コンピュータウィルスの拡散の観測および理論モデルとの比較

海和建大[†]　豊泉洋[*]
[†] 会津大学　[*] 早稲田大学
E-mail: hyro_tk@hotmail.com, toyoizumi@waseda.jp

本論文では、コンピュータワームがメールネットワーク上で拡散する様子をモデル化し、評価する方法を提案し、大量に拡散メール型ワームに対する効果的な防御手法を提案する。メールネットワーク上では、メールをやり取りしたノード間にはリンクが存在すると考える。メールネットワークについては、既にいくつかの論文でスケールフリーネットワークでモデル化することができることが指摘されている。メール型のワームは、このようなスケールフリーネットワーク上で拡散するとモデル化することができる。ワームに感染したメールの From フィールドは偽装されていることが多い。したがって、メールのログデータを使って、メールネットワークの構造やメール型ワームの拡散の様子を調べることはできない。本論文では、ワームの到着間隔データを使い、メールネットワークの構造や拡散の様子を解析する。また、メールネットワーク上のハブでのワームの活動を制限することにより、効果的にワームの拡散を防止できることを示す。

Observation of Spread Dynamics of Malicious Software and its Comparison to Theoretical Model

TATEHIRO KAIWA[†]　HIROSHI TOYOIZUMI[*]
[†]University of Aizu　[*]Waseda University
E-mail: hyro_tk@hotmail.com,toyoizumi@waseda.jp

In this paper, we present a way to build a model how computer worms propagate in mail network, and then propose an effective defense method against a mass-mailing worm. In mail network, two nodes are connected when emails are exchanged between them. As some papers pointed out already, mail network can be successfully regarded as a Scale-free Network. The mail worm spread can be regarded as the propagation in such Scale-free network. The From-field of an e-mail messages a worm sends often varies and/or is spoofed. The log data of emails reveal neither the mail network structure nor the worm dynamics. We use arrival intervals of worms to analyze the mail network structure and the worm dynamics. Also, we show that we can effectively stops the worm outbreak by regulating worm activity on hubs on the mail network.

1 Introduction

The most common way to spread malicious software programs is commonly referred to as computer worms like Blaster [1] and Netsky [2]. There are some methods that a worm uses for its propagation. One of the popular methods is to send e-mail messages with its own copy attached. The e-mail has some trap which, when activated, executes the copy of malicious software. For example, Swen [3] sends an e-mail message whose contents claim to be patches for Microsoft Internet Explore, or delivery failure notices from qmail. Although a lot of new viruses appear in few days, most of them are not nuisance because there are effective filters quickly available. However it is expected that in future the worms will be more virulent and, thus, result in significantly greater damage.

Currently, there are a lot of anti-virus softwares that have various effective countermeasures against various computer viruses and worms. The popular method is to use signatures. The system achieves almost 100% detection rate if the system has already the signature for the target. However, the system has some weak points. The system cannot detect a new virus or worm. The system also needs a large memory for signatures. Further, The system needs the time to detect the viruses and worms. Most importantly, not all computers are properly equipped with those anti-virus softwares.

The purpose of this research is to analyze and estimate how a computer virus or worm propagates in a mail network and to analyze and estimate an effective method against a mass-mailing worm. To estimate that, we focus attention on the arrival intervals of e-mail message with a specific computer worm attached. Because the From-field of an e-mail message a worm sends often varies and/or is spoofed, it is only arrival intervals that we can obtain from received e-mails. As the data of arrival time of many e-mail messages with each computer worms attached are observed on the mail server in University of Aizu.

Also, we have to consider a structure of the e-mail network in the Internet to research a propagation of a worm. In mail network, two nodes are connected when emails are exchanged between them. As some papers [4,

5] pointed out already, mail network can be successfully regarded as a Scale-free Network. Thus, we regard the structure of mail network as Scale-free.

2 Scale-free Network

Scale-free Networks are networks in which degree of a node is power-law distribution [4, 5]. Recently, there found that various networks such as the Internet, propagation of an epidemic and a genome have Scale-free Network structure.

Network grows daily. When a new node belongs to a network, the node does not connect to another randomly selected one. In real networks linking is never random. A subtle law of preferential attachment governs evolution of a network. Scale-free Network is the network such that.

In Scale-free Network, some nodes called " hub " nodes exhibit extremely high connectivity and almost all other nodes exhibit low connectivity. As there are a small number of hub nodes, a Scale-free Network has generally two characteristics. First one is that a Scale-free Network has tolerance to random attack to most nodes but it has weakness to attack to the hub nodes. Second one is that a Scale-free Network has also a characteristic of Small-world [6].

3 Activity of Worms

Most mass-mailing worms work as follows. When a mass-mailing worm is executed, it sends some e-mail messages with its own copy. After someone received one of those infected email, it may take some time for the recipient to read the e-mail message. He may execute mailer or something to read some e-mail messages, in which the e-mail message with attachment file of the worm is contained. At this time, the mass-mailing worm program runs if he executes the attachment file. Then, the worm will use the recipient's address book, inbox and/or other e-mail addresses saved in recipient's HDD to spread its own copies to other e-mail addresses. In addition to that, there are some worms which will infect other malicious programs, which are Trojan horse, back door, and so on.

4 Observation of Wild Worm

Because the From-field of an e-mail message a worm sends often varies and/or is spoofed, it is difficult to identify the PC a mass-mailing worm has infected and a correct data we can obtain from received e-mails is arrival intervals. To obtain arrival intervals of e-mail messages with a worm attached, we use a data observed on

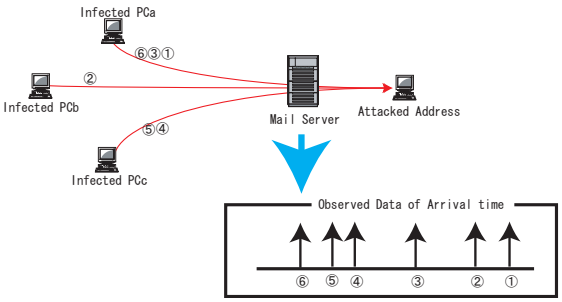

Figure 1: Observation Method. Each worm infecting PC sends an e-mail with own copy attached to other addresses independently. We can observe only the e-mails reaching our mail server.

a mail server in University of Aizu. The data contains what kind of worm an e-mail message has and when the e-mail message arrives. When we can observe arrival of one or more e-mail messages with a worm attached, there are one or more infected PCs having an e-mail address of a student of University of Aizu. (See Figure 1)

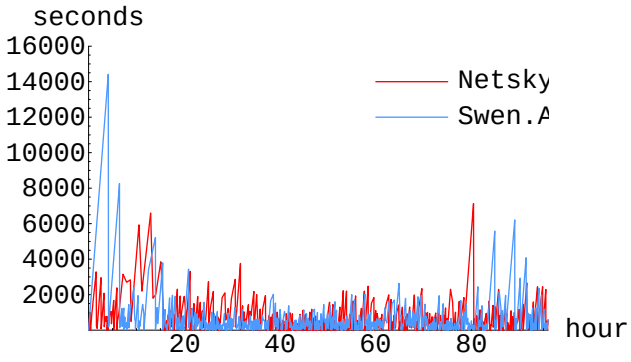

Figure 2: Arrival intervals of e-mail attached some specific viruses observed on the mail server in University of Aizu. Time 0 means the time on which the first e-mail attached each virus arrived at University of Aizu.

Virus Name	discovered	First observed at u-aizu
Netsky.P	Mar. 21, 2004	Mar. 23, 2004 - 01:37:20
Swen.A	Sep. 18, 2003	Sep. 18, 2003 - 23:07:31

Table 1: Information of observed viruses.

Figure 2 shows arrival intervals of e-mail messages with each worm attached. We observe them on the mail server in University of Aizu. Table 4 shows the date on which each worm is discovered by Symantec corp. and

first arrival time observed on the mail server in University of Aizu.

5 Network Topology in Simulation

In order to simulate the dynamics underlying mass-mailing worm's propagation, we need to create an mail network. There are some papers[7, 6, 8] discussed an e-mail network or social network, so we assume that an e-mail network are a Scale-Free network. We use a BRITE tool [9] to create a Scale-free Network which we use in this simulation. Figure 3 is GUI of BRITE.

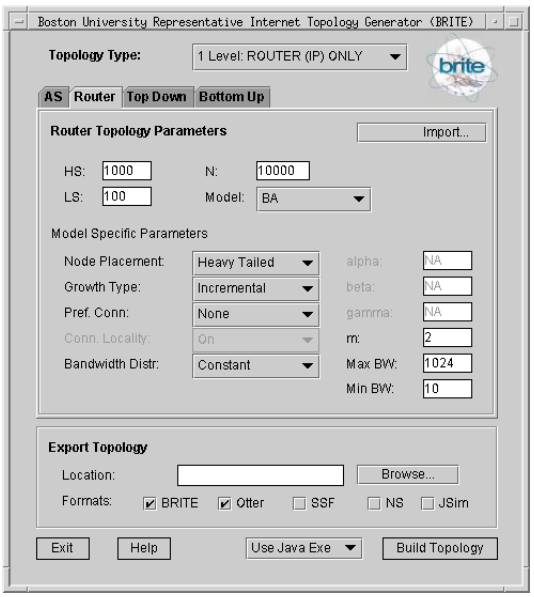

Figure 3: This is GUI of BRITE. We can change some parameters when we create something. For example, Node Placement, Growth Type, the number of nodes and so on.

Now, we suppose that an e-mail address network obeys a Scale-free Network, and we employ Barabasi-Albert (BA) model[4] to create a Scale-free Network. Creating a Scale-free Network with BRITE, we obtain that a network obeys a power-law with exponent -3. It is too hard to assess whether that exponent value is similar to a network a virus or worm use actually. However, being the number of link distributions of computer and social networks lie in the range of -2.0 to -3.4 [10], the exponent value is -3 we assume. This Figure 4 is an example of a Scale-free Network with BA model. The Figures 5 show a relationship between a degree of links of each node and the number of nodes. We can see that the relationship obeys power-law because this double-logarithmic plot is similar to liner. The number of nodes exhibits a power-law with exponent -3.

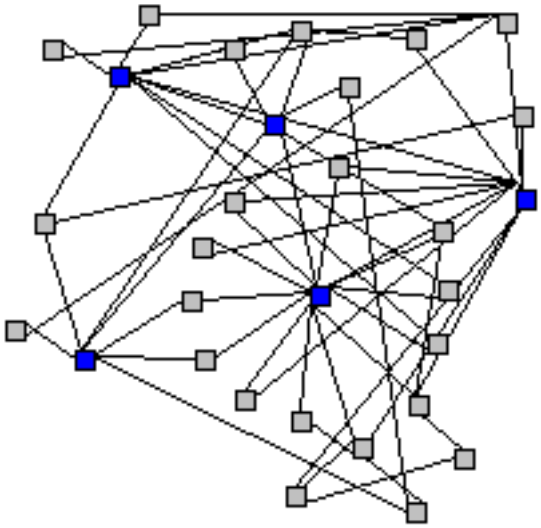

Figure 4: A example of a Scale-free Network. The number of nodes is 30 and the number of links per newly added node is 2. There are a few hub nodes and the almost other nodes only have a few link. Some colored nodes are hub nodes.

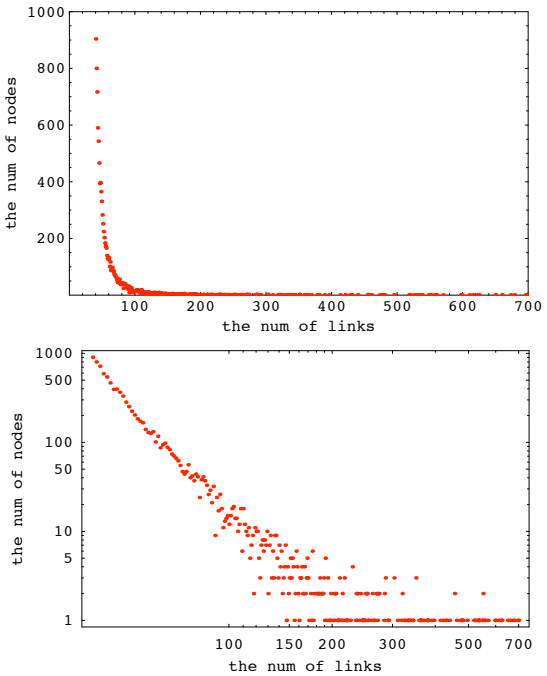

Figure 5: These figures show a relationship between a distribution of the number of links of each node and the number of nodes. The number of nodes is 10000 and the number of links per newly added node is 20.

6 Simulation Model of Worm Spread Dynamics

As discussed in Section 4, we assume that we can observe the arrival time. There is a community and each node has an e-mail address (see Figure 6). If two nodes shares address each other, the nodes are considered to have a link between them. Then, there is a student belonging to the community in University of Aizu. As the student belongs to the community with the e-mail address of the university, all e-mail messages excluding e-mail messages with a worm attached to him is observed the mail server in the university.

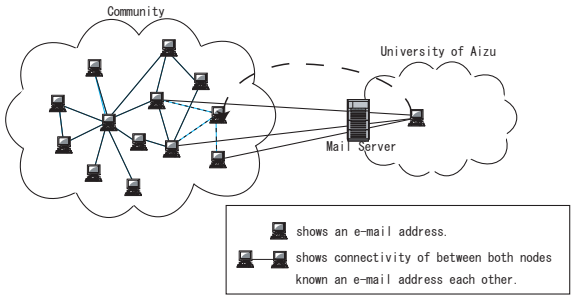

Figure 6: Image of a community sending e-mail messages with a worm attached.

We set the number of nodes in the community is 10000. We assume that the connectivity between the nodes in the community obeys power-law with BA model.

A node has many addresses also except ones of the same community. We assume that the number of the e-mail addresses a node has of other community is proportional to the number of links the node has and there is no e-mails come from other community. Then, each node has intervals at which mailer program is executed. We introducing three statuses at each node, that we call susceptible, infected and immune. Susceptible means that the node has a probability with which the node is potentially infected by a virus or worm. Susceptible contains when a user has not yet checked for new e-mails. Infected means that the user has opened the attachment and the virus has successfully infected the PC sending copies of itself to all of the neighboring nodes. In the simulation, we assume that a node, once infected, remains in this status forever. Immune means that the virus cannot infect the node. Immune contains when the user protect own PC with using some ways such as an anti-virus software, a patch, or no open the attachment. Once immunized, the node remains in this status forever.

A node a worm infects at once sends some e-mail messages with its own copy attached to other e-mail addresses per unit of time. We assume that each interval of e-mail messages which a worm sends follow an independent exponential distribution with mean $1/\lambda$. We assume that a worm select a link in an infected node randomly. After receiving the infected mail, the nodes will be infected with a click probability c. This parameter c is a probability with which a recipient receiving an e-mail messages with a worm attached executes the worm program. The probability c has an influence on a speed of propagation. In this simulation, we assume that all of nodes have a same c. We suppose that a susceptible node infected at once becomes an infected node and the node holds the own state ever. Then the node attacks the other susceptible nodes that connect the node such as the other infected nodes.

7 Fitting Model to Observation

To fit our worm spread model to observed arrival intervals of real worms, we pay attention to the part where arrival intervals stop decreasing. It means that most of neighbors of an observer are infected by that time. In our simulation correspondingly, most of neighbors of an observer have to be infected by the time. And we also have to consider decrease of arrival intervals from first arrival of real worm. Because a parameter λ is an estimated value empirically, we cannot change λ. Thus, we need to change c and the number of the e-mail addresses a node has of other community. Let n is the number of links per newly added node. Because all nodes have many e-mail addresses, it is wrong that n is lower value. Though we estimate the number of the e-mail addresses more, we cannot obtain fitting model. On the other hands, as there is a restriction least upper bound of the number of nodes, we cannot estimate that n is higher value because a higher value n makes our network topology a mesh topology. From results of many simulations, we decide that 's parameters. Assuming that the number of the e-mail each node has is ten times the number of links connected to the community, that n is 20 and that c is 0.01, we obtain Figure 7.

Figure 7 shows the result of simulation compared with observed data of Netsky.P. We can see that by adjusting our model we can make both the simulated dynamics and the real one to be similar. About 10 hours, arrival intervals of observed data become larger than simulation. This is because some infected neighbors stop to attack in the observation.

8 Local Network Structure Inference

We can estimate a network structure around a observer from observed data.

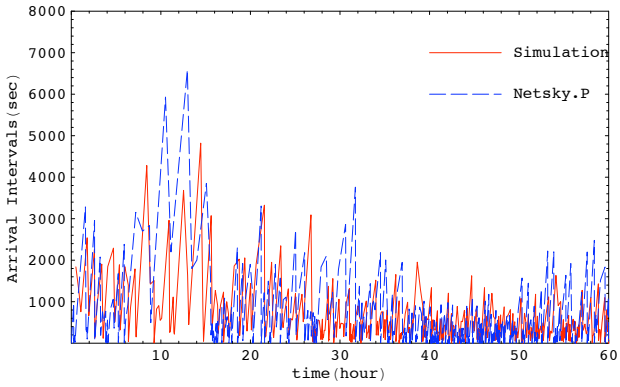

Figure 7: Comparing result of simulation at a node with observed data of Netsky.P. Time 0 means the time on which a first attack arrives at each observing point.

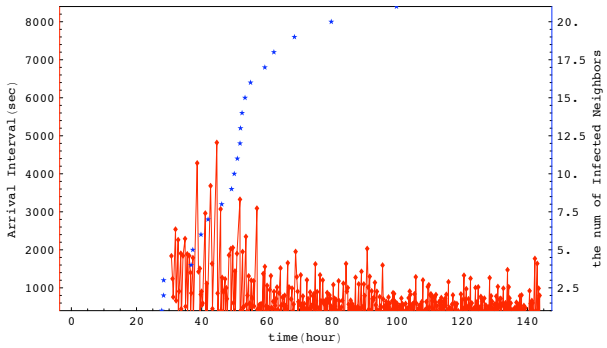

Figure 8: Figure of arrival intervals. An infected node send about five e-mail messages per a second, λ is 5. The number of neighbors of this observation node is 20. The mean of links the neighbors have is 1156.19. The figure set time 0 as the start of simulation.

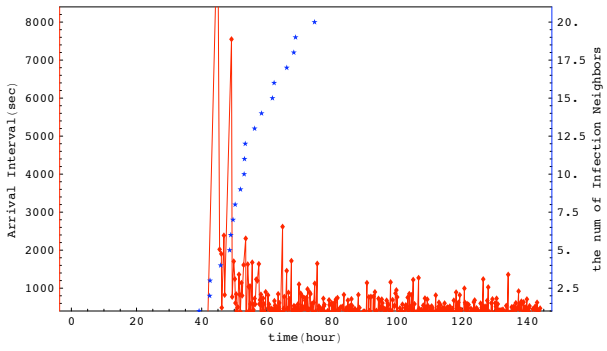

Figure 9: Figure of arrival intervals. An infected node send five e-mail messages per a second, λ is 5. The number of neighbors of this observation node is 20. The mean of links the neighbors have is 619.5.

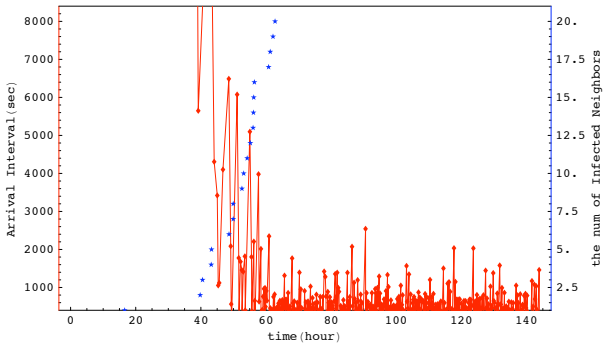

Figure 10: Figure of arrival intervals. An infected node send five e-mail messages per a second, λ is 5. The number of neighbors of this observation node is 20. The mean of links the neighbors have is 645.5.

Figure 8, Figure 9 and Figure 10 are results of Simulation. Each node has some neighbors and the number of neighbors is similar. In Figure 8, we can see that there are some infected neighbors in early time than others. We can estimate that is occurred by some infected hub nodes connected to the observer. On the other hand, we can see that the arrival intervals decrease slowly than others. The reason why that is that the number of infected neighbors increase more slowly than others. If there are some infected neighbors when first attack arrives, decrease of arrival intervals is more slowly. Comparing Figure 9 with Figure 10, we can see that there is a difference of outbreak time in same community.

Figure 9 shows arrival intervals of a node of which the number of neighbors are infected at short times is large. On the other hand, a case of a node of which the number of neighbors are infected increase slowly is Figure 10.

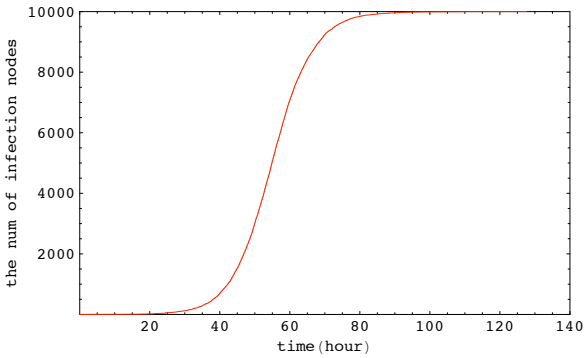

Figure 11: Figure of the number of infected nodes. The click probability c is 0.01. After 40 hours this simulation started, outbreak is occurred.

From Figure 11, we can see an outbreak time approximately. We can see that the infected neighbors of each

observation node increase after this time from Figure 8, 9 and 10. Thus, we can see that a virus may also occur an outbreak in the Internet when we observe a decrease of arrival intervals. Therefore, we have to do a countermeasure against a virus or worm before observed arrival intervals start to decrease.

Figure 12 shows a relationship between the number of all infected nodes and time. We can see that the graph is similar to logistic curve[11] and that a part when a virus occur outbreak. Moreover, we can see that a change of the parameter c makes the outbreak time earlier or later.

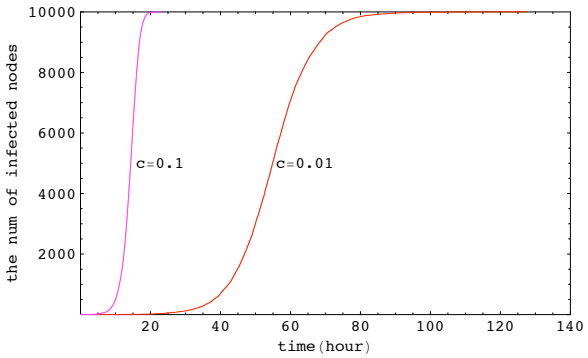

Figure 12: The graph shows the number of infected nodes with time. The c is 0.1 and 0.01. All nodes except an initial infected node is liable to infect and the number of total nodes is 10000.

9 Mathematical Model of Outbreak

Now we will consider a mathematical model [12] of worm dynamics on scale-free network. We will evaluate the total number of nodes a worm infects. Two nodes are considered to have a link when they share their email addresses. Then, we suppose that a computer virus or a worm propagates in the network the links of e-mail addresses generate. We assume that the number of links of each node is independently and identically distributed. And let the number of one of any nodes' links is K, and let the number of other links of a node connects a link chosen arbitrarily is K_e. Because the probability that those nodes with many links are chosen is high, the probability density of K_e is as follows. (using the probability density of K):

$$P\{K_e = k\} = \frac{kP\{K=k\}}{E[K]}. \tag{1}$$

Especially, the expectation $E[K_e]$ is as follows:

$$E[K_e] = \sum_{k=0}^{\infty} \frac{kP\{K=k\}}{E[K]}k \tag{2}$$

$$= \frac{E[K^2]}{E[K]}. \tag{3}$$

In a long-tailed distribution such as a power-law distribution, a relation between $E[K_e]$ and $E[K]$ is $E[K_e] \geq E[K]$. As a tail become more longer, $E[K_e]$ also become more larger. In case that a worm infects a node, we propose that other nodes will be infected with probability p when the node is connected to the first infected node. And then, we evaluate the total number of infected nodes when the propagation occurs a set of random nodes, and let the size of this initial set be S. This problem is called the percolation problem, and it is a well-known problem in physics. Now, we propose that we random choice a link in the network and assume that the propagation occurs from the link. Let the total outbreak size in that case is S_e. We assume that the network size is enough large and there is no infection loop. Then, we can show a relationship between S and S_e as follows:

$$S = 1 + \sum_{m=1}^{M} S_{e,m}, \tag{4}$$

where M is the number of infection links which the first chosen node has. S and S_e are independently and identically distributed. On the other hand, when the number of infection links of a node which is connected to the first chosen link is M_e, we obtain a recursive equation as follows:

$$S_e = 1 + \sum_{m=1}^{M_e-1} S_{e,m}. \tag{5}$$

Now, both M and M_e are integers which is greater than zero. The expectation of $S_{e,m}$, $E[S_e]$ is less than infinity because the number of nodes is not infinity. Then the expectation of M and M_e, $E[M]$ and $E[M_e]$ are both less than infinity because the number of links is also not infinity. As $S_{e,1}, S_{e,2}, \cdots, S_{e,M}$ be a sequence of M iid random variables distributed as random variable S_e, we obtain:

$$E(\sum_{i=1}^{M} S_{e,m}) = E(M)E(S_e). \tag{6}$$

The equation (6) is Wald's equation [13]. Evaluating the expectation of the equation (4) and (5) with using the equation (6), we obtain:

$$E[S] = 1 + E[M]E[S_e], \tag{7}$$

$$E[S_e] = 1 + (E[M_e] - 1)E[S_e]. \tag{8}$$

It follows that:

$$E[S] = 1 + \frac{E[M]}{2 - E[M_e]} \quad (9)$$

$$= 1 + \frac{pE[K]}{2 - pE[K_e]}. \quad (10)$$

If we estimate

$$\frac{pE[K]}{1 - pE[K_e]}, \quad (11)$$

we can obtain the outbreak size E[S]. On the other hand, from (10), we can see that the outbreak size $E[S]$ diverge when $E[M_e] = pE[K_e] = \frac{pE[K^2]}{E[K]} = 2$. Thus we need to control $E[M_e]$ to prevent the outbreak. Making $E[M_e]$ decrease, we need to decrease the variance of M or we increase the mean of M. Since it is impossible to increase $E[M]$, we have to decrease the variance of M when we consider an effective countermeasure.

10 Hub Defense Strategy

We consider that an effective method with the node degree distribution against a virus propagation. A Scale-free topology is known more commonly as a topology have a susceptibility to attacked to some hub nodes. In addition to that, we have learned an approach with mathematic in chapter 2. In this chapter, we simulate propagation of a worm in follow condition. Now, we consider a community having 10000 e-mail addresses. Let h is the number of immune hub nodes. We assume that an e-mail address corresponds with a node in the community and that each e-mail address has some links connecting other addresses in the community and that the degree of the links obeys power-law with BA model. Then, the number of links per newly added node is 20 and each node has 9 times links its own has. We assume that each infected node sends 5 e-mails per a second to own neighbors. Figure /reffig:defense shows that a difference of speed of propagation each case and that an effect of when we set the number of immune hub nodes changing.

In Figure 13, the more we increase the number of immune hub nodes, the more the method becomes more effective.

11 Conclusion

We derive a way to build a model of worm spread dynamics using Scale-free Network. From result of simulation, we found that arrival intervals also decrease slowly if the number of infected neighbors of an observer is increase slowly, and the rate of this decrease is depend on the link structure of the neighbors.

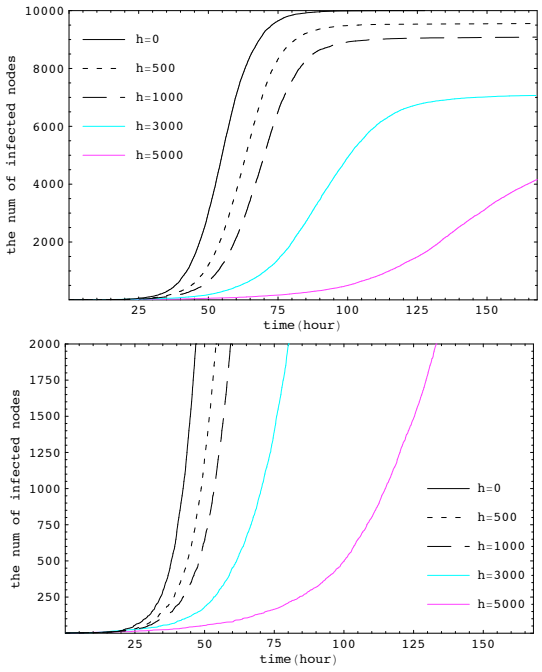

Figure 13: Effect of changing the number of immune hub nodes with parameter c=0.01. Each h is 0, 500, 1000, 3000 and 5000.

On the other hand, we confirm an effect of propagation of a worm by the number of immune hub nodes. Especially in early time, we can obtain enough advantage even though the number of immune hub nodes is small. However unfortunately, we cannot see a great effect as time goes by. However, the e-mail network in the Internet may be different from this paper. Therefore, we need to research how many links the e-mail address in the Internet has or we need to consider a method estimating that.

参考文献

[1] Symantec. W32.blaster.worm, http://securityresponse.symantec.com/avcenter/venc/data/w 32.blaster.worm.html. Technical report.

[2] Symantec. W32.netsky.p@mm, http://securityresponse.symantec.com/avcenter/venc/data/w32.netsky.p@mm.html.

[3] Symantec. W32.swen.a@mm, http://securityresponse.symantec.com/avcenter/venc/data/w32. swen.a@mm.html.

[4] Albert-Lazzio Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, Vol. 286, No. 5439, pp. 509–512, October 1991.

[5] A.-L. Barabasi. *Linked*. Plume Books, 2003.

[6] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, Vol. 393, No. 6684, pp. 440–442, June 1998.

[7] Holger Ebel, Lutz-Ingo Mielsch, and Stefan Bornholdt. Scale-free topology of e-mail networks, Feb 2002.

[8] Gueorgi Kossinets and Duncan J. Watts. Empirical analysis of an evolving social network. *Science*, Vol. 311, No. 5757, pp. 88–90, January 2006.

[9] Alberto Medina, Anukool Lakhina, Ibrahim Matta, and John Byers. BRITE: Universal topology generation from a user's perspective. Technical Report 2001-003, January 2001.

[10] R.Albert and A.-L. Barabasi. Statistical mechanics of complex networks. *Rev. Mod. Phys. 74*, pp. 47–97, 2002.

[11] Hiroshi Toyoizumi and Atsushi Kara. Predators: Good will mobile codes combat against computer viruses. *In Proceedings of the New Security Paradigms Workshop*, pp. 13–21, September 2002.

[12] Hiroshi Toyoizumi and Tatehiro Kaiwa. Observation and modeling method of dynamics of computer virus spread. *QoS Workshop "Reality of QoS in the Internet and Latest Measuring or Estimating Technologies."*, November 2005.

[13] Sheldom M. Ross. *Applied Probability Model With Optimization Applications*. Dover Pubns, 1992.