Scanning Computer Viruses with Reduced Virus Definition File

Daisuke Anzai 1090009 Supervised by Hiroshi Toyoizumi

Abstract

In spite that virus definition file has published several ten thousand kinds of virus information, almost overflowing viruses on the network are ten kinds in actually. We present the virus definition file can reduce and prove this is low risk, not reduce the performance of server even if number of received mails are many.

1 Introduction

One of the most serious problem on the Internet is the problem of computer virus. Currently, everyone can connect to network, in other words, we are always exposed to the threat of virus damage. The route of infection has mail, Web, a network, etc., and users are going to protect a machine from these attacks by using anti virus software, filtering, and firewall.

The report of damage by computer virus merely 14 kinds in 1990 acording to IPA [4]. However, exceed 1000 in 1994, and explose more 10000 in 2000. These things were persistently submit to IPA, the damege guess more several ten times than them in actually.

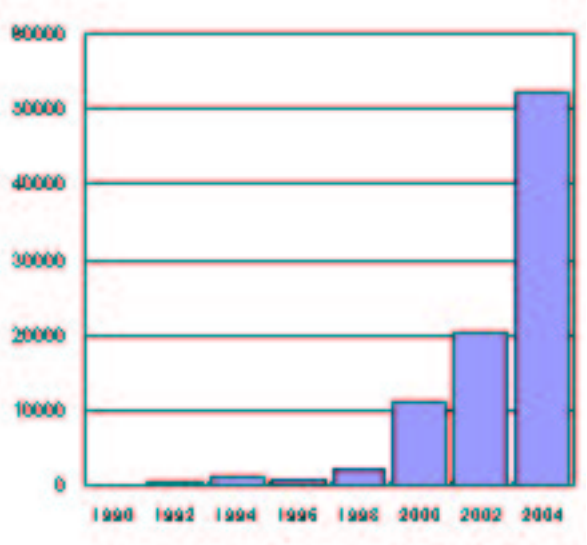

Figure 1: submmited Viruses to IPA

Computer viruses have been studied extensively over last several years. Our laboratory also studied some effective measure against computer viruses. Reference [1] describes analysis mechanism against computer viruses which camouflage sending source address by observation inter-arrival rate of mails with a virus. Also reference [2] shows new defense way which captured 3 types worm's algorithm and estimate the amount of mails with a virus based on exponential distribution, then success to stop the viruses at server.

The malicious program infected using mail is called mass mailing type worm, tends to transmit the mail which appended its duplicate to an unspecified user, make the user who received open an attached file, and it is going to infect it. To find or exterminate viruses, users generally use anti virus software. The anti virus software is made virus information an independent file, can correspond new viruses if only the file is updated. This virus information file is called virus definition file. Though the use of this file, mail server scans the mail whether it contains virus or not. In 1986, the first virus was discovered. Only about 20 years past but several ten thousands of kind of them are confirmed now, and they will continue to increase. Reference [3] tells more than 68000 kinds of viruses information have already appeared in the virus definition file. It makes virus definition file long. The scanning time is also same.

Some old viruses are almost dead, and if they are alive, some security holes (that is, what becomes a problem on the measure against security among the faults which exist in a program) and vulnerability (that is, the state where an aggressor can execute a command as another user in a single stand or a set of computing systems, the state where an aggressor can access the restricted date, the state where an aggressor can become an another user or an another group, or the state where an aggressor can devise a service refusal attack) are canceled by occurring windows update or version of Internet Explorer is innovated. Therefore, the degree of infected danger must be extremely low and may not become damage.

Even now, the updating of virus definition file only add new viruses. In near future, published virus in the definition file will exceed 100000 kinds. It is likely not to reach smoothly when mails arrival rate become higher. The purpose of my research is to assume server's system is M/D/1 queuing model, and use the virus definition file which carried reducing virus information, compares a usual system and this system, and guess the validity of

this system.

2 Reducing Virus Definition File

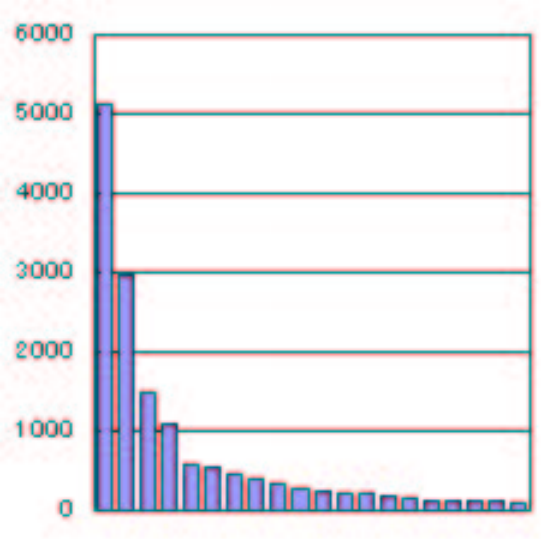

Figure 2: The detected viruses in October

2.1 Algorithm

Viruses detected by InterScan VirusWall which installed in the mail server of the University of Aizu are logged at refference [5]. This site told the probability that a specific virus come again is high if the virus arrived many in recently. In addition, it must have efficacy against the new type and the new type will appear one after another. Because of these consideration, we suggest updating condition which is decided by 3 patterns as top n of sum of arrived virus count from 1 month ago to yesterday (30 days), 1 week ago to yesterday (7 days), and in yesterday (1 day).

2.2 How Many Kinds of Virus We Should Scan

In last year, the most of kinds of virus had reached in this server was the first half of November because subspecies of *Worm.Netsky* and *Worm.Bagle* bled in large quantities. It simulates at this term and if it is possible to stop their invasion efficiently, it can be thought that it is effective at other term, too. Hence, we must use the information in October. This term had been confirmed about 70 kinds of viruses during 1 month, about 50 kinds during 1 week, and about 30 kinds during 1 day. By descending order, server extract top n virus information and try to stop malicious mails. Figure 4, 5, and 6 show

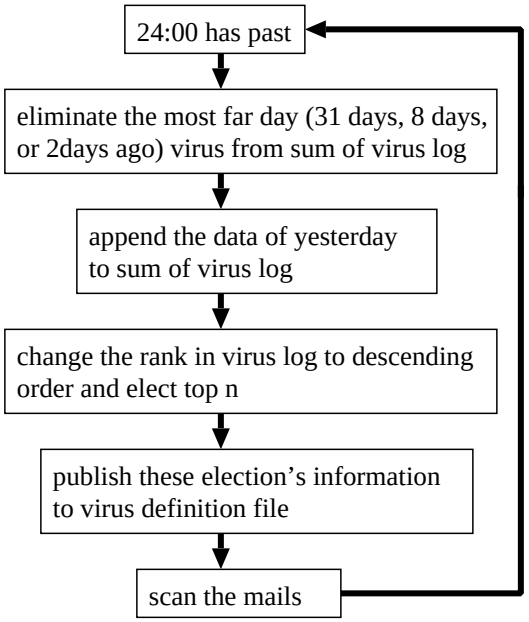

Figure 3: The flowchart of updating algorithm

the elimination rate against the all of the mail attached to the virus. As a reason that the rate is low in November 1st, a new type virus had come many. The month method cannot deal with the virus. These average are

	top 10	top 20	top 30	top 40
month	76.6%	91.7%	96.6%	99.0%
week	87.5%	94.5%	98.1%	99.1%
day	87.4%	96.1%	97.7%	——

	top 50	top 60	top 70	——
month	99.6%	99.8%	99.8%	——
week	99.7%	——	——	——
day	——	——	——	——

Table 1: the elimination rate of virus mails

As long as looking Table 1, the virus definition file which published only 30 kinds of virus information can stop more than 95 percent of the entire mails attached virus, and published only 40 kinds can stop more than 99 percent. These result shows published virus in virus definition file is enough to less than 100.

3 Model construction

The amount of arrived mails determine the aspect of server's congestion. After scanning, server deliver popular mails to users or eliminate malicious mails. The

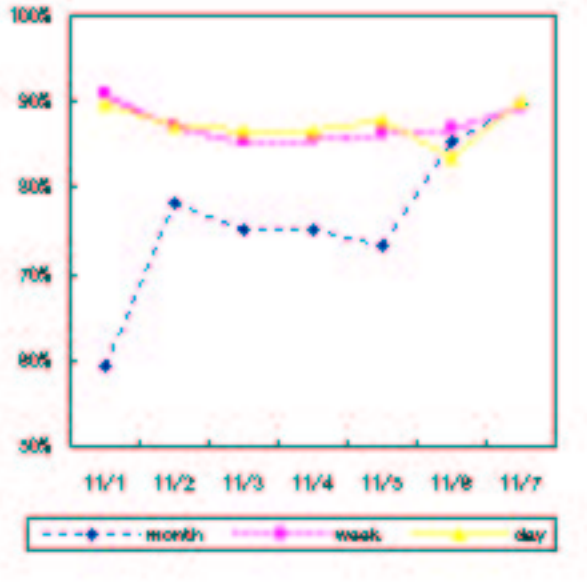

Figure 4: The rate of elimination virus mail (n=10)

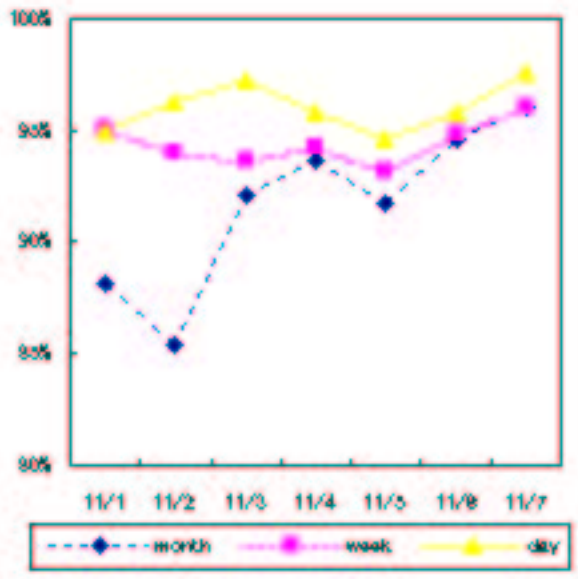

Figure 5: The rate of elimination virus mail (n=20)

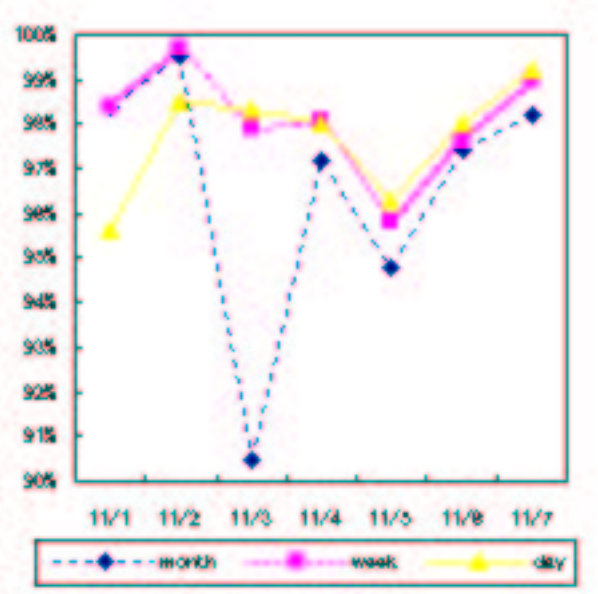

Figure 6: The rate of elimination virus mail (n=30)

mechanism seems queue. Since the interval of mail arrival seems Poisson and scanning time of a mail by a server is almostly as long as other mails, we capture the server as M/D/1 queuing system, reset each parameter to server system. The arrival time rate replace the mail received time interval at a server, the service time replace the time that the server scan one mail, number of windows replace number of servers, and without procession limit.

Table 2 is the amount of E-mail circulation including virus or spam mail from outside the university to inside or from inside the university to outside by the mail server (mail0) of the University of Aizu.

In last yaer, the day which comes the most of the amount of the mails is 15 in Octover and 41441 mails have come. It can compute the number of the mails who arrive in 1 second (= λ : the average of arrival rate) by using above table.

$$\lambda = \frac{41441}{24 \times 60 \times 60} \cong 0.48. \tag{1}$$

If the time that server scan one mail is S, the average of service rate (= μ) is

$$\mu = \frac{1}{S}. \tag{2}$$

Therefore, taking ρ (= λ / μ : the average of operating

DATE	INBOUND	OUTBOUND	TOTAL
2004.12	149573	281547	431120
2004.11	172534	321635	494169
2004.10	189718	327142	516860
2004.09	138861	280945	419806
2004.08	129507	263556	393063
2004.07	144116	267748	411864
2004.06	141637	279584	421221
2004.05	145633	293032	438665
2004.04	159881	305393	465274
2004.03	117199	254470	371669
2004.03	11709	29389	41179
2004.03	10546	26729	37275

Table 2: The amount of e-mail traffic during last year

rate of a server)

$$\rho = \frac{\lambda}{\mu} \tag{3}$$

Equation (3) leads the average of length in queue (=L) and L equals to

$$L = \rho + \frac{\rho^2}{2(1-\rho)}. \tag{4}$$

Also, the average of sojourn time (=W) equals to

$$W = \frac{1}{\mu} + \frac{\rho}{2\mu(1-\rho)}. \tag{5}$$

4 Result

Taking only n kinds of information published definition file, rate of new service time S' is defined as

$$S' = \frac{n}{68000}S. \tag{6}$$

The service rate μ' is

$$\mu' = \frac{1}{S'} \cong \frac{68000}{nS}. \tag{7}$$

Using same method as equation (3), (4), (5), each of parameter is expected. The average of server's operating rate ρ is changed by μ',

$$\rho' = \frac{\lambda}{\mu'} = 7.06 \times 10^{-6} nS. \tag{8}$$

The average of length in queue used reducing definition file define L'.

$$L' = \rho' + \frac{\rho'^2}{2(1-\rho')} = \frac{3.53 \times 10^{-6} nS(2 - 7.06 \times 10^{-6} nS)}{1 - 7.06 \times 10^{-6} nS}. \tag{9}$$

As the same, the average of sojourn time used reducing definition file define W'.

$$W' = \frac{1}{\mu'} + \frac{\rho'}{2\mu'(1-\rho')} = \frac{7.35 \times 10^{-6} nS(2 - 7.06 \times 10^{-6} nS)}{1 - 7.06 \times 10^{-6} nS}. \tag{10}$$

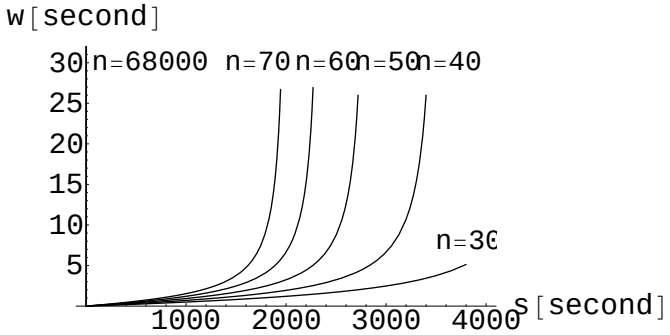

Figure 7: The relation S and W'

When compare Figure 7, n=30, 40, 50... shows extremely low the probability that yield waiting against n=68000 shows waiting arise where S is more 1 second. However, because of the time of scanning one mail is not so long, Figure 5 also may not waiting.

5 Conclusion and Future work

From above result, there are long differences of waiting time and it is understood to be able to correspond enough even if we don't use this system when amount of mail traffic is like the server of University of Aizu, however, if scanning viruses are several ten kinds, can deal with distress viruses for scanning such as disguise type which tell not virus. Therefore, instead of using the waste time and toil to lookup old viruses, server can use the processing ability to scan new type virus which hard to detect. My future work is research the measure against the attacking viruses in a special day.

Acknowledgment

I would like to thank my supervisor Prof. Toyoizumi and master in Performance Evaluation lab for helpful comments, suggestions, and supporting my works. Also I'd like to thank Prof. Brain for guidance of English.

References

[1] H Toyoizumi, Detect the Source of Worms with Spoofed Email Address, May, 2004

[2] Keiichi Kato, Modeling Penetration of Viruses at the Gateway, Graduation Thesis, March, 2004

[3] Symantec, http://www.symantec.com/

[4] IPA, http://www.ipa.go.jp/

[5] Information Processing Center, http://web-int/labs/istc/ipc/index.html

[6] D.P.Heyman, M.J.Sobel, Stochastic Models, 1990

[7] Performance Evaluation, http://www.u-aizu.ac.jp/toyo/lectures/PE/PerforamnceEvaluation.htm

[8] trendmicro, http://www.trendmicro.com/

[9] Sheldon M.Ross, Stochastic Process Second Edition, 1996