Internet Virus Spreading and Control Based on Complex Network Theory

Wang Tao [1, *]

[1]Xijing University, No.1, Xijing Road, Chang'an District, Xi'an City, Shaanxi Province, China

[*]dfyang_dfyang@126.com

Key words: complex network; internet; computer virus; spreading; control

Abstract: Based on complex network theory and a full understanding of the characteristics of internet complex network, this article gives an overview of epidemiological models of computer virus spreading, the virus spreading model of e-mail, the spreading models of stochastic constants, interval models and the spreading critical value theory of computer viruses on associative network and non-associative network. On this basis, the article has analyzed the two mechanisms including stochastic immunization and optional immunization controlled by internet viruses. In addition, by making use of stochastic immunization and targeted immunization, viruses are difficult to spread, thus improving the reliability of the whole network system.

1. Introduction

With the emergence and rapid development of the internet, the dependence of human society on the internet is becoming higher and higher. Therefore, the safe reliability of the internet appears to be very important. With the rapid development of the internet, computer viruses spring up accordingly, which has posed a grave threat to the safety of the internet. For the purpose of strengthening the internet's resistance to the attack from viruses and resisting computer viruses in a more effective way, it is necessary to have a deep understanding of the spreading mechanism of computer viruses in the internet. Based on complex network theory and a full understanding of the complex network characteristics of the internet, the article has researched the spreading and control of computer viruses on the internet.

2. Complex Network and the Internet

By complex network, it refers to a graph G (V, E) composed of a vertex set V(G) and an edge set E(G). From the perspective of the distribution of network node degree, a complex network has two main different topological structures. One is uniform network, for example, the stochastic network put forward by Erdös and Rényi (ER). By uniform network, it means that the degree of each node in the network is almost equal to that of any other node. As a rule, the distribution of the degrees of the nodes is exponential distribution. The opposite of uniform network is non-uniform network with remarkable non-uniform characteristics. One of the typical examples is the Scale-Free (SF) network proposed by Barabasi and Albert (BA). Its distribution of degrees is power-law distribution.

Up to now, the structure of the internet is described from two main different angles. One is based on the level of router. That is to say, each router is regarded as a node and the links between routers are considered as edges. The other is based on Autonomous System (AS). By AS, it refers to various sub-networks allowing for adopting different internal routing algorithms. Each AS stands for a node. If there is a BGP (Border Gateway Protocol) peer-to-peer connection between two Autonomous Systems, it indicates that there is an edge between the two nodes. Since the internet is characteristic of strong power-law distribution, it is a non-uniform network whose node's degree is of great volatility [1,2]. It is well-known that the increase of the internet follows the principle of "the rich get richer". New users always tend towards looking for servers or routers with higher reputation and a large number of links to set up connection with them. With the passing of time, the nodes and edges in the system keep on increasing. However, the characteristics of network topology cannot undergo great changes. Besides, the internet has its hierarchical structure, which allows

interconnections between network nodes. Therefore, correlation is a feature of the internet that cannot be ignored. Besides, in the internet, when new connections are added, routers are always prone to connect to the internet by way of the shortest distance, thus leading to more connections between the routers in the same area while there are relatively a small number of connections between routers in different areas. Therefore, there is a larger clustering coefficient between the routers in the same area. For this reason, Wang Xiaofan et al have proposed a MLW (multi-local-world) model [3].

3. Models of Computer Virus Spreading

Computer viruses are a program with special functions made by some people who take advantage of the inherent vulnerability of computer software and hardware. The number of computer viruses increases exponentially every year. In addition, due to the change of transmission medium and the wide application of the internet in recent years, virus infection objects have began to shift from workstation (terminal) to network components (agency, protection and service settings). Accordingly, the types of viruses have also shifted from document type to the type of network worms. Models of computer virus spreading are always the issue that arouses people's interest. Some related models in this field have been set up.

3.1 Epidemiological Models

In the early 1990s, Jeffrey O. Kephart and Steve R. White noticed some general characteristics of viruses (whether biological viruses or computer viruses). They introduced the analytical method in biology into the research on computer viruses and analyzed the spreading of computer viruses based on network topological structure [4]. Up to now, the basic thought of epidemiological models is still the important basis on which the spreading models of computer viruses are established. In an epidemiological model, the individuals in population are abstractly defined as some limited typical states. An epidemiological model includes the following basic states: S: susceptible state, I: infective state and R: recovered state. As a rule, the conversion process between these states is used to name the model. For example, if the susceptible population is infected before recovering health and having immunity, it is called SIR Model. However, if the susceptible population is infected before returning to susceptible state, it is called SIS Model. Besides, some simple probabilities are employed to describe the shift between the states of such models.

The uniformity and symmetry of SIR model and SIS model can influence the analysis of computer viruses. There are often the exchanges of programs or documents between computers. These computers with exchange relations are usually in groups that are closely related to each other. The connection of such network is of obvious non-uniformity. Besides, the spreading of computer viruses has its preferential direction. Therefore, efforts should be made to improve the analysis of computer virus spreading by way of SIR model and SIS model and further to consider factors such as network connections and correlation.

3.2 E-mail Virus Spreading Model

Zou et al. has put forth an e-mail virus spreading model [5]. In the model, e-mail virus spreading follows this process below: Users always check their e-mails at fixed periods. When a user checks his mailbox and meets with an e-mail whose appendix contains viruses, he is likely to give up this letter or open the appendix of this letter. Once the appendix with viruses is opened, the viruses will immediately infect the user and send e-mails with viruses to all the e-mail addresses in the user's e-mail address book. In the model, main consideration is given to the following two human factors that influence viruses: the check time of e-mails and probability of opening these e-mails. A user's check time of his e-mails depends on his individual habit while the probability of opening his e-mails is determined by his knowledge and awareness of e-mail viruses.

However, this model has some shortcomings. To each and every user, the probability of opening appendix with viruses on is a fixed value and a function about the user, which has nothing to do with time. This cannot excellently describe the actual behavior on the part of the user. In fact, most

of experienced users can evade all the appendixes with viruses. That is to say, the probability for them to open appendixes with viruses is zero while inexperienced users are likely to open any appendix and the probability for them to open appendixes with viruses is almost one. With the passing of time, some inexperienced users will come to realize that appendixes may carry viruses so that they sharpen their precaution consciousness. As a result, the probability for them to open appendixes with viruses will decrease gradually. In other words, the probability of opening appendixes with viruses is changeable with time instead of being changeless. On the other hand, the reliability that actual e-mail network is inferred to conform to the power-law distribution only on the basis of the group of e-mails that complies with the power-law distribution is to be verified.

3.3 Stochastic Constant Spreading Models

By using some empirical data obtained in the course of the outbreak of Code Red Worms, S. Staniford , V. Paxson and N. Weaver set up a model—Stochastic Constant Spread Model which is called RCS for short [6]. This model contains the following three approximate conditions: firstly, that the system may be patched or closed or its connection is severed is ignored; secondly, the number of the targets vulnerable to attack is regarded as a constant in this model and thirdly, the internet is considered as an undirected and completely connected graph.

Set k as an average initial infection rate, namely, the number of the hosts vulnerable to infection which are attacked by infected hosts in unit time when each round begins. In the model, k is assumed as a constant, thus ignoring the differences caused by different processing speeds, network bandwidths and positions of infected hosts. Set α as the proportion of the infected hosts at the moment t and $N\alpha$ indicates the number of infected hosts. These hosts scan other infected hosts at the rate k in unit time. Since the hosts with the proportion α have been infected, each infected host will make $k(1-\alpha)$ new hosts infected in unit time. Therefore, the expression of n (the number of hosts to be infected) in the time slot dt can be obtained as follows (here α is assumed as a constant):

$$n = (N\alpha) * k(1-\alpha)dt$$

Since the address space is as large as 2^{32} and the list of the targets scanned by worms is stochastic, the possibility that two different worms will attack the same target at the same moment is very small, which is negligible. Under the assumption that N is a constant, $n = d(N\alpha) = Nd\alpha$ is obtained. Finally, the following differential equation model is obtained:

$$\frac{d\alpha}{dt} = k\alpha(1-\alpha)$$

3.4 Interval Models

Facing the analysis of the worms called “flash” or “Warhol”, most of traditional models can be confronted with barriers. The common explanation given to the phenomenon is as follows: a large part of the network is unable to guarantee the sufficient required bandwidth when worms are spreading at full speed. In other words, the bandwidth has been saturated before the number of infected hosts is saturated. In order to better analyze the computer viruses whose spreading is restricted by bandwidth, a new model—interval model based on the AS structure of the internet is put forward in document [10]. In this document, the factor of the bandwidth between nodes is taken into consideration. In the same AS, the spreading of viruses is unimpeded, which conforms to RCS model. It is necessary to give some reconsideration between different Autonomous Systems.

N_i denotes the number of infected hosts in No. i AS （AS_i） while a_i indicates the proportion of infected hosts in AS_i i. Set k as the average speed of the virus spreading. In each independent *AS*, k is approximate to a constant. $P_{IN,i}$ stands for the probability of the hosts in a AS_i that are attacked by other hosts in the same AS_i while $P_{OUT,i}$ represents the probability of the hosts in another *AS*.

$$P_{IN,i} = N_i / N$$

$$P_{OUT,i} = 1 - P_{IN,i} = N_j / N, j \neq i$$

N differential equations have been obtained.

$$\frac{da_i}{dt} = [\sum_{j=1}^{n} N_j a_j](1 - a_i)\frac{K}{N}$$

This is a group of non-linear differential equations. In No. k *AS*, viruses increase on the basis of RCS model while other Autonomous Systems are in a poised state for the time being. After this state lasts for a period of time, the viruses in this *AS* spread beyond the *AS* for the first time. Designate s as the size of the viruses, namely, the size of the bandwidth occupied by virus attack. r_j is the number of attack launched in AS_j in unit time. M refers to the total number of computer systems that have appeared in the internet and M_i indicates the number of the computer systems in AS_i.

The average value of the speed of the attack launched by a single infected host is $R \cong K\frac{M}{N}$ （K is the speed of successful outward attack）. The total bandwidth consumed by AS_i is garnered by adding the bandwidth $b_{i,inco\min g}$ consumed by external viruses to the bandwidth $b_{i,outco\min g}$ consumed by outward attack:

$$b_i = \frac{sK}{N}[M\sum_j N_j a_j - M_i a_i N_i]$$

As the extension and expansion of RCS models, interval models give a better explanation of the phenomenon that in simulation, attacks keep on increasing before ceasing abruptly. This phenomenon is produced by the internet connection interruption caused by the network overload due to insufficient bandwidth. Of course, this is only the mitigation of attacks and cannot be inferred as the slowdown in virus increase.

4. Critical Value Theory

From the perspective of critical value, Pastor-Satorras et al. conducted a series of research of the spreading of computer viruses in the network, including different properties embodied in uniform networks and non-uniform networks as well as the similarities and differences In associative networks and non-associative networks.

4.1 Spreading of Computer Viruses in Non-associative Networks

Firstly, we put aside the correlation between different nodes and only take into consideration the virus spreading in non-associative networks. Through the comparison between the spreading characteristics of uniform networks and those of non-uniform networks, we can describe the spreading mechanism of computer viruses in the internet in a clearer way.

4.2 Virus Spreading in Uniform Networks

Here, the standard SIS model in epidemiology is taken to make an analysis. In SIS model, all nodes belong to two states: vulnerable to infection and infected. The probability of the nodes from the state of being vulnerable to infection to the state of being infected is set as p_v and the probability of returning to the state of being vulnerable to infection from the state of being infected is designated as p_δ. The effective spreading speed is set as λ, and $\lambda = p_v / p_\delta$. The specific value of the probability of being infected to that of being cured on the part of nodes is used to measure the spreading performance of viruses in the network. In addition, The density of infected nodes at the moment t is set as $\rho(t)$, indicating the proportion the number of infected nodes to the total number of nodes. When the time t tends to infinite, the steady-state density ρ of the system is obtained.

Here are three assumptions. Since consideration is given to uniform network with a very small

fluctuation of node degree, each node is assumed to have the same degree k, which is tantamount to the average degree of the network $< k >$. The second assumption is called uniform-mixing hypothesis which means that infected intensity is proportional to the density $\rho(t)$ of infected individuals. The last assumption is that the time scale for describing viruses is much less than the life cycle of individuals. So both the birth of individuals and their natural death are beyond consideration. Therefore, we can obtain the expression of the steady-state density of infected individuals as follows:

$$\rho = \begin{cases} 0 & \lambda < \lambda_c \\ \lambda - \lambda_c / \lambda & \lambda \geq \lambda_c \end{cases}$$

In the expression, $\lambda_c = < k >^{-1}$ is defined as a spreading critical value or threshold value. In the case that the effective spreading speed is above the critical value, infection can spread and end up being stabilized in a equilibrium state. However, in the case that the effective spreading speed is less than the critical value, the infection undergoes an exponential decrease, unable to spread extensively. Therefore, in a uniform network, there is a positive critical value which separates the infected state from the healthy state. However, the result of the aforementioned theory is inconsistent with the actual data analysis of computer virus spreading. This inconsistence indicates that computer viruses do not spread in the uniform network.

4.3 Virus Spreading in Non-uniform Network

Now, let's put aside the assumption about uniformity and research the characteristics of computer virus spreading in SF network. Set the relative density $\rho_k(t)$ as a probability of an infected node whose degree is k. In SF network, the higher the degree of node is, the higher the probability of being infected is. In this case, the expression of λ_c turns into: $\lambda_c = < k > / < k^2 >$. When the exponent γ of the power-law distribution $p(k) - k^{-\gamma}$ satisfies the condition $2 < \gamma \leq 3$, the unbounded situation of degree's fluctuation in the infinite SF network is expressed as $\lambda_c = 0$. That is to say, in the infinite SF network, the critical value of spreading speed disappears. In SF network, whatever the spreading speed is, viruses can spread and end up staying in an equilibrium state as long as the spreading speed is greater than zero. This is the so-called "vulnerability" when SF network is faced with the spreading of computer viruses. As the internet has the characteristics of SF network, computer viruses can spread extensively by infecting a small part of nodes on the internet. Thankfully, since the spreading speed λ has a large variation range, its value can be small and λ tends to zero in the actual internet. Therefore, generally speaking, computer virus spreading only reaches a low degree of infection in the end. The conclusion that the critical value disappears in infinite SF network is just identical to the data from actual observation, thus explaining the phenomenon characterized by the inconsistence between theory and practice in uniform networks.

It should be pointed out that the conclusions of critical value obtained in SIS model hold water in SIR model. In other word, different infection processes do not influence the critical value characteristics of the virus spreading in both a uniform network and a non-uniform network.

4.4 The Spreading of Computer Viruses in Associative Networks

The analysis above is about the situation in non-associative networks. Correlation is a feature that cannot be ignored in the internet, which is all the same important to understanding the spreading of computer viruses on the internet. Boguná and Pastor-Satorras carried out a further research on the spreading of computer viruses in associative network. In an associative network, the expression of critical value formula is: $\lambda_c = 1/\Lambda_m$, in which Λ_m stands for the maximum eigenvalue of connection matrix.

Moreno et al. pointed out that the event probability of virus spreading in associative networks is smaller that that in non-associative networks with the same degree distribution. The related properties of spreading critical value do not change because of the emergence of correlation.

However, the spreading life cycle in associative networks is longer than that in non-associative networks. Besides, in an associative network whose size is limited, the critical value of spreading speed is a little greater, indicating that these networks demonstrate a stronger robustness than non-associative networks.

4.5 Immune Mechanism of Complex Networks

Immunization is an important method of restraining computer viruses. At present, there are mainly two immunization strategies: random immunization (uniform immunization) and selected immunization (targeted immunization) [7].

4.6 Random Immunization

Random immunization equally treats the nodes with a greater degree (a high risk of being infected) and those with a less degree. While being immunized, nodes are chosen in a random way without priority. In document [7], the proportion of immunized nodes to the total number of nodes is set as g and a critical value g_c is obtained. The expression of g_c is written as $g_c = 1 - \lambda_c / \lambda$. When g is greater than the value, the number of being infected nodes in the end is zero. Set ρ_g as the final infection degree (the final infected density) and the expression is as follows:

$$\rho_g = \begin{cases} 0 & g > g_c \\ (g_c - g)/(1-g) & g \le g_c \end{cases}$$

The method of random immunization performs almost no function in SF networks. At whatever speed λ computer viruses spread, the speed of random immunization method is very slow thought it can partially restrain infection. At the same time, in infinite SF networks, the disappearance of the critical value λ_c means that here the critical value $g_c = 1$, that is, the method of random immunization needs all the nodes in an immune network before ensuring that infection is wiped out in the end, which is obviously not realistic. Since the internet is a kind of SF network, the outbreak of computer viruses cannot be prevented even though a lot of nodes in the network are immunized in a random way.

4.7 Targeted Immunization

Consideration should be given to how to make use of the non-uniform characteristics of the internet to conduct selective targeted immunization. The nodes with greater degrees are selected to be immunized. Once these nodes are immunized, it means that the edges connected with them can be wiped out from the network, thus reconstituting the structure of the network. In addition, the critical value of immunization obtained from targeted immunization will be much smaller.

At present, users will keep on installing some anti-virus software on the internet. However, the life cycle of computer viruses is rather long. The reason lies in that the processes of file scanning and anti-virus updates are in fact a process of random immunization. In the overall internet, even though a lot of nodes are immunized, the spreading of computer viruses cannot be exterminated. Therefore, although the method of targeted immunization is an effective method, this method requires an understanding of the overall information of the network, at least a clear understanding of the degree of each node in the network. Only in this way can we find the nodes with higher degrees and immunize them. As far as the giant and complicated internet is concerned, this is not a very easy thing.

5. Conclusion

The complex network theory has provided new ideas and methods for the research on the spreading of computer viruses on the internet. In terms of the computer viruses that spread on the same internet, although the research is still based on the relevant theories about the spreading of the internet viruses, we cannot analyze their spreading characteristics blindly based on the structural characteristics of the internet. It is necessary for us to make a concrete analysis of a concrete

problem. Consideration should be given to the specific network topology under different virus spreading mechanism on the basis of the internet topological structure. As with different network structures, different control strategies produce different effects. In actual work, we can make it difficult for viruses to spread by combining random immunization and targeted immunization so as to improve the reliability of the entire network system.

References

[1] Vázquez A , Pastor-Satorras R , Vespignani A. Large-scale topological and dynamical properties of the Internet [J] . Phys. Rev E , 2002 ,65 (6) : 066130.

[2] Faloutsos M , Faloutsos P , Faloutsos C. On power2law relationships of the Internet topology[J] . Computer Communication Review , 1999 , 29 (4) : 251 - 262.

[3] Chen Guanrong , Fan Zhengping ,Li Xiang. Modelling the complex Internet topology[M] . Complex Dynamics in Communication Networks[M] , Springer Publisher , in press , 2004.

[4] Kephart J O , White S R. Directed2graph epidemiological models of computer viruses[A] . Proceedings of the 1991 IEEE Symposium on Security and Privacy[C] . Oakland ,California ,USA : IEEE Computer Society Press ,1991. 343 - 359.

[5] Cliff C Z ou , Towsley D , G ong Weibo. Email virus propagation modeling and analysis[R] . Technical Report TR2CSE203204 , University of Massachussets , Amherst .

[6] Staniford S , Paxson V , Weaver N. How to own the Internet in your spare time[A] . Proceedings of the 11th USENIX Security Symposium[C] . San Francisco ,USA , 2002.

[7] Pastor2Satorras R ,Vespignani A. Epidemics and immunization in scale-free networks[Z] . Bornholdt S. Handbook of Graphs and Networks : From the Genome to the Internet [M] , Wiley2VCH ,2002.