the layers and finally spanning across hosts software in a development environment to predict the state of a software development over time [4].

When new software and software component are introduced into cyberspace or in software development environment, there are two categories in which they could be placed (1, 7). Hosts by their very nature could be immune to a particular pathogen (virus) or non-immune to it (Fig. 2). For example some pathogens are operating system dependent. Hence a host introduced with the favorable operating system would not be immune and vice versa.[2] It is assumed for this model that all new software components introduced are in the negative state of infection for any infectious agent in the software development environment [5].

An immune or a non-immune host from the negative stages (1, 6) could then receive an agent and move into the incubating stage (2) where it is just containing the agent but the agent has not been triggered and hence the host is non-infectious. The agents in non-immune hosts could then be triggered either by user activities or by their own properties to infect the host in stage 3. Non-infectious stages (4, 5) could be attained either by immune and non-immune hosts' component where the agent could actually have been triggered but is unable to cause active infections. For example software virus infecting a particular version of an operating system could be contained within the host software by a few network traffic deterrent tools, thereby rendering them to be non-infectious [6].

Vaccination is taken care of by the connector between stages 4 and 5 where a non-immune host is immunized based on the infectious agents infecting it [1].

Note: The above characterization ignores reduction in number of host's software component due to deaths (host taken down) due to either infectious agents or due to any other reason, but does include births in the form of new software component as they join the negative infection stage.

3. Some basic terminologies

1. *Computer virus* is a program that can "infect" other programs by modifying them to include a possibly evolved version of it. With this infection property, a virus can spread to the transitive closure of information flow, corrupting the integrity of information as it spreads.
2. *Vaccine* is a software program designed to detect and stop the progress of computer viruses.
3. *Malicious agent* is a computer program that operates on behalf of a potential intruder to aid in attacking a system or network. Historically, an arsenal of such agents consisted of viruses, worms, and Trojanized

Please cite this article in press as: B.K. Mishra, D. Saini, Mathematical models on computer viruses, Appl. Math. Comput. (2006), doi:10.1016/j.amc.2006.09.062

programs. By combining key features of these agents, attackers are now able to create software that poses a serious threat even to organizations that fortify their network perimeter with firewalls.

4. Mathematical model 1

The main aim of this model is to find the probability that at any time t how many software components are infected by virus, assuming the recovery rate and proportion of un-infected population receiving infection per unit time does not change with time. We also assume that this model does not differentiate between infectious and non-infectious in the group of affected components, nor between susceptible and immune in the unaffected group.

4.1. Mathematical analysis

Let,

x_t proportion of software component population infected at time t
r recovery rate
h proportion of unaffected population receiving infections per unit time

$$\therefore x_{t+\Delta t} - x_t = [h(1 - x_t) - r(x_t)]\Delta t. \tag{1}$$

This in the limit $\Delta t \to 0$ gives

$$\frac{dx_t}{dt} = h - (r + h)x_t, \tag{2}$$

$$\text{At time } t = 0, x(0) = x_0, \tag{3}$$

$$r \geqslant 0, \quad h \geqslant 0. \tag{4}$$

Note:

1. It is assumed that r and h do not change with time.
2. The model does not differentiate between infectious and non-infectious in the group of affected computers, nor between susceptible and immune in the unaffected group.

From Eq. (3),

$$x_t = \frac{h}{r+h} - \left(\frac{h}{r+h} - x_0\right)e^{-(r+h)t}, \quad t \geqslant 0. \tag{5}$$

As

$$t \to \infty, \quad x_\infty = \frac{h}{r+h}. \tag{6}$$

Eq. (6) corresponds to the proportion of infected population (epidemic situation).

Special cases:

1. $r = 0$, $h > 0 \to x_\infty = 1$ [Whole software component population infected].
2. $r > 0$, $h = 0 \to x_\infty = 0$ [Infection disappears].
3. $r = 0$, $h = 0 \to x_\infty = x_0$ [No change].

For new systems: $x_0 = 0$, then from Eq. (5)

$$x_t = \frac{h}{r+h}[1 - e^{-(r+h)t}]. \tag{7}$$

Here x_t represents the proportion of new software component t population infected at time t.

Please cite this article in press as: B.K. Mishra, D. Saini, Mathematical models on computer viruses, Appl. Math. Comput. (2006), doi:10.1016/j.amc.2006.09.062

Let,

p_t: population of infected software component detected at time t using some diagnosis procedure of testing. Then $p_t = x_t$.

If infected computer software are detected with probability $k(0 < k \leqslant 1)$, then

$$p_t = kx_t = \frac{kh}{h+r}[1 - e^{-(r+h)t}]. \tag{8}$$

Let,

n_i number of new software component observed

a_i number of software component found to be infected at time t_i where $(1 \leqslant i \leqslant I)$

Then the estimates can be obtained by minimizing

$$\sum_{i=1}^{I} \left(p(t_i) - \frac{a_i}{n_i} \right)^2, \tag{9}$$

where $p(t_i)$ is given in Eq. (8).

5. Mathematical model 2

The main aim of this model is to estimate the proportion of software component population infected at any time and at any indefinite time t (i.e. $t \to \infty$) under different cases. The cases are as follow:

Case 1: Recovery rate less than or equal to proportion of unaffected population receiving infection per unit time.

Case 2: Recovery rate greater than or equal to proportion of unaffected population receiving infection per unit of time.

It is also assumed that the host carries multiple viruses and is in infected state as long as there is at least one virus present.

5.1. Mathematical analysis

Assumption: Host carries multiple viruses and is in infected state as long as there is at least one virus present

$$\frac{dx_t}{dt} = h - rx_t, \quad h \leqslant r, \tag{10a}$$

$$\frac{dx_t}{dt} = h(1 - x_t), \quad h \geqslant r. \tag{10b}$$

Note:

1. If $h < r$, then in time Δt all software component whether infected or not exhibit new infection at the rate $h\Delta t$, hence change in x_t over Δt is given by

$$(h - rx_t)\Delta t. \tag{11}$$

2. If $h > r$, then once infected the system would never recover, hence change in x_t over Δt is given by

$$h(1 - x_t). \tag{12}$$

From (Eqs (10a) and (10b))

$$x_t = \frac{h}{r}(1 - e^{-rt}), \quad h \leqslant r, \tag{13a}$$

$$x_t = (1 - e^{-ht}), \quad h \geqslant r. \tag{13b}$$

Please cite this article in press as: B.K. Mishra, D. Saini, Mathematical models on computer viruses, Appl. Math. Comput. (2006), doi:10.1016/j.amc.2006.09.062

As $t \to \infty$

$$x_\infty = \frac{h}{r}, \quad h \leqslant r, \tag{14a}$$

$$x_\infty = 1, \quad h \geqslant r. \tag{14b}$$

6. Mathematical model 3

The main aim of this model is to find out the rate of change of proportion of total population with exactly j viruses ($1 \leqslant j < \infty$) and proportion of total population with zero virus, assuming that the total population is distributed into different groups based on the number of viruses present in a particular module.

Total population segregated into different groups based on the number of viruses present in a particular module. Thus computer software belonging to group j carries j number of viruses within itself.

6.1. Mathematical analysis

If $j = 0$ then the corresponding group is a collection of computer software unaffected by viruses.

Let,

$f_j(t)$ proportion of total population with exactly j viruses, $(1 \leqslant j < \infty)$

x_t proportion of population infected at time t

$$x_t = \sum_{j=1}^{\infty} f_j(t) \tag{15}$$

and the unaffected proportion is given by

$$1 - x_t = 1 - \sum_{j=1}^{\infty} f_j(t). \tag{16}$$

In an interval Δt, a change in $f_j(t)$ can occur when

1. One or more computer software in group $(j-1)$ is attacked by an infectious viruses.
2. Two or more computer software in-group $(j+1)$ recovers partially so that one reduces the number of viruses in the infected host.

Let,

r recovery rate for computer software

h rate of new infections being introduced

Then the increase in $f_j(t)$, in time Δt is given by

$$[hf_{j-1}(t) + r(j+1)f_{j+1}(t)]\Delta t \tag{17}$$

and the decrease in $f_j(t)$ is given by

$$(h + rj)f_j(t). \tag{18}$$

By limiting arguments,

$$\frac{\mathrm{d}f_0(t)}{\mathrm{d}t} = -hf_0(t) + rf_1(t), \tag{19a}$$

$$\frac{\mathrm{d}f_j(t)}{\mathrm{d}t} = hf_{j-1}(t) - (h + rj)f_j(t) + r(j+1)f_{j+1}(t), \quad j \geqslant 1, \tag{19b}$$

$$\text{At } t = 0,\ f_j(0) = f_j^0 (0 \leqslant j \leqslant \infty) \text{ with } f_j^0 \geqslant 0 \quad \text{and} \quad \sum_{j=0}^{\infty} f_j^0 = 1. \tag{20}$$

Please cite this article in press as: B.K. Mishra, D. Saini, Mathematical models on computer viruses, Appl. Math. Comput. (2006), doi:10.1016/j.amc.2006.09.062

7. Mathematical model 4

The main aim of this model is to find out what is the probability that at any time t, z number of software components are infected, assuming that initially (i.e. at $t = 0$), a number of component are infected and also there is a change from infected to uninfected or vice versa.

Attempt has also been made to find the mean value for the fraction of components, infected at time t.

Assumption: A change from infected to un-infected or vice versa by an uncertain chance mechanism.

7.1. Mathematical analysis

Let,

N total population

X_t number of infected computers at time t

$\therefore N - X_t$ number of susceptible at time t

P (one new infection occurring in time t to$t + \Delta t) = h(N - X_t)\Delta t$

P (one new infection occurring in time t to $t + \Delta t) = rX_t\Delta t$

Define

$$p_j = P(X_t = j), \quad 0 \leqslant j \leqslant N \tag{21}$$

and

$$P(z,t) = \sum_{j=0}^{N} p_j(t) Z^j. \tag{22}$$

Now,

$$\frac{\mathrm{d}p_0(t)}{\mathrm{d}t} = -hNp_0(t) + rp_1(t),$$

$$\frac{\mathrm{d}p_j(t)}{\mathrm{d}t} = -\{h(N-j) - rj\}p_j(t) + h(N-j+1)p_{j-1}(t) + r(j+1)p_{j+1}(t). \tag{23}$$

For $1 \leqslant j \leqslant N - 1$

$$\frac{\mathrm{d}p_N(t)}{\mathrm{d}t} = -rNP_N(t) + hp_{N-1}(t)$$

and

$$\frac{\partial p}{\partial t} = (1-z)(hz + r)\frac{\partial p}{\partial z} + rh(z-1)P, \tag{24}$$

$$\text{Let at } t = 0, \text{ the number of infected computers is given by } x(0) = a. \tag{25}$$

Then

$$p_j(0) = 1; \quad \text{If } j = a, \tag{26a}$$

$$p_j(0) = 0; \quad \text{Otherwise} \tag{26b}$$

and

$$p(z,0) = z^a.$$

Using the initial conditions of Eq. (25) and Eqs. (26a) and (26b), Eq. (24) can be solved to yield:

$$p(z,t) = (r+h)^{-N}\{(r+hz) + r(z-1)\mathrm{e}^{-(r+h)t}\}^a\{(r+hr) - h(z-1)\mathrm{e}^{-(r+h)t}\}^{N-a}. \tag{27}$$

If $m(t)$ be the mean value for the fraction of computers infected at time t, then

$$m(t) = E[X(t)]|N, \tag{28}$$

Please cite this article in press as: B.K. Mishra, D. Saini, Mathematical models on computer viruses, Appl. Math. Comput. (2006), doi:10.1016/j.amc.2006.09.062

where $E[X(t)] = \left(\frac{\partial p}{\partial z}\right)_{z=1}$

$$E[X(t)] = \left(\frac{Nh}{r+h}\right)(1 - e^{-(r+h)t}). \tag{29}$$

Substituting Eq. (29) in Eq. (28), we get,

$$m(t) = \left(\frac{h}{r+h}\right)(1 - e^{-(r+h)t}). \tag{30}$$

8. Conclusion

The mathematical models developed above will help in finding the probability of a system being infected by any computer virus or a group of computer viruses at any time specifically dealing with the speed of breeding of the viruses. Models can be useful in designing defenses against non-harmful and malignant computer virus attacks which are of considerable importance in the present day context. These models will help in carrying out the sensitivity analysis and can be verified by simulation.

9. Limitations of models and research challenges

Our model does not differentiate between susceptible and immune in the unaffected group. Models developed above do not talk about speed and transient of virus spread. Better sensitivity analysis can be developed which our models do not address.

References

[1] E. Makinen, Comment on a framework for modelling Trojans and computer virus infection, Computer Journal 44 (2001) 321–323.
[2] F. Cohen, Computer viruses – theory and experiments, in: DOD/NBS 7th Conference on Computer Security, originally appearing in IFIP-sec 84, also appearing in Computers and Security, vol. 6, 1987, pp. 22–35.
[3] Harlod Timbley, Stuart Anderson, Paul Cains, A framework for modelling Trojans and computer virus infection, The Computer Journal 41 (7) (1998) 445–458.
[4] J. Balthrop, S. Forrest, M.E.J. Newman, M.M. Williamson, Technological networks and the spread of computer viruses, Science 304 (5670) (2004) 527–529.
[5] J.L. Aron, M.O. Leary, R.A. Gove, S. Azadegan, M.C. Schneider, The benefits of a notification process in addressing the worsening computer virus problem: results of a survey and simulation model, Computer and Security V21 (2002) 142–163.
[6] J.O. Kephart, S.R. White, Measuring and modelling computer virus prevalence, in: Proceeding of the 1993 IEEE Computer Society Symposium on Research in Security and Privacy, Oakland, California, 1993, May 24–25, pp. 2–14.
[7] L. Billings, W.M. Spears, I.B. Schwartz, A unified prediction of computer virus spread in connected networks, Physics Letters a297 (2002) 261–266.
[8] M. Newman, S. Forrest, J. Balthrop, Email networks and the spread of computer viruses, Physical Review E 66 (2002) 035101.
[9] N.J.T. Baily, The Mathematical Theory of Infectious Diseases and its Application, Griffin, London, 1975.
[10] V. Capasoo, Mathematical Structure of Epidemic Systems, Springer Verlag, 1993.

Please cite this article in press as: B.K. Mishra, D. Saini, Mathematical models on computer viruses, Appl. Math. Comput. (2006), doi:10.1016/j.amc.2006.09.062