Journal of Theoretical and Applied Information Technology
28th February 2013. Vol. 48 No.3
© 2005 - 2013 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A HUMAN-ORIENTED VIRUS PROPAGATION MODEL IN EMAIL NETWORKS

[1]XIANGHUA LI, [2]CHAO GAO

[1] Assoc. Prof., College of Computer and Information Science, Southwest University, Chongqing, China
[2] Assoc. Prof., College of Computer and Information Science, Southwest University, Chongqing, China
E-mail: [1]li_xianghua@163.com, [2]cgao@swu.edu.cn

ABSTRACT

Our world is currently threatened by digital viruses, such as email viruses and mobile viruses. These viruses are mainly activated by users' operations. Therefore, it's important for us to understand the pattern of user's operational behaviors and estimate the effect of such behaviors on virus propagation. This paper first reveals the statistical characteristics of human behaviors, especially the email-checking intervals of the same user based on the Enron email dataset. After that, we analyze the effect of human operational behaviors and network topologies on virus propagation in a human-oriented virus propagation model. The empirical results from real dataset show that the waiting intervals of each user to check mailbox follow a long-tail distribution. Combining this finding, our experiments accurately describe the process of email-virus propagation. The results show that viruses can fast spread in a network if the email-checking intervals follow a long-tail distribution with a higher power-law exponent. Meanwhile, our results find that the infected nodes with the highest-degree may speed up the virus propagation through analyzing the effects of network structure on virus propagation.

Keywords: *Human Dynamics, Waiting Intervals, Virus Propagation*

1. INTRODUCTION

It is important for us to understand and reveal the dynamic characteristics of human behaviors that have significant scientific and commercial potential [1]. The research about the statistical analysis of the laws of human behaviors from mass data has brought extraordinary and glorious progress in sociology [2].

Currently, the most of research about human behaviors are qualitative description, i.e., the human activities are assumed as a random model on the whole and depicted by the Poisson process. As a typical method to characterize human behaviors, the Poisson process is widely used in many real models to quantify the process of human activities, and depict the statistical regularity on the frequency of certain events in a period of time [3]-[5]. With the application of computer science, more and more statistical results from the log files and database show that the most of human behaviors deviate the Poisson process, i.e., user frequently has a short period of focused activity followed by a long period of inactivity. Through analyzing the logs that recording the information of human activities, some studies have found that when a user engages with certain activities, the waiting intervals of the same user follow a power-law distribution with a long-tail characteristic [6]-[7]. Although some researches have studied the behavioral characteristics of sending email [8]-[10], they didn't address the effect of human behaviors on virus propagation. Meanwhile, the correctness of theory needs more empirical research to verify, especially there are lots of variances among different types of users. This paper further mines the characteristics of human behaviors from the Enron email dataset [11], and analyzes the statistical properties of sending email by the same user. Inspired by these finding, this paper estimates and reveals the effect of human operational behaviors and network topologies on virus propagation.

The research about the effect of non-Poisson characteristics of human actives on collective behaviors (e.g., the virus/rumor propagation in a computer/social network) is a popular topic. There are two typical propagation models: epidemic models based on the mean-field theory (SI, SIS, SIR, etc.) [12], and individual-based models based on the multi-agent simulation [3]. The traditional epidemic models provide a macroscopic understanding of the propagation by some differential equations. However, some assumptions such as full mixing and equiprobable contacts are

© 2005 - 2013 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

unreliable in the real world [13]. Therefore, some microscopic characteristics cannot be observed through these models. In order to overcome these shortfalls, Zou et al. have built an individual-based email model to analyze worm propagation [3]. Some human behaviors, i.e., checking email-boxes and clicking suspected emails, are added into their model in order to examine the effects of human behaviors on worm propagation. The interactive behaviors in Zou's model are characterized by Poisson distribution that deviates from current research about human dynamics [6]-[10]. Therefore, some conclusions in Zou's models are not based on the real situation. In this paper, we improve the accuracy of a human-oriented virus propagation model by combining the empirical studies on human operational behaviors.

Although Vazquez et al. have tried to integrate the sending email intervals that are characterized by probability into virus propagation model [14], the model assumes that a user will be instantly infected after he/she receives a virus email. Therefore, their model doesn't address the effect of user's security awareness on virus propagation. Therefore, there exists a gap between their model and the real-world scenario. In this paper, we construct a human-oriented propagation model and extract activities rules of sending emails among different employees through analyzing communication logs in the Enron email dataset. Based on the above analyses, we simulate the process of virus spreading in both synthetic and benchmark networks in order to observe the effect of human behaviors and network topologies on virus propagation. The numerical results show that viruses fast propagate at the initial stage, and then slowly diffuse if the distribution of email-checking intervals follows a power-law distribution which is similar to the real-world scenario. At the same time, our experiments also explain why some old viruses can also propagate in a network for a long time.

Table 1: The Information of People in the Enron Dataset

Id	Name	Job	Position	Messages	Begin Time	End Time
44	John Arnold	Vice president	-	1587	2000-02-27	2002-01-18
48	Tana Jones	N/A	N/A	4437	1999-05-03	2002-02-08
52	Kay Mann	Employee	-	5100	2000-06-02	2002-05-28
53	Joho Lavorato	CEO	Enron America	1122	2001-01-26	2001-06-08
73	Jeff Dasovich	Employee	Enron government relations executive	6272	1999-12-03	2002-09-22
107	Louise Kitchen	President	Enron Online	1504	1999-05-24	2002-02-06
109	Vince Kaminski	Manager	Risk Management Head	1219	2001-05-15	2002-01-30
122	Sally Beck	Employee	Chief Operating Officer	1596	1999-12-13	2002-02-06
125	Eric	Eric Bass	Trader	1641	1999-12-13	2002-02-07

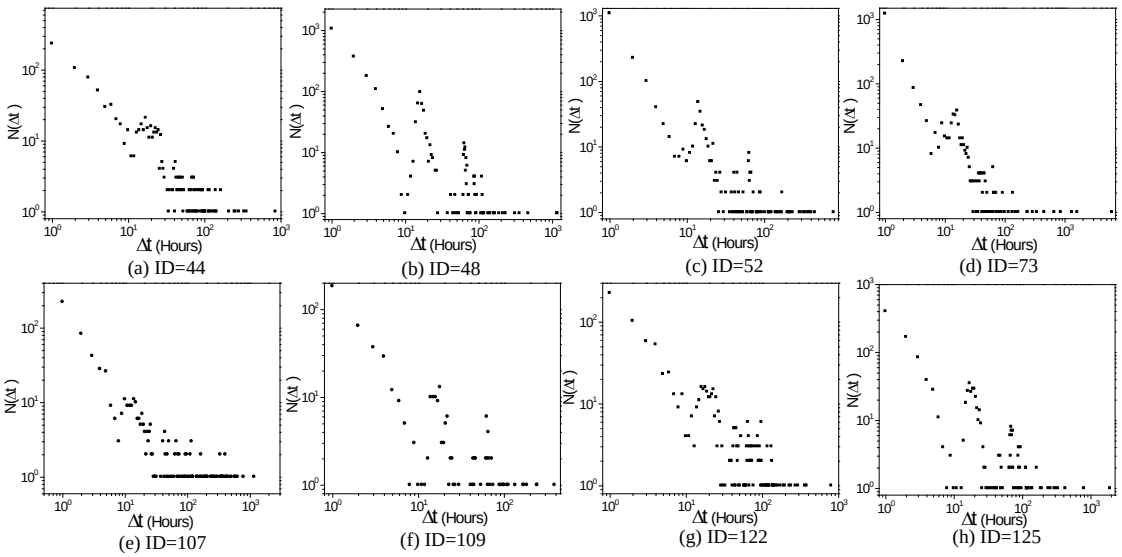


Figure 1: The Inter-event Distribution of Users. The X-scales are Hours (Logarithmic Charts)

Journal of Theoretical and Applied Information Technology
28th February 2013. Vol. 48 No.3
© 2005 - 2013 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

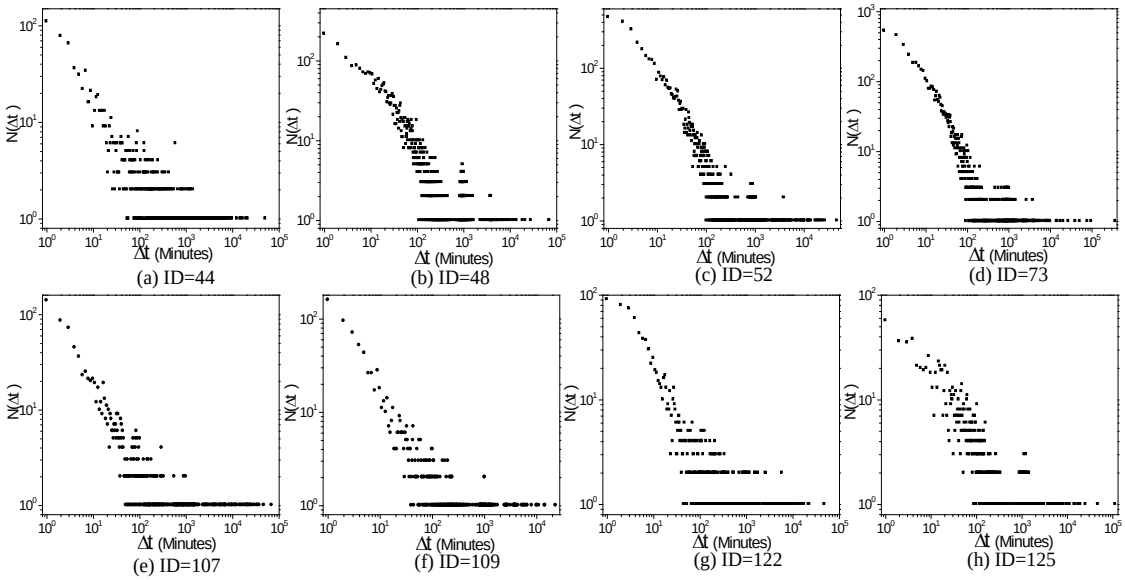


Figure 2: The Inter-event Distribution of Users. The X-scales are Minutes (Logarithmic Charts)

2. THE COLLECTIVE BEHAVIORS IN THE EMAIL COMMUNICATION

The Enron email dataset is released at 2003 by the Federal Energy Regulatory Commission during the investigation. There are lots of versions about the Enron dataset. The version used in our study is built and cleaned by Jitesh Shettyand and Jafar Adibi at USC/ISI[1]. There are 151 employees and 252759 emails stored in the MySql database [11]. Table 1 only presents some typical employees' information because of paper limitation. The interval distributions of sending emails by the same user are respectively measured by Hour and Minute. Figs. 1-2 show that the waiting intervals of one person follow a long-tail distribution. Comparing with Fig. 1, Fig. 2 emerges the power-law distribution with the increment of nodes. At the same time, there is a peak at $\triangle$t=16 which is the interval between when people leave work and when they return to their offices. Based on the fitting curves in Figs.1-2, the exponent of waiting interval is near 1.3, *i.e.*, $\alpha \approx 1.3 \pm 0.5$.

Based on the above analyses, we find that the waiting intervals follow a power-law distribution. But we cannot assert that all users' waiting time follows a power-law distribution. However, we can assert that the distribution of waiting intervals has a long-tail characteristic. Meanwhile, we cannot measure the email-checking intervals because the login time is not recorded in the Enron dataset. Combing the research about the human behaviors in the Web browsing [15] and the effect of non-Poisson activities on the propagation in the CCNR group [14], we find that there are similar between the distribution of email-checking intervals and sending emails intervals. In the next section, we use a power-law distribution to characterize the email-checking behaviors in order to analzye the effect of human behaviors on virus propagation.

3. SIMULATION RESULTS

In this section, we perform several experiments to uncover the effect of some factors on virus propagation. Sec. 3.1 presents the main parameters setting in our experiments. Sec. 3.2 introduces some networks as used in this paper. Sec. 3.3 evaluates the effects of users' operational patterns and network topologies on virus propagation.

3.1 Propagation Model

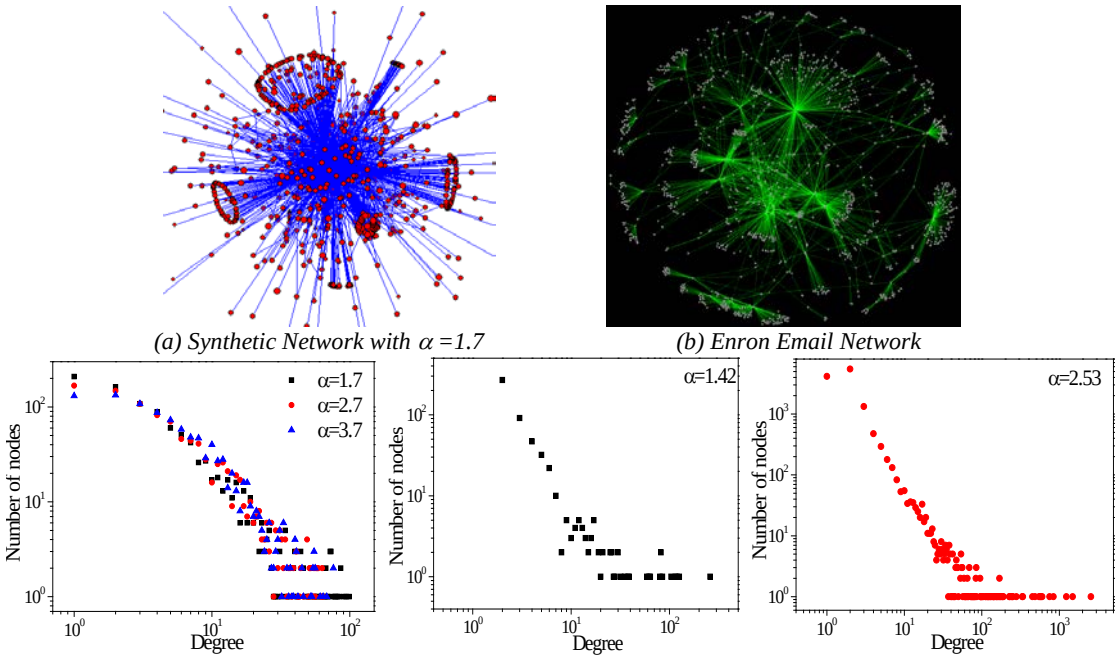
Based on traditional epidemic models, if there is an edge between two nodes in a network, the virus can propagate from an infected node to healthy neighbors at the next time. However, the process of virus propagation, depicted by traditional epidemic models, is different from real digital viruses in the real world. In order to accurately depict the propagation of email worm, Zou et al. have built an individual-based email model [3], in which the viruses are triggered by human behaviors, rather than the contact probability. That is to say, the user will be infected only if he has checked his email-box and clicked the email with virus attachment. Therefore, virus propagation is mainly based on two user factors: *the email-checking intervals* and

[1] http://sgi.nu/enron/corpora.php

Journal of Theoretical and Applied Information Technology
28th February 2013. Vol. 48 No.3
© 2005 - 2013 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the clicking email probabilities. The email-checking time and clicking email probability of user i is denoted as T_i and P_i (i=1,2,…,N, N denotes the total number of users in a network), respectively. T_i is determined by user own habits. And the P_i is determined by user's security awareness for the risk of viruses. If a user clicks an infected email, the node is infected and will automatically send viruses to all friends in its hit-list. Meanwhile, we assume that the user will delete suspected emails if a user does not click the virus email.

The checking email interval of user i in Zou's model [3] is depicted by Poisson distribution, i.e., $T_i(\tau) \sim \lambda e^{-\lambda\tau}$. However, we observe that the email-checking intervals of one user follows a power-low distribution based on the analyses in Sec.2, i.e., $T_i(\tau) \sim \tau^{-\alpha}$. In order to observe and quantitatively analyze the effect of email-checking intervals on virus propagation, the distribution of clicking email probabilities (p_i) in our model is consistent with Zou's model, i.e., the security awareness among different users in a network follows a normal distribution, $p_i \sim N(0.5, 0.3^2)$. If users have higher security awareness, they would not be infected even if they receive an infected message.



(a) Synthetic Network with α =1.7 *(b) Enron Email Network*

(c) Three Synthetic Networks *(d) Enron Email Network* *(e) AS Network*

Figure 3: (a) (b) Network Structures and (c-e) Degree Distribution of Synthetic and Benchmark Networks

3.2 The Structures of Networks

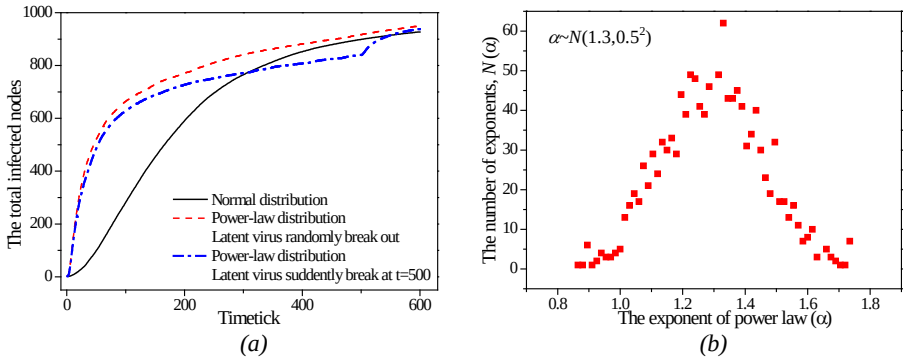
Many studies on social networks have shown that email networks present a phenomenon of scale-free [16], where nodes' degrees follow a power-law distribution [17]. Therefore, this paper utilizes synthetic and benchmark networks, both of which have a long-tail statistical distribution, to simulate virus propagation. Based on GLP [18], three synthetic networks are generated where the power-law exponent can be tuned. The three synthetic networks all have 1000 nodes with the power-exponent α =1.7, 2.7 and 3.7, respectively. Besides the synthetic networks, we also extract two publicly available benchmark networks, i.e., Enron email network and autonomous system network (AS network). The Enron email network is built based on communization logs in the Enron dataset. There are 1238 nodes and 2106 edges in such network that includes both interior employees and exterior users. The AS network is created by Vaishnavi Krishnamurthy[2]. There are 12741 nodes and 26888 edges. The topological snapshots and degree distribution of these networks are shown in Fig.3.

Initially, two nodes are selected randomly from a network as infected nodes in order to simulate a multiple-seed attack that often occurs in the real world. We mainly measure the final infected nodes in a network after the whole system runs 600 steps. Since the process of email worm propagation is stochastic, all experimental results are average

[2] http://www.cs.ucr.edu/%7Evkrish/

Journal of Theoretical and Applied Information Technology
28th February 2013. Vol. 48 No.3
© 2005 - 2013 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

values over 100 simulation runs. The more details about simulation process are shown in Ref. [3]



(a) (b)

Figure 4: (a) Processes of Virus Propagation in the Enron Email Network. (b) The Power-law Exponent among Users.

3.3 The Effect of Human Behaviors on Virus Propagation

3.3.1 The effect of checking intervals

In order to accurately depict the process of virus propagation, the email-checking intervals of a user are simulated by a power-law distribution, rather than a normal distribution, based on our previous statistical results. Fig. 4 shows that the viruses propagate quickly in a network if the email-checking intervals follow a power-law distribution. The results are more accordance with the observed trends in the real computer network [19], i.e., the viruses at the initial stage are explosive growth and then latent for a long time in order to be activated by user again. On the other hand, from the perspective of human dynamics, the reason that causes the above results is that users frequently have a short period of focused activity followed by a long period of inactivity [20]. In other words, there are very long periods of inactivity which are separated by bursts and intensive activity. Therefore, although some old viruses are killed by the anti-software, they could also intermittently outbreak in the network so far. That is because some viruses are hidden in some inactive users and not be found by anti-software [20]. When those inactive users are re-activated, the virus will propagate again.

Furthermore, we estimate the effect of different distributions of user's email-checking intervals on virus propagation in both synthetic and benchmark networks. Since the ability of users follows a normal distribution based on the empirical studies in [5], Fig. 5 plots three normal distributions of power-law exponents (α) among users, which determine users' checking email intervals. The email-checking intervals of each user will be generated based on their own power-law exponent.

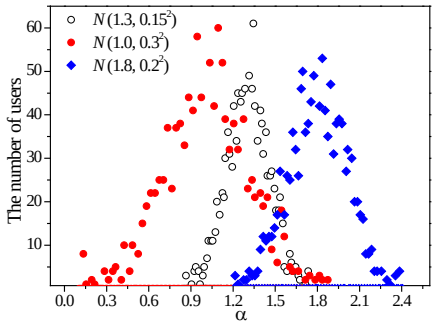


Figure 5: The Distribution of the Power-law Exponent among Different Users.

Figure 6 illustrates the simulation results in different networks. These results shows that the more frequent users check their email-boxes (i.e., the power-law exponent is higher), the more infected the nodes are. Meanwhile, the nodes with the highest-degree will enlarge the scale of virus propagation through comparing the final number of infected nodes in three synthetic networks (i.e., Fig.6(c)(d)(e)). That is because the higher power-law exponent a network has, the more highest-degree nodes the network will have.

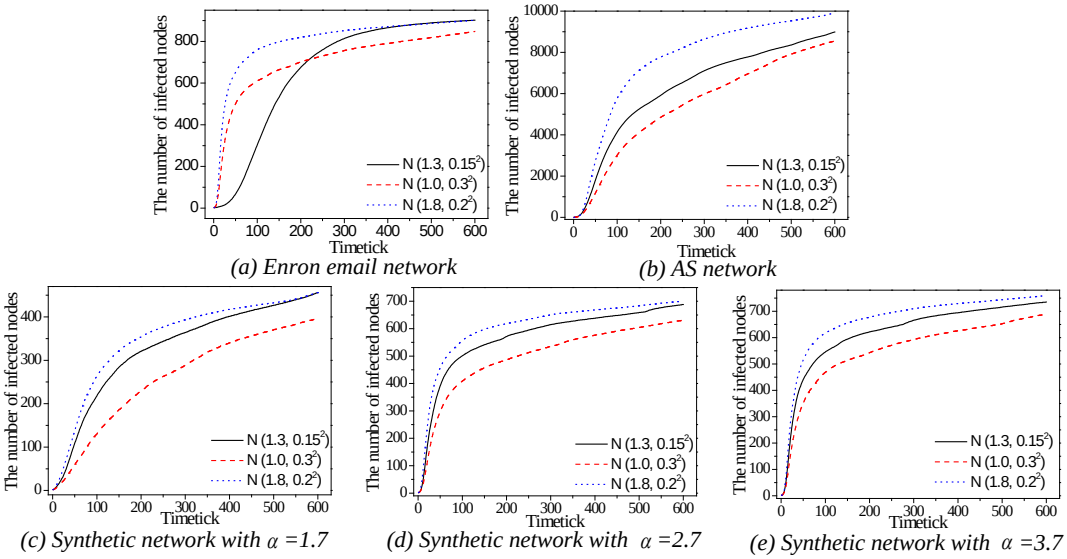
3.3.2 The effect of security awareness

In a human-oriented model, virus propagation is activated by human operations, such as clicking on a suspicious email. Therefore, user's own security awareness plays an important role on virus propagation. If a user has enough knowledge background about viruses (i.e., the higher security awareness), the user will have a lower probability (i.e., p_i) to click on a suspicious message.

Figure 7 shows the effect of users' security awareness on virus propagation where the power-law exponents of users' email-checking intervals follows $N(1.3, 0.15^2)$ as shown in Fig.5. The results show that if users' security awareness is higher (i.e., the mean value of p_i is smaller), the propagation scope will become smaller (i.e., the total number of

© 2005 - 2013 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

infected nodes are smaller). Therefore, it's important to improve users' risk awareness about viruses through public security education (i.e., the public campaigns on the risks of viruses to users) or warning messages, which has been already used for restraining mobile virus propagation [21].



(a) Enron email network *(b) AS network*

(c) Synthetic network with α =1.7 *(d) Synthetic network with α =2.7* *(e) Synthetic network with α =3.7*

Figure 6: The Processes of Virus Propagation with Different Email-Checking Intervals

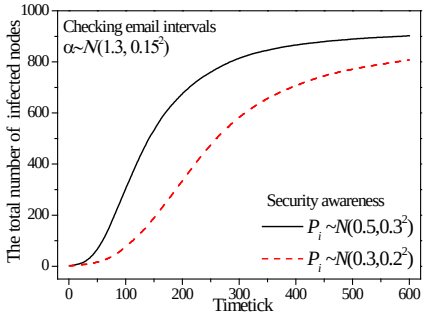


Figure 7: The Effect of Individual Security Awareness (p_i) on Virus Propagation

4. CONCLUSION

This paper starts with the human dynamics and analyzes the effect of human operations on virus propagation in a human-oriented email model. Based on the email dataset from Enron Corporation, we have found that the sending email intervals among different users follow a power-law distribution and the exponents of different users are around 1.3. Inspired by this finding, we have simulated virus propagation in a human-oriented virus propagation model. The results have shown that viruses can fast spread in a network if email-checking intervals follow a power-law distribution. In such situation, the viruses are explosive growth at the initial stage and then slow growth. That is because viruses will stay at latent state and await activation by users. Meanwhile, we can effectively restrain virus propagation through improving the public campaigns on the risks of viruses to users.

ACKNOWLEDGEMENT

This work is supported by the Natural Science Foundation Project of CQ CSTC (No.cstc2012jjA40013, cstc2012jjB40012), the Specialized Research Fund for the Doctoral Program of Higher Education (20120182120016), the Fundamental Research Funds for the Central Universities (XDJK2012B016, XDJK2012C018), and the PhD fund of Southwest University (SWU111024, SWU111025).

REFRENCES:

[1] T. Zhou, H.A.T. Kiet, B.J. Kim, B.H. Wang and P. Holme, "Role of activity in human dynamics", *EPL*, Vol. 82, No. 2, 2008, pp. 28002.

[2] D.J. Watts, "A twenty-first century science", *Nature*, Vol. 445, No. 7127, 2007, pp. 489.

[3] C.C. Zou, D. Towsley and W. Gong, "Modeling and Simulation study of the propagation and defense of Internet E-mail worms", *IEEE Transactions on Dependable and Secure Computing*. Vol. 4, No. 2, 2007, pp. 105-118.

© 2005 - 2013 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[4] J.M. Liu, S.W. Zhang and J. Yang, "Characterizing web usage regularities with information foraging agents", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 5, 2004, pp. 566-584.

[5] F.A. Haight, "Handbook of the Poisson Distribution", *New York: John Wiley and Sons*, 1967.

[6] A. Vazquez, J.G. Oliveira, Z.Dezso, K.-I.Goh, I. Kondor and A.-L. Barabasi, "Modeling bursts and heavy tails in human dynamics", Physical Review E, Vol. 73, No. 3, 2006, pp. 036127.

[7] T. Zhou, X. P. Han and B. H. Wang, "Towards the understanding of human dynamics", *Http://arxiv.org/pdf/0801.1389*.

[8] J.P. Eckmann, E. Moses and D. Sergi, "Entropy of dialogues creates coherent structure in email traffic", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 101, No.40, 2004, pp. 14333-14337.

[9] A. Johansen, "Probing human response times", *Physica A*, Vol. 338, No. 1-2, 2004, pp.286-291.

[10] A.L. Barabasi, "The origin of bursts and heavy tails in human dynamics", *Nature*, Vol. 435, No. 7039, 2005, pp.207-211.

[11] J. Shetty and J. Adibi, "The Enron email dataset database schema and brief statistical report", 2004, http://www.isi.edu/~adibi/Enron /Enron_Datas et_Report.pdf.

[12] A.L. Lloyd and R.M. May, "How viruses spread among computers and people", *Science*, Vol. 292, No. 5520, 2001, pp. 1316-1317.

[13] M.E.J. Newman, "The spread of epidemic disease on network", *Physical Review E*, Vol. 66, No. 1, 2002, pp. 016128.

[14] A. Vazquez, B. Racz, A. Lukacs and A.L. Barabasi, "Impact of non-poissonian activity patterns on spreading process", *Physical Review Letters*, Vol. 98, No. 15, 2007, pp. 158702.

[15] Z. Dezso, E. Almaas, A. Lukacs, B.Racz, I Szakadat and A.-L, Barabasi, "Dyanmics of information access on the web", *Physical Review E*, Vol. 73, No. 6, 2006, pp.066132.

[16] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez and D.-U. Hwang, "Complex networks: Structure and dynamics", *Physics Reports*, 2006, Vol. 424, No. 4-5, pp. 175-308.

[17] L.F. Costa, O.N.O. Jr, G. Travieso, et al., "Analyzing and modeling real-world phenomena with complex networks: A survey of applications", *Advances in Physics*, 2011, Vol. 60, No. 3, pp. 329-417.

[18] T. Bu and D. Towsley, "On distinguishing between internet power law topology generators", *Proceedings of the Twenty First Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM02)*, 2002, pp. 638-647.

[19] D. Moore, C.Shannon and J. Brown, "Code-red: A case study on the spread and victims of an internet worm", *Proceedings of the ACM SIGCOMM/USENIX Internet Measurement Workshop*, 2002, pp. 273-284.

[20] C. Gao, J.M. Liu and N. Zhong, "Network immunization and virus propagation in email network: Experimental evaluation and analysis", *Knowledge and Information System*, Vol. 27, No. 2, 2011, pp.253-279.

[21] E.V. Ruitenbeek and F. Stevens, "Quantifying the effectiveness of mobile phone virus response mechanisms", *Proceedings of 37th Annual IEEEIFIP International Conference on Dependable Systems and Networks (DSN07)*, 2007, pp. 790-800.