Proposal for a Communication Link Model Based on Resonance Frequency of Network Users

Masato Uwajima[†] Sasaki Toru[†] Chisa Takano[‡] Masaki Aida[†]
[†] Graduate School of System Design, Tokyo Metropolitan University
6-6, Asahigaoka, Hino-shi, Tokyo 191-0065, Japan
E-mail: {uwajima-masato,maida}@sd.tmu.ac.jp

[‡] Graduate School of Information Sciences, Hiroshima City University
3-4-1, Ozuka-higashi, Asa-minami, Hiroshima 731-3194, Japan
E-mail: takano@hiroshima-cu.ac.jp

Abstract

Communication network data, such as the volume of traffic and the number of users, has been mainly used for facility engineering of communication networks. This use focuses on quantitative data. The qualitative data on a communication network can be seen as analogous to the characteristics of human activity in society. If such sociological information can be extracted from communication network data, it would be possible to develop technology that supports enterprises in developing their marketing strategies. We have found that the power law is applicable to the volume of cellular phone traffic and the number of SNS users, and, using this property, we have identified the structure of a social network used for the exchange of information between SNS users. In this paper, we investigate the characteristics of the communication frequency on links in social networks by using the power law observed in the process of computer virus infection, and propose an information propagation model. Our simulation result shows that the proposed information propagation model exhibits characteristics similar to those in the data obtained from a real SNS.

1. Introduction

Knowledge about trends in user preferences and behavior is critical for enterprises in developing their business operations or marketing strategies. If enterprises can find out how users are responding to their products or services, they can make appropriate responses to these reactions. For example, they can use such information to determine whether their new products have been widely accepted by users and consequently whether it is wise to withdraw from the market rather than make further investment. Enterprises usually ascertain user responses through questionnaire surveys or by asking research companies to conduct a survey. However, these methods are costly, time-consuming and often not very reliable. Therefore, enterprises still find it difficult to reflect user responses in their business strategies.

If we can assume that trends in user behavior reflect the universal characteristics of the structure of a social network, it can be expected that by studying this structure we can obtain the basic understanding required for the development of marketing strategies that are independent of specific products or services. However, the use of questionnaires is too costly and time-consuming for a detailed survey of the structure of a large social network.

We can assume that data on a communication network, such as the volume of traffic and the number of users, reflect the characteristics of human societal behavior in one way or another. If such sociological information can be extracted from data on a communication network, it should be possible to develop technology that supports enterprises in developing their marketing strategies. The aim of our present study is to examine a variety of data on a communication network in order to ascertain the structure of the social network of its users, and to develop technology useful for the development of marketing strategies. We have found that the power law is applicable to the volume of cellular phone traffic and the number of SNS users,and, using this property, we have elucidated the structure of a social network used for the exchange of information between SNS users [4],[3].

In this paper, we use the power law observed in the process of computer virus infection, and the power law observed in the communication time for exchange of files

978-0-7695-3266-0/08 $25.00 © 2008 IEEE
DOI 10.1109/SSIRI.2008.36

using file sharing software to examine the frequencies at which people access the network. In addition, we use the analogy of the resonance frequency of a band pass filter to examine the frequencies at which communication links between users are used. We have conducted a simulation of information propagation in the network model given in Reference [3], and confirmed that the simulation gives results similar to the data obtained from a real SNS.

2. Analysis of the frequency at which people use the network

This section examines how people use the network by analyzing the process by which computer viruses are spread.

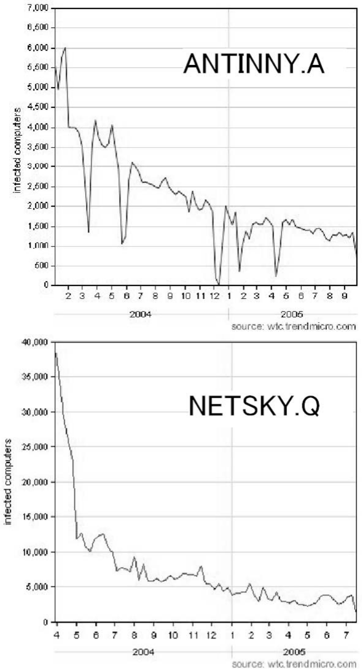
There are several types of computer viruses. In this paper, we will consider only worm-type viruses (hereinafter simply referred to as worms). A worm operates independently rather than parasitizing another program. It spreads itself over the network by taking advantage of emails and file exchange protocols. The user's computer is infected only when he or she executes an infected file (filename.exe) received via email or a file exchange protocol. In other words, infection is triggered by an intentional action by the user to obtain information through the network. Therefore, the process of worm propagation can be considered as reflecting the characteristics of user behavior when he or she actively uses the network.

2.1. Characteristics of worm-type computer virus propagation

We analyzed the data on changes over time in the number of hosts infected with computer viruses as posted on the Web site of Trend Micro Incorporated [2]. Since there are different types of worms, we decided that worms that satisfy the following two conditions met the objective of our study.

- Many hosts were infected;
- Hosts were infected through the network via email or file exchange protocols.

Figure 1 shows changes over time in the number of hosts infected by ANTINNY.A and NETSKY.Q, both of which satisfy the above two conditions. The infection of ANTINNY.A was spread by a file sharing program called Winny, while that of NETSKY.Q was spread mainly by email. The horizontal axis shows time, and the vertical axis shows the number of newly infected hosts detected by Trend Micro. The figures show that the number of newly infected hosts tends to decrease over time. In order to examine this trend more closely, we plotted the data on a double logarithmic chart as shown in Figure 2.



source: wtc.trendmicro.com

source: wtc.trendmicro.com

Figure 1. Changes over time of newly infected hosts

On the horizontal axis, $t = 0$ is the first time the worm appeared (which is also the date when an anti-virus program began to be distributed). A line with a gradient of 2/3 is also shown as a reference. Figure 3 shows similar charts for NETSKY.P and HTML_NETSKY.P, which also satisfy the above-mentioned two conditions.

These charts show that the number of newly infected hosts decreases at a rate of t to the power of 2/3 irrespective of the type of worm, where t is the elapsed time from the first appearance of the worm [5]. Since this characteristic is not dependent on the type of worm, we can conclude that it reflects the behavior of users rather than the behavior of the worm.

2.2. Distribution of network access frequencies

In order to roughly explain the above decrease in the number of infected hosts, we consider f to be the frequency at which the user accesses the network, and assume an ideal situation in which each user has his or her specific f. The factor that directly affects f is the user's active attempts to execute an infected file. Indirect factors that determine the frequency at which the user takes such an action are as follows:

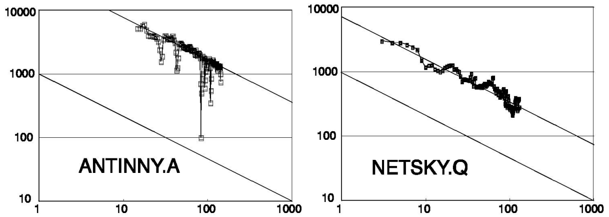


Figure 2. Number of newly infected hosts on a double logarithmic chart (1)

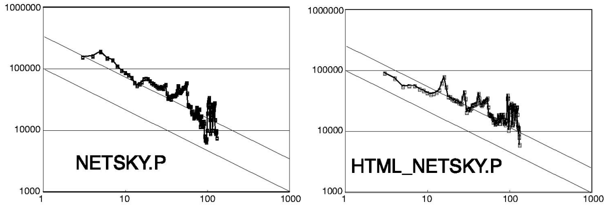


Figure 3. Number of newly infected hosts on a double logarithmic chart (2)

- In the short term, the number of emails sent or received and the number of exchanges of files using a file exchange application per unit time;
- In the medium term, the frequency at which the user's computer is started;
- In the long term, changes of the computer or the user, such as may occur when the user buys a new computer, changes his or her ISP, or is transferred to a new organization.

Let $\eta(t)$ be the number of users accessing the network at time t. Let us consider the following Fourier transform of $\eta(t)$ with the network access frequency f as the frequency.

$$H(f) = \int_{-\infty}^{\infty} \eta(t)\, e^{-2\pi i f t}\, \mathrm{d}t \qquad (1)$$

We assume that time $t = 0$ has no specific meaning, and that $\eta(t)$ is an even function having the same property in both the future and past directions, namely $\eta(t) = \eta(-t)$. Then $H(f)$ becomes a real function. Here, $H(f)\,\mathrm{d}f$ is the number of users in the network access frequency range of $(f, f + \mathrm{d}f]$.

2.3. Relation between the number of newly infected hosts and $H(f)$

Let $I(t)\,\mathrm{d}t$ be the number of newly infected hosts during period $(t, t + \mathrm{d}t]$ starting after time t has elapsed from the first appearance of the worm.

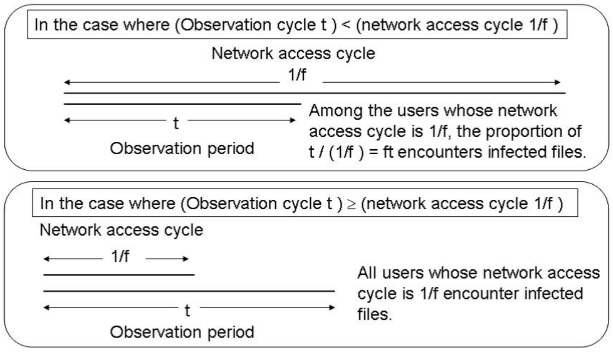

In the case where (Observation cycle t) < (network access cycle 1/f)
Network access cycle
1/f
t
Observation period
Among the users whose network access cycle is 1/f, the proportion of t / (1/f) = ft encounters infected files.
In the case where (Observation cycle t) ≥ (network access cycle 1/f)
Network access cycle
1/f
t
Observation period
All users whose network access cycle is 1/f encounter infected files.

Figure 4. Rate at which infected files are encountered

The above section analyzed worms that infected many hosts. It is likely that such worms are widespread throughout the network. (It is to be noted that not all hosts that have infected files in them are infected because they are infected only after an infected file is executed.) Therefore, the probability of a network user encountering an infected file is determined by how often he or she accesses the network. We assume that a certain percentage of users execute an infected file inadvertently (or deliberately).

First, we consider the cumulative number of infected hosts $\int_0^t I(s)\,\mathrm{d}s$ for a period $(0, t]$. If the network access frequency f is $1/t$ or greater (i.e., $1/f \le t$), all hosts whose network access frequency is f encounter an infected file during time t. On the other hand, if f is lower than $1/t$ (i.e., $t < 1/f$), the rate at which all hosts whose network access frequency is f encounter an infected file during time t is ft (< 1) (see Figure 4). From this, the cumulative number of infected hosts becomes

$$\int_0^t I(s)\,\mathrm{d}s = c \int_0^{1/t} ftH(f)\,\mathrm{d}f + \int_{1/t}^{\infty} H(f)\mathrm{d}f \qquad (2)$$

where c is constant, and is the proportion of users who execute an infected file inadvertently. We differentiate both sides of the equation with t and use $I(t) = O(t^{-2/3})$ to obtain

$$I(t) = c \int_0^{1/t} fH(f)\,\mathrm{d}f = O(t^{-2/3}) \qquad (3)$$

The integrant becomes $H(f) = O(f^{-1/3})$. Therefore, we get

$$H(f) = O(f^{-4/3}) \qquad (4)$$

3. Network access frequency as seen from the distribution of Winny connection times

This section aims to obtain additional evidence that supports Equation 4 by analyzing the distribution of Winny connection times.

3.1. Distribution of Winny connection times

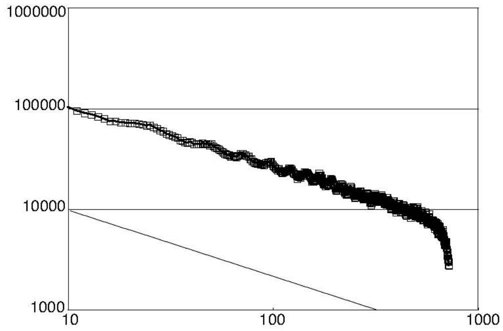


Figure 5. Distribution of Winny connection times

Ohzahata et al. measured the connection times of Winny users on a real network [6]. Figure 5 shows the result of their measurement, which was made over a 30-day period. The double logarithmic chart in Figure 5 shows Winny connection time (in seconds) on the horizontal axis and the number of users for the corresponding connection time on the vertical axis. The density function $w(t)$ of the Winny connection time for t becomes

$$w(t) = O(t^{-2/3}) \tag{5}$$

The behavior at the right bottom corner of the chart does not match Equation 5 because the measurement period was finite.

3.2. Analysis of network access frequency

As we did with the analysis of worms, we will study the distribution of the connection time of the file exchange application. We assume that a file exchanging host terminates its connection as soon as it has obtained the desired file. This means that both the recipient and the host that has the desired file have connections at the same time. Let $w(t)$ dt be the probability at which the desired file is found during time $(t, t + \mathrm{d}t]$, where t is the time elapsed from the beginning of the connection setup. Let X be the connection time, then $w(t)\,\mathrm{d}t = P[t < X \leq t + \mathrm{d}t]$.

First, we will consider the case of $P[X \leq t]$. If the network access frequency f of the host that has the desired file is $1/t$ or greater (i.e., $1/f \leq t$), the connection time is t or smaller. On the other hand, if the network access frequency f of the host having the file is smaller than $1/t$ (i.e., $t < 1/f$), the probability at which the recipient searching for the file and the host having the file are connected at the same time within the connection time t becomes $ft(< 1)$. Thus, the distribution of the connection times becomes

$$P[X \leq t] = \int_0^{1/t} fta(f)H(f)\,\mathrm{d}f + \int_{1/t}^{\infty} a(f)H(f)\,\mathrm{d}f \tag{6}$$

where $a(f)$ is a weighting function used to take into consideration the case where the hosts that have the desired files tend to have a specific network access frequency, and thus $a(f)H(f)$ indicates the probability density that the host that has the desired file has the network access frequency of f. As in the case of worms, we differentiate both sides of the equation to get

$$w(t) = \int_0^{1/t} fa(f)H(f)\,\mathrm{d}f \tag{7}$$

Since we know Equation 5 from the measurements, we get

$$\int_0^{1/t} fa(f)H(f)\,\mathrm{d}f = O((1/t)^{2/3}) \tag{8}$$

Therefore, the integrand becomes

$$a(f)H(f) = O(f^{-4/3}) \tag{9}$$

From the analysis of computer viruses, we know that $H(f) = O(f^{-4/3})$,

$$a(f) = O(1) \tag{10}$$

This indicates that there is no noticeable deviation in the distribution of the network access frequencies of the hosts holding desired files. In other words, the rate at which files spread through file exchange using Winny is proportional only to users' network access frequency as is the case with the spread of worms.

Conversely, as can be inferred from the behavior of the increase in the number of recipients newly infected with ANTINNY.A via Winny, if we assume that the files exchanged by Winny propagate as computer viruses do, we can say that the data on the connection time for Winny provides corroborative evidence that the distribution of network access frequencies is

$$H(f) = O(f^{4/3})$$

4. Relationship between the degree distribution of human relations and network access frequencies

From the data on the relationship between the number of mobile phone users and the amount of traffic and on the change in the number of SNS users over time, we know the following about the degree distribution in a users' information exchange network [1].

$$n(k) = O(k^{-4}) \tag{11}$$

where k is the degree of users, and $n(k)$ is the number of users having degree k.

We assume the following in order to examine the relationship between the structure of a users' information exchange network and the network access frequency.

Assumption: Network access frequency f is a function of degree k.

Substituting $f(k)$ into f in Equation 4, and considering that $H(f(k)) = (H \circ f)(k)$ is a function of k, we get

$$H(f(k)) = (H \circ f)(k) = O(f(k)^{-4/3}) \tag{12}$$

$(H \circ f)(k)$ is the number of users having degree k, and thus is

$$n(k) = (H \circ f)(k) \tag{13}$$

From Equation (11), we get

$$f(k) = O(k^3) \tag{14}$$

This result is also reasonable from the intuitive expectation that users with a large k, i.e., a large number of entities to communicate with, tend to be heavy users with a high level of network access frequency.

Let $L_f(k)$ be the communication frequency on a link of the user having degree k. From the above discussion, $L_f(k)$ becomes

$$L_f(k) = \frac{f(k)}{k} = O(k^2) \tag{15}$$

Therefore, it is in the order of the degree squared.

5. Proposal of a model for the communication frequency on a link

The previous sections examined the network access frequency of a specific user. The characteristics found are specific to the user and do not describe the communication frequency on links between users. In this section, we consider the interaction between the communicating parties, and discuss a model that determines the communication frequency on a link.

In examining the communication frequency on a link, we do not consider the type of communication that is completed by the user unilaterally sending information to the other party. Rather, we consider an interactive communication that is completed only when the other party responds to the originating party. We consider that this is "real-world" information propagation that can be applied to word-of-mouth and other types of marketing. For example, in a large SNS called mixi [1], a user can join the SNS only at the invitation of the existing user. In an information propagation process like this in which communication is considered completed only when an invited user accepts the invitation of an existing user, it can be considered that the real information propagation process is determined by a relation that takes into consideration not only the communication frequency of one party but also that of the other party.

Let us consider a link that connects two nodes, H and L. Let the degrees of H and L be k_h and k_l ($k_h \geq k_l$) respectively. Let the communication frequencies of H and L be $L_f(k_h)$ and $L_f(k_l)$ respectively. It is reasonable to assume that the communication frequency on the link does not exceed the access frequency $L_f(k_h)$ of the user on the high frequency side and that it does not go below the access frequency $L_f(k_l)$ of the user on the low frequency side. Therefore, the communication frequency on a link should be a value between $L_f(k_h)$ and $L_f(k_l)$. Since human sense obeys logarithmic scale, we assume that the communication frequency on a link is given by the geometric average of the frequencies of two ends as

$$\sqrt{L_f(k_h) \cdot L_f(k_l)} = O(k_h \cdot k_l) \tag{16}$$

This model means that the communication frequency on a link is determined by the product of the degrees of the two ends of the link.

Incidentally, we can consider an analogy. Since the communication frequency on a link is in range $[L_f(k_l), L_f(k_h)]$, we may be able to regard the link as a band pass filter with cutoff frequencies of $L_f(k_h)$ and $L_f(k_l)$. The frequency gain of a band pass filter for an analog circuit takes the maximum value at the resonance frequency, and it is obtained as the geometric average of the cutoff frequencies, (16). In this sense, the communication frequency may be the resonance of users at two ends.

6. Evaluation and consideration of the results

We created a network model that reflects the relationship between the number of mobile phone users and the volume of traffic on the one hand and the social network structure

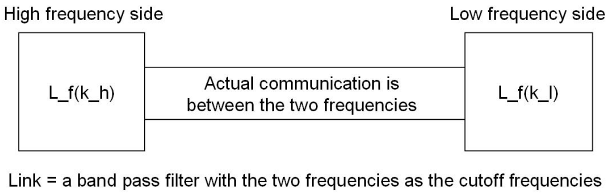

High frequency side
Low frequency side
L_f(k_h)
Actual communication is between the two frequencies
L_f(k_l)
Link = a band pass filter with the two frequencies as the cutoff frequencies

Figure 6. Model for communication frequency on a link using resonance frequency

obtained from the chronological change in the number of SNS users [3] on the other, and simulated information propagation on that network model. In this network model, a node represents a user, and a link represents the information exchange relation between users.

We adopted a simulation model in which the number of users increases only as a result of an existing user inviting a new user as is the case in the large SNS, mixi. We call this model a mixi type model. In this simulation, we examined how the number of users increases over time. In a mixi type model, the entry of a new user is heavily dependent on the structure of the social network of the users. Therefore, differences in the structures of social networks should have a strong effect on the characteristics of the increase in the number of users.

The simulation was executed under the following conditions. The number of nodes was 50,000. At the initial state, only those nodes that have the highest degree hold information. The state of a node having information corresponds to the state that the user is already a member of a mixi type SNS service. Information is propagated from the node that already has information to a node that does not over the link connecting them. This means that no information propagation occurs between nodes that do not have a link connecting them. The propagation of information to a node that does not have information corresponds to the subscription of a new user to a mixi type SNS service. To conduct the evaluation, we caused information propagation at a rate proportional to the communication frequency on a link, and observed how the number of nodes receiving information changes over time. We considered three information propagation models depending on how the communication frequency on a link is determined:

- (k^0 model) : The communication frequency is the same on any link, and information propagation occurs at random.
- (k^2 model) : The communication frequency is determined by the product of the degrees of the two ends of a link ($\sqrt{L_f(k_h) \cdot L_f(k_l)}$).

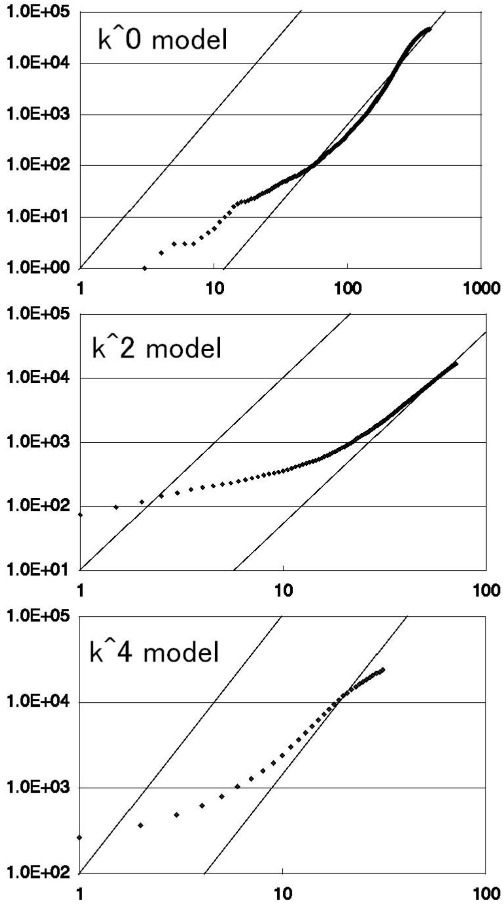


Figure 7. Results of the evaluation of different information propagation models

- (k^4 model) : The communication frequency is determined by the square of the product of the degrees of the two ends of a link ($L_f(k_h) \cdot L_f(k_l)$).

Figure 7 shows, on a double logarithmic chart, how the number of nodes receiving information changes over time for each information propagation model. Figure 8 similarly shows the change in the number of mixi users over time [1]. Lines with a gradient of 3 are included in the charts for comparison. These results reveal that the k^2 model reproduces the actual information propagation well, and produces a result that matches what can be expected from Equation 15.

7. Conclusions

In this paper, we have used the power law observed in the process of computer virus infection, and the power law observed in the communication time of file sharing software,

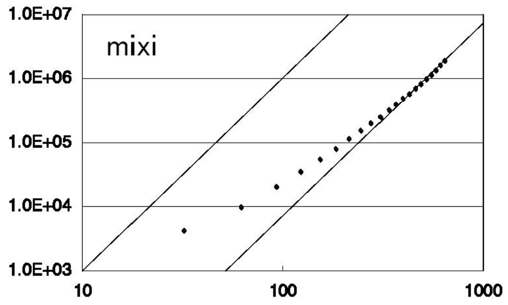


Figure 8. Change in the number of mixi users over time

to examine the frequencies at which people use the network. In addition, using an analogy of the resonance frequency of a band pass filter to the characteristics of network access frequency, we proposed a mode for the communication frequency on a link connecting users. In order to validate the model for communication frequency on a link, we have simulated information propagation on a social network model. We have confirmed that the simulation gives results similar to the change in the number of users of a real SNS. We will verify the structure of a social network using other types of data. Also, based on the proposed information propagation model, we will study application technologies, such as those that can be used in developing marketing strategies.

Acknowledgement

Part of this study was financed by the 2005 Research Grant from the International Communications Foundation (ICF), and a grant from Core Research Grant (S) 18100001 funded by the Ministry of Education, Culture, Sports, Science and Technology.

References

[1] mixi, Inc. http://mixi.co.jp/.

[2] TrendMicroHomepage(Japan). http://www.trendmicro.co.jp/home/.

[3] M.Aida, J.Sasaki. Structural analysis of a social network using the process of the spread of communication services. In *IEICE Technical Report*, number IN2006-41, 2006.

[4] M.Aida, K.Ishibashi, C.Takano, H.Miwa, K.Muranaka, A.Miura. Cluster structures in topology of large-scale social networks revealed by traffic data. *IEEE GLOBECOM 2005*, 2005.

[5] M.Aida, T.Sasaki, C.Takano. Study of the number of users vulnerable to computer viruses. In *IEICE Technical Report*, number TM2005-41, 2005.

[6] S.Ohzahata, K.Kawashima. Pure p2p network size estimation method using network measurements and computer simulation experiments. In *Proc. of the 19th International Teletraffic Congress (ITC19)*, pages 69–78, 2005.