Non-Poisson Processes of Email Virus Propagation

Miroslav Mirchev[1] and Ljupco Kocarev[1, 2, 3]

[1] Faculty of Electrical Engineering and Information Technologies, Skopje, Macedonia
[2] Macedonian Academy for Sciences and Arts, Skopje, Macedonia
[3]University of California, San Diego, CA
{[1]miroslavm, [2]lkocarev}@feit.ukim.edu.mk

Abstract. Email viruses are one of the main security problems in the Internet. In order to stop a computer virus outbreak, we need to understand email interactions between individuals. Most of the spreading models assume that users interact uniformly in time following a Poisson process, but recent measurements have shown that the intercontact time follows heavy-tailed distribution. The non-Poisson nature of contact dynamics results in prevalence decay times significantly larger than predicted by standard Poisson process based models. Email viruses spread over a logical network defined by email address books. The topology of this network plays important role in the spreading dynamics. Recent observations suggest that node degrees in email networks are heavy-tailed distributed and can be modeled as power law network. We propose an email virus propagation model that considers both heavy-tailed intercontact time distribution, and heavy-tailed topology of email networks.

Keywords: Computer viruses, Dynamical systems, Complex networks.

1 Introduction

The concept of a computer virus is relatively old in the young and expanding field of information security. It was first developed by Cohen in [1, 2], and it still is an active research area. Computer viruses still accounts for a significant share of the financial losses that large organizations suffer for computer security problems, and it is expected that future viruses will be even more hostile.

According to The WildList Organization International [3] there were 70 widespread computer viruses in July 1993, and that number have increased up to 953 in July 2009 (Fig. 1). With the proliferation of broadband ‘‘always on’’ connections, file downloads, instant messaging, Bluetooth-enabled mobile devices, and other communications technologies, the mechanisms used by viruses to spread have evolved as well [4, 5]. Still, many viruses continue to spread through email. Indeed, according to the Virus Bulletin [6], the email viruses (email worms) still accounts for large share of the virus prevalence today.

Email viruses spread via infected email messages. The virus may be in an email attachment or the email may contain a link to an infected website. In the first case the virus will be activated when the user clicks on the attachment and in the second case when the user clicks on the link leading to the infected site.

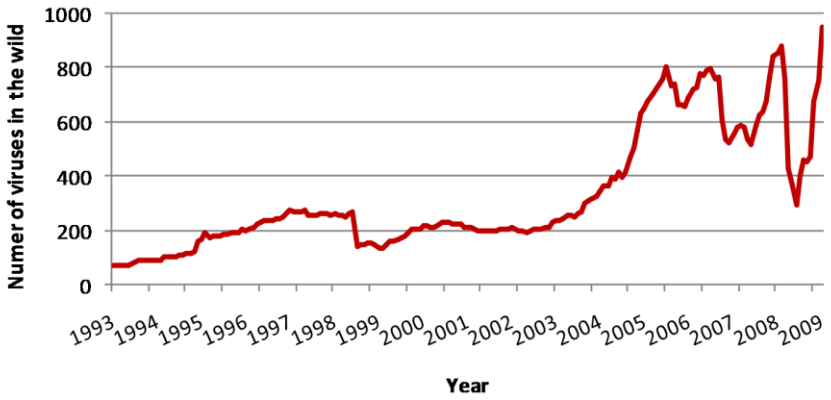

Fig. 1. Number of viruses in-the-wild according to The WildList Organization International. This is a cooperative listing of viruses reported as being in the wild by virus information professionals. The list includes viruses reported by multiple participants, which appear to be non-regional in nature. The WildList is currently being used as the basis for in-the-wild virus testing and certification of anti-virus products by the ICSA, Virus Bulletin and Secure Computing.

When an email virus infects a machine, it sends an infected email to all addresses in the computer's email address book. This self-broadcast mechanism allows for the virus's rapid reproduction and spread, explaining why email viruses continue to be one of the main security threats. While some email viruses used only email to propagate (e.g. Melissa), most email viruses can also use other mechanisms to propagate in order to increase their spreading speed (e.g. W32/Sircam, Love Letter).

Although virus spreading through email is an old technique, it is still effective and is widely used by current viruses. It is attractive to virus writers, because it doesn't require any security holes in computer operating systems or software, almost everyone uses email, many users have little knowledge of email viruses and trust most email they receive (especially email from friends) and email is private property so correspondent laws or policies are required to permit checking email content.

Email viruses usually spread by connecting to SMTP servers using a library coded into the virus or by using local email client services. Viruses collect email addresses from victim computers, in order to spread further, by: scanning the local address book, scanning files with appropriate extensions for email address and sending copies of itself to all mail in the user's mailbox. Some viruses even construct new email addresses with common domain names.

In order to eradicate viruses, as well as to control and limit the impact of an outbreak, we need to have a detailed and quantitative understanding of the spreading dynamics and environment. In most email virus models have been assumed that the contact process between individuals follows Poisson statistics, and, the time between two consecutive contacts is predicted to follow an exponential distribution [7-13]. Therefore, reports of new infections should decay exponentially with a decay time of about a day, or at most a few days [7–11]. In contrast, prevalence records indicate that new infections are still reported years after the release of antiviruses [4, 7, 14], and their decay time is in the vicinity of years, 2–3 orders of magnitude larger than the Poisson process predicted decay times. This discrepancy is rooted in the failure of the Poisson

approximation for the interevent time distribution. Indeed, recent studies of email exchange records between have shown that the probability density function of the time interval between two consecutive emails sent by the same user is well approximated by a fat tailed distribution [15-19]. In [20] the authors prove that this deviation from the Poisson process has a strong impact on the email virus's spread, offering a coherent explanation of the anomalously long prevalence times observed for email viruses.

The email network is determined by users' email address books, and its topology plays important role in the spreading dynamics. In [21] the authors use Yahoo email groups to study the email network topology. Although the topology of email groups is not the complete email network topology, they use it to figure out what the topology might be like. Their findings suggest that the email groups are heavy-tailed distributed, so it is reasonable to believe that email network is also heavy-tailed distributed. The problem of virus spreading in networks with heavy-tailed distribution has been studied in [7, 10, 21].

An epidemic threshold is a critical state beyond which infections become endemic. In [22, 23], the authors have presented a model that predicts the epidemic threshold of a network with a single parameter, namely, the largest eigenvalue of the adjacency matrix of the network.

In this paper, we propose an email virus propagation model with nonlinear dynamical system, which considers both heavy-tailed intercontact time distribution and heavy-tailed topology of email networks. We use this model to reveal new form of the epidemic threshold condition.

The rest of the paper is organized as follows. In Section 2, we define the network model, and analyze the email network topology and communication patterns. After that in Section 3, we propose a discrete stochastic model for Non-Poisson virus propagation in email network with power law topology, and we introduce the epidemic threshold. Simulation results and analyses are given in Section 4 and Section 5 concludes the paper.

2 Email Network Model

Let $G = (V, E)$ be a connected, undirected graph with N nodes, which represent the email users, and m edges, which represent the contacts between the users. Every user has an address book in which he has all the users he contacts with. These address books are represented with the adjacency matrix $\mathbf{A}$ of the graph G, i.e., $a_{ij} = 1$ if $(i, j) \in E$ (user i have user j in his address book) and $a_{ij} = 0$ otherwise.

At time k, each node i can be in one of two possible states: **S** (susceptible) or **I** (infected). The state of the node is indicated by a status vector which contains a single 1 in the position corresponding to the present status, and 0 in the other position:

$$\mathbf{s}_i(k) = [s_i^S(k) \;\; s_i^I(k)]^T \tag{1}$$

and let

$$\mathbf{p}_i(k) = [p_i^S(k) \;\; p_i^I(k)]^T \tag{2}$$

be the probability mass function of node i at time k. For every node i it states the probability of being in each of the possible states at time k.

The network topology is determined by the adjacency matrix **A**, i.e. by the users' email address books. The size of a user's email address book is the degree of the corresponding node in the network graph. Since email address books are private property, it is hard to find data to tell us what the exact email topology is like.

We use the Enron email dataset, described in [24] and available at [25], to study the email network topology. This set of email messages was made public during the legal investigation concerning the Enron Corporation. It is the only publicly available email dataset and consists of 158 users (mostly senior management) and 200,399 messages (from which 9728 are between employees). The dataset contains messages from a period of almost three years. On Fig. 2 the degree distribution of the users' address books from this dataset is shown. We see that the power law $P(k) \sim k^{-3.5}$ approximates well a substantial part of the users' degree distribution, but fails to approximate well for small and large degree values. This is mostly due to the fact that the number of users in the dataset is small, but nevertheless it gives us an insight into the real email network topology. Because of this degree distribution, and the findings from [21], it is best if we model the email network as a power law network.

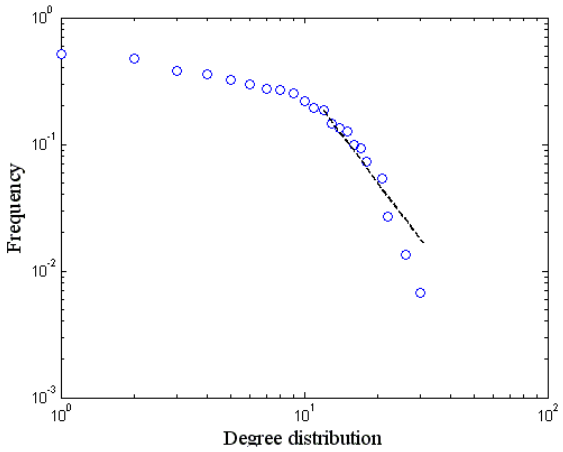

Fig. 2. Degree distribution of the address books from the Enron email dataset. The dashed line represents the power law decay $P(k) \sim k^{-3.5}$.

The contact dynamics responsible for the spread of email viruses is driven by the email communication and the usage patterns of individuals. To characterize these patterns we also use the Enron email dataset. We use only the messages between the employees (9728 messages), and it is sufficient for accurate analysis. Let τ (interevent time) denote the time between two consecutive emails sent by a single user. The dis-

tribution of the aggregate interevent of all the users approximately follows a power law with exponent $\alpha \approx 2.4$ and a cut-off at large τ values (Fig. 3).

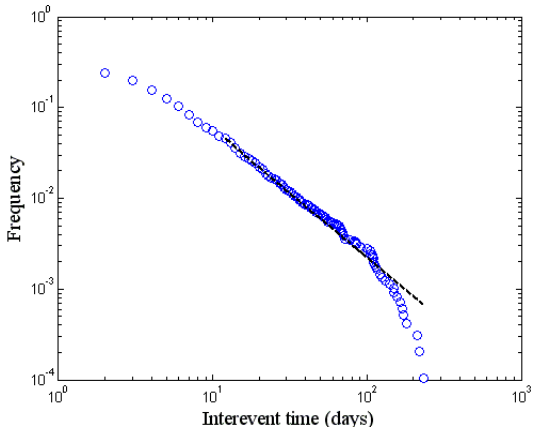

Fig. 3. Distribution of the interevent time between two consecutive emails sent by an email user in the Enron dataset. We aggregate the interevent times of all users (the distribution for single users is similar). The dashed line represents the power law decay $P(k) \sim k^{-2.4}$.

3 Email Virus Propagation Model

The spreading dynamics is jointly determined by the email activity patterns and the topology of the corresponding email communication network. We propose a discrete stochastic model for virus propagation in email network with power law topology and communication pattern with heavy-tailed interevent time distribution.

The Barabasi-Albert model [26] is used for generating email networks with power law topology, which is one of several proposed models that generate power law networks. The model is using a preferential attachment mechanism and generates network which has degree distribution with the power law form $P(k) \sim k^{-3}$.

In order to compare power law networks against random networks, we use the Erdos-Renyi model [27] for generating random networks. In this model, a graph $G(N, p)$ is constructed by connecting N nodes randomly. Each edge is included in the graph with probability p, with the presence or absence of any two distinct edges in the graph being independent.

When an email user have received message with a virus attachment by some of his contacts, he may discard the message (if he suspects the email or detects the email virus by using anti-virus software) or open the virus attachment if unaware of it. When the virus attachment is opened, the virus immediately infects the user and sends out virus email to all email addresses on this user's email address book. Different users open virus attachments with different probabilities, depending on their computer security knowledge. We assume that the probability that an email user opens the infected attachment, after he has received some infected message is constant and denote it with

β. The infected user will not send out virus email again unless the user receives another copy of the email virus and opens the attachment again.

It takes time before a recipient receives a virus email sent out by an infected user, but the email transmission time is usually much smaller comparing to a user's email checking time. Thus in our model we ignore the email transmission time. In most cases received emails are responded to in the next email activity burst [15, 17], and viruses are acting when emails are read, approximately the same time when the next bunch of emails are written. According to this email users' activity can be represented as follows. Let $b_j(k)$ represent users' j activity at time k. If user j is active at time k $b_j(k) = 1$, otherwise $b_j(k) = 0$. We assume that a user reads all his emails at the moment he is active.

We model email users activity by using chaotic-maps. This method is used in [28, 29] for modeling packet traffic. The following map is convenient for our purposes:

$$x_j(k+1) = \begin{cases} \dfrac{x_j(k)}{(1-c_1 x_j(k)^{m_1-1})^{\frac{1}{m_1-1}}}, & \text{if } x_j(k) < d \\ 1 - \dfrac{1-x_j(k)}{(1-c_2(1-x_j(k))^{m_2-1})^{\frac{1}{m_2-1}}}, & \text{if } x_j(k) \geq d \end{cases} \tag{3}$$

where

$$c_1 = \frac{1-d^{m_1-1}}{d^{m_1-1}} \tag{4}$$

$$c_2 = \frac{1-(1-d)^{m_2-1}}{(1-d)^{m_2-1}}, \tag{5}$$

and $d \in [0, 1]$. At each time k, the value of $x_j(k)$ is evaluated for each user j, and then:

$$b_j(k) = \begin{cases} 0, & \text{if } x_j(k) < d \\ 1, & \text{if } x_j(k) \geq d \end{cases} \tag{6}$$

We choose this chaotic map, because for values of m_1 and/or m_2 in the range (3/2, 2) the map generates interevent times that have heavy tailed distribution. More precisely for d=0.7, $m_1 = 1.53$ and $m_2 = 1.96$ the distribution approximately follows a power law with exponent $\alpha \approx 2.4$ and a cut-off at large τ values, very similar to the true interevent time distribution (this can be achieved with other values also).

At the beginning ($k = 0$) there is a small number of initially infected users. Let $V(k)$ denote the infected inbox matrix, where $v_{ij}(k) = 1$, if user j have unread infected email message from user i at time k, and otherwise $v_{ij}(k) = 0$. At time $k = 0$, $v_{ij}(0) = 1$, if user j have initially infected user i in his address book, and otherwise $v_{ij}(0) = 0$. At each time k:

$$v_{ij}(k+1) = (a_{ij}h_i(k)) + v_{ij}(k)(1-h_i(k)))(1-b_j(k)) \tag{7}$$

$$h_{ij}(k+1) = \begin{cases} 1, & \text{if } s_i^I(k) = 0 \text{ and } s_i^I(k+1) = 1 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

Previously we assumed that a user reads all his emails at the moment he is active. So if user j is active at time k ($b_j(k) = 1$), all the messages from the infected inbox matrix V should be removed, $v_{ij}(k+1) = 0$ for all i.

We introduce another parameter δ, which represents the curing probability. After some user gets infected, he may use some means (such as virus removal tool) to remove the virus from his computer. As with β we assume constant curing probability among users. Having defined all this, the equations describing the evolution of our email virus propagation model are:

$$\begin{aligned} p_i^S(k+1) &= b_i s_i^S(k)(1-f_i(k)) + (1-b_i)s_i^S(k) + s_i^I\delta \\ p_i^I(k+1) &= b_i s_i^S(k) f_i(k) + s_i^I(k)(1-\delta) \\ s_i^T(k+1) &= Multirealize[\mathbf{p}_i^T(k+1)] \end{aligned} \tag{9}$$

where *Multirealize*[.] performs a random realization for the probability distribution given with $\mathbf{p}_i^T(k+1)$, and:

$$f_i(k) = 1 - \prod_{j=1}^{N}(1-\beta v_{ji}(k)) \tag{10}$$

4 Simulations and Analyses

For our simulations, we use email networks with 1000 nodes representing the email users and 3000 links representing the users' address books. First, we compare the spreading of email viruses in power law and random (Erdos-Renyi) network, by using both Poisson process approximation and true interevent distribution. For this simulation we use $\delta = 0$, because we are interested in the spreading dynamics, i.e. the number of new infections, instead of the total number of infected users. The other parame-

ter values are d=0.7, $m_1 = 1.53$, $m_2 = 1.96$ and $\beta = 0.5$. From Fig. 4 we see that the spreading process in the power law email network evolves more rapidly than in random network, i.e. the number of new infections at the beginning is higher. If we compare the different interevent distributions, we see that the Poisson process approximation evolves much faster and the spreading process ends in one order of magnitude faster than in the true interevent time distribution. The number of new infections in power law networks, after the initial period, slightly deviates from exponential decay, while in random networks the decay is clearly exponential.

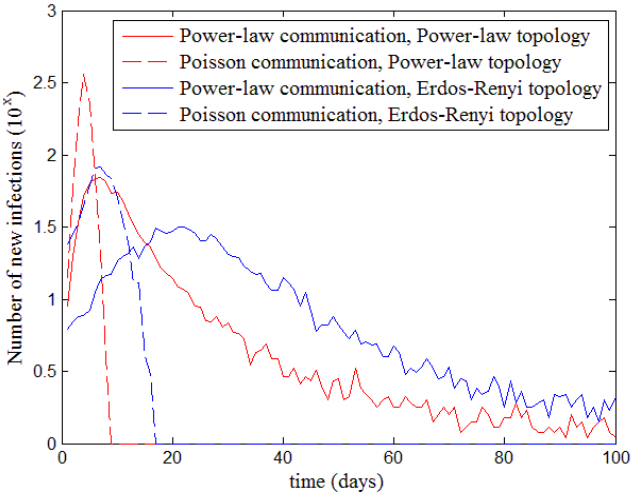

Fig. 4. Average number of new infections in Power-law network (red) and Erdos-Renyi (blue). After the initial period, the lines correspond to an exponential decay predicted by the Poisson process approximation (dash lines) and the true interevent distribution (solid lines).

Predicting the epidemic threshold condition is an important part of a virus propagation model. In [22, 23] the authors predict the epidemic threshold with a single parameter $\lambda_{1,\mathbf{A}}$, the largest eigenvalue of the adjacency matrix **A** of the network. They prove that if an epidemic dies out, then it is necessarily true that:

$$\frac{\beta}{\delta} < \frac{1}{\lambda_{1,\mathbf{A}}}. \tag{11}$$

The epidemic threshold in power law networks is zero [22], so we make the epidemic threshold analysis on random networks. We analyze the dependencies between the parameters, β, δ, $\lambda_{1,\mathbf{A}}$ and d at their threshold values, i.e. the values for which the system moves from a state where the virus prevails, to a state where the virus diminishes). The parameter d captures the characteristics of the communication pattern. We see (Fig. 5) that as in [22, 23] β and δ have linear dependency with $\lambda_{1,\mathbf{A}}$, while the threshold value of d exponentially increases, as $\lambda_{1,\mathbf{A}}$ increases. According to this, the

epidemic threshold condition would have the form given in (12), which captures the essence of both network topology and communication patterns.

$$\frac{\beta}{\delta} < \frac{d^x}{\lambda_{1,\mathbf{A}}} \tag{12}$$

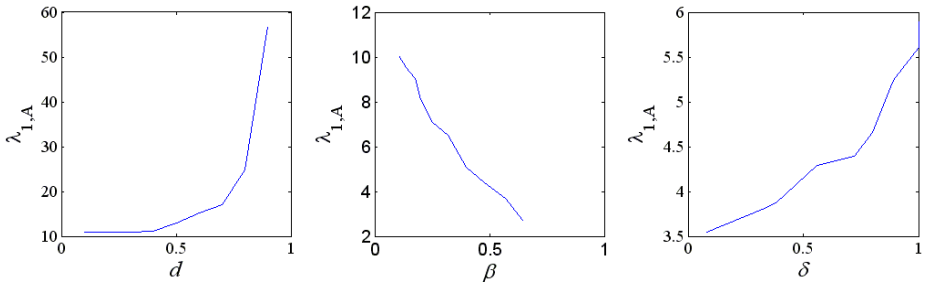

Fig. 5. The dependencies of the parameters β, δ, $\lambda_{1,\mathbf{A}}$ and d at the epidemic threshold.

5 Conclusion

In this paper we analyzed the email network topology and the email communication patterns. We proposed a model for virus propagation in email network with power law topology and communication pattern with heavy-tailed interevent time distribution. The analysis showed that the prevalence time for true interevent time distribution is much longer than predicted by standard Poisson based models, which is coincident with real data. Although the number of new infections exponentially decays in random networks, for email networks it slightly deviates from straight exponential decay.

The epidemic threshold analysis has revealed a new form of the condition under which an epidemic diminishes, which captures the essence of both network topology and communication patterns. This form will be further analyzed.

References

1. Cohen, F.: Computer Viruses. PhD Thesis, University of Southern California (1985)
2. Cohen, F.: Computer viruses – theory and experiments. Computers & Security 6(1):22--35 (1987)
3. The WildList Organization International, http://www.wildlist.org
4. Wang, P., González, M.C., Hidalgo, C.A., Barabási, A.-L.: Understanding the Spreading Patterns of Mobile Phone Viruses. Science 324, 1071 --1076 (2009)
5. Hu, H, Myers, S., Colizza, V., Vespignani, A.: WiFi networks and malware epidemiology. Proceedings of the National Academy of Sciences 106, 1318--1323 (2009)

6. Virus Bulletin, http://www.virusbtn.com
7. Pastor-Satorras, R., Vespignani, A.: Epidemic Spreading in Scale-Free Networks. Physical Review Letters 86(14):3200--3203 (2001)
8. Meyers, L.A., Pourbohloul, B., Newman, M.E.J., Skowronski, D.M., Brunham, R.C.: Network theory and SARS: Predicting outbreak diversity. Journal of Theoretical Biology 232: 71--81 (2005)
9. Moreno, Y., Pastor-Satorras, R., Vespignani, A.: Epidemic outbreaks in complex heterogeneous networks. European Physical Journal 26:521--529 (2002)
10. Barthe´lemy, M., Barrat, A., Pastor-Satorras, R., Vespignani, A.:Velocity and Hierarchical Spread of Epidemic Outbreaks in Scale-Free Networks. Physical Review Letters 92:178701 (2004)
11. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.U.: Complex Networks: Structure and Dynamics. Physics Report 424:175-308 (2006)
12. Moreno, Y., Nevokee, M., Pacheco, A.F.: Dynamics of rumor spreading in complex networks. Physical Review E 69, 066130 (2004)
13. Nekovee, M., Moreno, Y., Bianconi, G., Marsili, M.: Theory of rumor spreading in complex social networks. Physica A 374(1):457--470 (2007)
14. Pastor-Satorras, R., Vespignani, A.: Evolution and Structure of the Internet: A Statistical Physics Approach. Cambridge University, Cambridge (2004)
15. Eckmann, J.P., Moses, E., Sergi, D.: Entropy of dialogues creates coherent structures in e-mail traffic. In: Proc. of Natl. Acad. Sci. USA, 101:14333--14337 (2004)
16. Johansen, A.: Probing human response times. Physica A: Statistical Mechanics and its Applications 338:286--291 (2004)
17. Barabasi, A.L.: Modeling bursts and heavy tails in human dynamics. Nature 435: 207--211 (2005)
18. Vazquez, A.: Impact of memory on human dynamics. Physica A: Statistical and Theoretical Physics 373:747--752 (2007)
19. Vazquez, A.: Exact results for the Barabási Model of human dynamics. Physical Review Letters 95, 248701:1-4 (2005)
20. Vazquez, A., Racz, B., Lukacs, A., Barabasi, A.L.: Impact of Non-Poissonian Activity Patterns on Spreading Processes. Physical Review Letters 98, 158702, (2007)
21. Zou, C., Towsley, D., Gong, W.: Email Virus Propagation Modeling and Analysis. Technical Report TR-CSE-03-04. Department of Electrical and Computer Engineering. Univ. of Massachusetts. Amherst
22. Wang, Y., Chakrabarti, D., Wang, C., Faloutsos, C., "Epidemic Spreading in Real Networks: An Eigenvalue Viewpoint", In: Proceedings of 22nd International Symposium on Reliable Distributed Systems, 25--34 (2003)
23. Wang, Y., Chakrabarti, D., Wang, C., Leskovec, J., Faloutsos, C.: Epidemic Thresholds in Real Networks. ACM Transactions on Information and System Security (TISSEC) 10(4), (2008)
24. Klimt, B., Yang, Y.: Introducing the Enron Corpus. In: Proceedings of the 1st Conference on Email and Anti-Spam (CEAS '04). Mountain View, CA (2004)
25. Enron Email Dataset, http://www.cs.cmu.edu/~enron/
26. Barabási, A.-L., Albert. R.: Emergence of scaling in random networks. Science 286, 509--512 (1999)
27. Erdős, P.; Rényi, A.: The Evolution of Random Graphs. Magyar Tud. Akad. Mat. Kutató Int. Közl. 5, 17--61 (1960)
28. Erramilli, A., Roughan, M., Veitch, D., Willinger, W.: Self-Similar Traffic and Network Dynamics. Proceedings of the IEEE 90(5):800--819 (2002)
29. Erramilli, A., Singh, R.P., Pruthi, P.: An application of deterministic chaotic maps to model packet traffic. Queueing Systems 20:171--206 (1995)