Analysis of an evolving email network

Chaopin Zhu,[1] Anthony Kuh,[1] Juan Wang,[2] and Philippe De Wilde[3]
[1]*Department of Electrical Engineering, University of Hawaii at Manoa, 2540 Dole Street, Holmes 483, Honolulu, Hawaii 96822, USA*
[2]*1011 Electrical Engineering Building, Imperial College London, South Kensington Campus, Exhibition Road, London SW7 2BT, United Kingdom*
[3]*Department of Computer Science, Heriot Watt University, Edinburgh EH14 4AS, United Kingdom*
(Received 17 September 2005; revised manuscript received 2 May 2006; published 11 October 2006)

In this paper we study an evolving email network model first introduced by Wang and De Wilde, to the best of our knowledge. The model is analyzed by formulating the network topology as a random process and studying the dynamics of the process. Our analytical results show a number of steady state properties about the email traffic between different nodes and the aggregate networking behavior (i.e., degree distribution, clustering coefficient, average path length, and phase transition), and also confirm the empirical results obtained by Wang and De Wilde. We also conducted simulations confirming the analytical results. Extensive simulations were run to evaluate email traffic behavior at the link and network levels, phase transition phenomena, and also studying the behavior of email traffic in a hierarchical network. The methods established here are also applicable to many other practical networks including sensor networks and social networks.

DOI: 10.1103/PhysRevE.74.046109 PACS number(s): 89.75.Da, 89.20.Hh, 89.75.Fb

I. INTRODUCTION

This paper, which extends [1], proposes a variation of the evolving email network model proposed in [2], whose topology evolvement is driven by Poisson traffic. An analytical analysis is presented to study the network topology evolvement and verify previous empirical work in [2]. Markov chains and random graphs are used to analyze the model. These analytical tools are also useful in analyzing a variety of practical networks from the Internet to social networks to sensor networks. The analytical results obtained not only confirm the empirical results established in [2], but using embedded Markov chains, many of the email network parameters are characterized by steady state distributions. These results are useful as well to help understand the statistical behavior of real email networks observed in [3,4].

Email communications have become an indispensable form of communication in our present information technology society. Of growing concern is the amount of "spam" or unwanted emails that the Internet users receive. Some of these unwanted emails are harmful and spread electronic viruses by automatically sending emails to the addresses in the address books (or files containing email addresses) of infected computers. The viruses have caused serious damage in the past and measures are now being taken to combat these viruses. Hence, it is important to know how email viruses are spread and then to devise effective strategies to control the spreading or alleviate the damage. This paper studies the first issue by looking at a simple model of email networks and analyzes and observes the behavior of these networks. Early research based on real email networks [3,4] at universities show that these networks have small world and scale-free features.[1] Three undirected models: random graph network model, small world network model, and scale-free network model are used to study email virus propagation [7]. Our goal is to study a simple and directed email network model that can naturally explain the behavior and properties of email networks. Wang and De Wilde [2] first proposed a simple evolving email network model and obtained numerous interesting simulation results providing some useful insights into the behavior of email networks. The model that they presented is difficult to analyze. In this paper, we propose an evolving email network model in which the Bernoulli traffic assumption in Wang and De Wilde's model is replaced by a Poisson one. This Poisson assumption allows us to thoroughly analyze both link and network dynamics for both transient and steady state cases. The steady state behavior is virtually identical to the behavior observed in models proposed by [2]. Our analytical results show that when using embedded Markov chains the steady state behavior of many of the email network parameters follows a steady state distribution that can be calculated.

The study of the behavior of email networks and other large scale networks is a growing research area. The survey paper [8] is a good reference for complex and/or large scale networks. In 1959, Erdös and Rènyi introduced random graph theory [9], which defines random networks as a set of nodes connecting to each other with equal probability. These network models have been successfully applied to analyzing

[1]A small world network model is introduced by Watts and Strogatz [5]. The small world model features the combination of high clustering coefficient and small average path length. The scale-free model is proposed by Barabási and Albert [6]. The degree distribution of this model follows a power law tail distribution. More discussion about degree distribution, clustering coefficient, and average path length will be given in Sec. III.

various social and physical networks. It was found that the degree distributions of these networks are Poisson. Beyond the scope of random graph theory, recently it has been found that a lot of networks have power law tail degree distributions. Examples are the world wide web [10,11]; in which a set of HTML (HyperText Markup Language) documents (nodes) are connected by HTML links (links), the actor network in movie industry [5,6], in which each actor represents a node and any two actors are connected if they collaborated in the production of a same movie, and the citation network in academia [12], in which published papers are nodes while citations between papers form links. This power law tail distribution means that these networks have scale-free features. Namely, they are robust to random node failures while vulnerable to targeted attacks to some *hub* nodes. Barabási and Albert [6] explained the scale-free feature as consequences of two mechanisms: network expansion by adding new nodes and the new nodes preferring to attach to well-connected nodes. For many *ad hoc* networks, whose topology is determined by their traffic characteristics, a scale-free feature is highly desired. However, *ad hoc* network traffic is limited by power, geographic distance, bandwidth, protocol overhead, etc. This scalability issue of *ad hoc* networks, including sensor networks, remains as a challenging open problem [13]. This paper focuses on modeling and analysis of email networks while providing a new perspective where *ad hoc* networks can also be studied using many of the same approaches developed here.

The remainder of this paper is organized as follows. An evolving email network model is formulated in Sec. II in which we replace the Bernoulli assumption in [2] with a Poisson one. In Sec. III we use tools from Markov analysis and random graphs to obtain analytical results of this network. Simulations in Sec. IV confirm our analytical results and coincide with the behavior observed in [2]. Section V concludes this paper and discusses further related work and other applications, including sensor networks.

II. EMAIL NETWORK MODEL

In [2], two models are proposed to define the number of emails sent from one node to all other nodes: the equal contact model and the degree-related contact model. In the equal contact model the outgoing email traffic rate of a node is constant while in the degree-related contact model it is directly dependent on the out-degree of the node (number of links to other nodes). In the degree-related contact model, a node with higher out-degree sends more emails to other nodes, it has more users in its address book, which may lead to a scale-free network as the network grows. For a fixed network size, [2] showed that the average number of links converges to a steady state value. Unfortunately the time-varying traffic in the degree-related model makes an analysis of this model difficult. This paper focuses on the equal contact model. An analysis of the time-varying traffic model is a direction of ongoing research.

For the equal contact model discussed in [2] email traffic is generated from a binomial distribution. Using this distribution makes it difficult to analyze email network properties. However, the binomial distribution can be approximated by a Poisson distribution when the number of trials (degree of nodes) is large and the Poisson distribution is amenable to the analysis. In this paper we therefore assume that the amount of email traffic between two users is a Poisson random process. We will see later in the simulation section that the behavior of the model presented here is very similar to those presented in [2].

Assume an email network has N nodes. Each node is assigned an integer from 1 to N as its ID. Let $\mathcal{N}=\{1,2,\ldots,N\}$ denote the set of nodes in the network. Each node stores an address book to represent its connection with all nodes. Let (i,j) denote the ordered node pair of node i and node j and $\mathcal{L}=\{(i,j):i,j\in\mathcal{N}\}$ denote the set of all ordered node pairs in the network. The topology of the network at any time can be described as a directed graph (this graph is the address link connectivity graph). Let an $N\times N$ random adjacency matrix $C(t)=[C_{ij}(t)]$ denote the topology of the network at time instance t. If node j is in the node i's address book at time t there exists a *link* from node i to node j at time t, i.e., $C_{ij}(t)=1$; otherwise $C_{ij}(t)=0$.

Each node is capable of sending emails to all the nodes including itself. For each pair (i,j), denote by $K_{ij}(t,s)$ the number of emails sent from node i to node j during the time interval from t to s. Assume that, for all $(i,j)\in\mathcal{L}$, $K_{ij}(0,t)$ are mutually independent Poisson processes with intensity $\lambda_{ij}>0$, respectively.

Two operations, *generation* and *deletion* discussed in [2], define the network evolvement. Both operations are taken periodically. Let T_g and T_d denote the lengths of *generation period* and *deletion period*, respectively. At the end of each generation period (let this time be $t+T_g$), each node checks the number of emails it sent to a certain destination node during this period. If the amount of traffic is greater than *generation threshold* g, the source node shall put the destination node in its address book; otherwise the network remains unchanged, i.e., for any $(i,j)\in\mathcal{L}$,

$$C_{ij}(t+T_g)=\begin{cases}1, & K_{ij}(t,t+T_g)>g\\ C_{ij}(t), & K_{ij}(t,t+T_g)\leqslant g.\end{cases}$$

Similarly, at the end of each deletion period (let this time be $t+T_d$), each node checks the number of emails it sent to a certain destination node during this period. If the amount of traffic is less than or equal to *deletion threshold* d, the node shall remove the destination node from its address book; otherwise the network remains unchanged, i.e., for any $(i,j)\in\mathcal{L}$,

$$C_{ij}(t+T_d)=\begin{cases}C_{ij}(t), & K_{ij}(t,t+T_d)>d\\ 0, & K_{ij}(t,t+T_d)\leqslant d.\end{cases}$$

A link from node i to node j is established if node j appears in node i's address book.

For any $(i,j) \in \mathcal{L}$, the number of emails sent from node i to node j during a generation period, $K_{ij}(t,t+T_g)$, is a Poisson random variable with parameter $\lambda_{ij}T_g$. Then we define the *generation probability*

$$P_{ij}^g = P\{K_{ij}(t,t+T_g) > g\}.$$

Similarly, we define the *deletion probability*

$$P_{ij}^d = P\{K_{ij}(t,t+T_d) \leqslant d\}.$$

III. STEADY STATE ANALYSIS

The topology of the network model in Sec. II is described by a time-varying adjacency matrix. Each element of this adjacency matrix is a random process. Note that both the generation operation and deletion operation are dependent on the same Poisson traffic. By carefully sampling the random process at discrete times we can define an embedded Markov chain. We can then analyze the Markov chain to find its stationary transition probabilities. To simplify the problem, we first assume that the first generation period and the first deletion period begin at the same time instance 0. In other words, generation and deletion operations are synchronized at the beginning. At the end of this section we give arguments that show our results still remain valid even when this assumption is dropped. Before presenting our formulation, we define the following term.

Definition. A random process has *cyclic steady state distribution* with period T if $\lim_{n\to\infty} P(nT+t_0)=\Pi(t_0)$ for each t_0, where $P(t)$ is the probability mass function of a random process at time t,$\Pi(t_0)$ is a probability mass function depending on t_0, and $0 \leqslant t_0 < T$.

A. Single ordered node pair

First we consider one ordered node pair. Let T_{max} and T_{min} be the larger one and the smaller one of generation period and deletion period, respectively (i.e., $T_{max}=\max\{T_g,T_d\}$ and $T_{min}=\min\{T_g,T_d\}$). We consider two cases.

(i) Case 1: T_{max} is a multiple of T_{min}.

(ii) Case 2: T_{max}/T_{min} is a rational number.

The former is a special case of the latter. Because of its simplicity, we start with the first case.

1. Case 1

For (i,j) being the tag ordered node pair, we have the following theorem.

Theorem 1. If T_{max} is a multiple of T_{min}, the event of node j in node i's address book has a unique cyclic steady state distribution with period T_{max}.

Proof: Without loss of generality, assume $T_d \geqslant T_g$ and let $T_d=mT_g$, where m is a positive integer. Also without loss of generality, we assume that at time instance $t=nT_d$ (n is a nonnegative integer), the generation operation is taken right before the deletion operation if both operations are taken simultaneously. The tag ordered node pair (i,j) has two states: 0 and 1, i.e., $C_{ij}(t)=1$ if there is a link from node i to node j; $C_{ij}(t)=0$ otherwise. So $C_{ij}(nT_d)$ $(n=\{0,1,2,\ldots\})$ is a Markov chain. Let $s_{ij}(n)$ denote the state of (i,j), i.e., $s_{ij}(n)=C_{ij}(nT_d)$. Denote the state distribution by $P_{ij}(n)=[\Pr\{s_{ij}(n)=1\}\ \Pr\{s_{ij}(n)=0\}]'$ where $\Pr\{A\}$ is the probability of event A occurring.

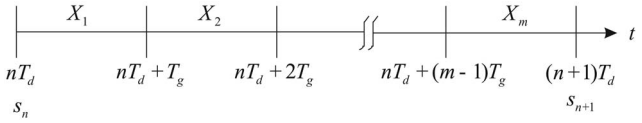

FIG. 1. Illustration of the embedded Markov chain defined in case 1: $T_d=mT_g$. It shows the relations between email traffic $X_1,X_2,\ldots,X_m$ and time instances $nT_d,nT_d+T_g,\ldots,(n+1)T_d$.

For the embedded Markov chain we have the transition matrix Q_{ij} satisfying

$$P_{ij}(n+1) = Q_{ij}P_{ij}(n), \tag{1}$$

where

$$Q_{ij} = \begin{bmatrix} q\{s_{ij}(n+1)=1|s_{ij}(n)=1\} & q\{s_{ij}(n+1)=1|s_{ij}(n)=0\} \\ q\{s_{ij}(n+1)=0|s_{ij}(n)=1\} & q\{s_{ij}(n+1)=0|s_{ij}(n)=0\} \end{bmatrix}.$$

To define these transition probabilities, consider time interval $(nT_d,(n+1)T_d]$. Let X_k be the number of emails sent out from node i to node j during the kth generation period, i.e., $K_{ij}[nT_d+(k-1)T_g,nT_d+kT_g]=X_k(k\in\{1,2,\ldots,m\})$. Relations between time instances and email traffic amounts are shown in Fig. 1. Note that X_ks are independently identically distributed Poisson random variables with parameter $\lambda_{ij}T_g$. Let G denote the complement of event $\{X_1 \leqslant g, X_2 \leqslant g, \ldots, X_m \leqslant g\}$ (i.e., $\bar{G}=\{X_1 \leqslant g, X_2 \leqslant g, \ldots, X_m \leqslant g\}$), and D denote the event of $\{\Sigma_{k=1}^m X_k \leqslant d\}$. When event G occurs a generation operation is taken if there is no link from node i to node j; whereas when event D occurs a deletion operation is taken if there is a link from node i to node j. Then we have

$$\begin{aligned} q\{s_{ij}(n+1)=0|s_{ij}(n)=0\} &= \Pr\{\bar{G}\cup(G\cap D)\} \\ &= \Pr\{\bar{G}\} + \Pr\{G\cap D\} \\ &= \Pr\{\bar{G}\} + \Pr\{D\} - \Pr\{\bar{G}\cap D\} \\ &= (1-P_{ij}^g)^m + P_{ij}^d - \Pr\{\bar{G}\cap D\}, \end{aligned}$$

and similarly

$$q\{s_{ij}(n+1)=1|s_{ij}(n)=0\} = 1 - P_{ij}^d - (1-P_{ij}^g)^m + \Pr\{\bar{G}\cap D\},$$

$$q\{s_{ij}(n+1)=0|s_{ij}(n)=1\} = P_{ij}^d,$$

$$q\{s_{ij}(n+1)=1|s_{ij}(n)=1\} = 1 - P_{ij}^d.$$

Since the state space of the Markov chain is finite and every element of transition matrix Q_{ij} is positive, it is a recurrent Markov chain. Obviously it is also irreducible and aperiodic. So the Markov chain has a unique steady state distribution [14] $\Pi_{i,j}=\lim_{n\to\infty}P_{ij}(n)$ such that

$$\Pi_{ij} = Q_{ij}\Pi_{ij}.$$

In a deletion period, only the generation operation is taken. We could study embedded Markov chains at different time $C_{ij}(nT_d+kT_g)$, where $k=\{1,2,\ldots,m\}$, $(n=\{0,1,2,\ldots\})$. This also has a unique steady state distribution

$$\begin{bmatrix} 1 & P_{ij}^g \\ 0 & 1-P_{ij}^g \end{bmatrix}^{k-1} \Pi_{ij}.$$

So we say that the event of node j in node i's address book has a unique cyclic steady state distribution with period T_d.■

Remark 1 (Steady state link probabilities). In the above proof, the upper element of column vector Π_{ij}, denoted by $\Pi_{ij}(1)$, is the minimal steady state probability that there exists a link from node i to node j. The maximal steady state probability, which is the upper element of vector

$$\begin{bmatrix} 1 & P_{ij}^g \\ 0 & 1-P_{ij}^g \end{bmatrix}^{m-1} \Pi_{ij},$$

should occur after the $(m-1)$th generation period in the T_d-length cycle.

Generally the term $\Pr\{\bar{G}\cap D\}$ is complicated. In the case of $g \geqslant d$, we have

$$\Pr\{\bar{G}\cap D\} = \Pr\{D\} = P_{ij}^d.$$

Before moving forward, let us take a look at a simple example.

Example 1. In an email network with ten nodes, consider two nodes: node 1 and node 2. The traffic from node 1 to node 2 is Poisson with intensity 0.1 messages per day. Every 10 days node 1 inspects if the number of emails it sent to node 2 is more than one message. If so, node 2's address should be added in node 1's address book. Every 20 days, if no message is sent to node 2, node 2's address should be deleted from node 1's address book.

In this case, we have $\lambda_{12}=0.1$, $T_g=10$, $T_d=20$, $g=1$, and $d=0$. Hence

$$P_{12}^g = 1 - e^{-(\lambda_{12}T_g)}(1+\lambda_{12}T_g) = 0.26,$$

$$P_{12}^d = e^{-(\lambda_{12}T_d)} = 0.14,$$

and

$$Q_{12} = \begin{bmatrix} 1-P_{12}^d & 1-(1-P_{12}^g)^2 \\ P_{12}^d & (1-P_{12}^g)^2 \end{bmatrix} = \begin{bmatrix} 0.86 & 0.46 \\ 0.14 & 0.54 \end{bmatrix}$$

Finally we have

$$\Pi_{12} = \begin{bmatrix} 0.77 \\ 0.23 \end{bmatrix}.$$

This means asymptotically the probability that node 2 appears in node 1's address book is 0.77 after at the end of every 20 days and the statistical average number of links emanating from node 1 is 7.7. Ten days later another possible generation operation may be taken, then the probability that node 2 appears in node 1's address book changes to 0.83 and the statistical average number of links emanating from node 1 becomes 8.3.

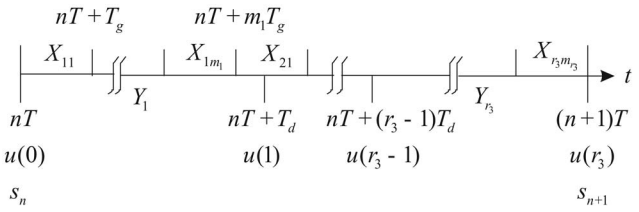

FIG. 2. Illustration of the embedded Markov chain defined in case 2: $T=r_3T_d$. It shows the relations between email traffic X's, Y's, and time instances.

2. Case 2

In this case, T_{max}/T_{min} is a rational number. Let T be the least common multiple of the generation period and the deletion period, i.e., $T=\mathrm{lcm}\{T_g,T_d\}=\mathrm{lcm}\{T_{max},T_{min}\}$.

Theorem 2. If T_{max}/T_{min} is a rational number, the event of node j in node i's address book has a unique cyclic steady state distribution with period T.

Proof. Without loss of generality, assume $T_d > T_g$. Let r_1 be the greatest integer less than $T_d/T_g(r_1=\lfloor T_d/T_g\rfloor)$ and r_2 be the smallest integer greater than $T_d/T_g(r_2=\lceil T_d/T_g\rceil)$. Let $r_3=T/T_d$. For an ordered node pair (i,j), note that $C_{ij}(nT)$ is a Markov chain. This Markov chain also has two states 0 and 1. Let $s_{ij}(n)$ denote the state of the Markov chain at time instance nT, i.e., $s_{ij}(n)=C_{ij}(nT)$. Let m_h be the number of generation periods in hth $(h\in\{1,2,\ldots,r_3\})$ length-T_d time interval in the tag interval $(nT,(n+1)T]$. So $m_h=r_2$ if $(h-1)T_d\leqslant(\Sigma_{f=0}^{h-1}m_f)T_g+r_2T_g\leqslant hT_d$ otherwise $m_h=r_1$. Similar to the proof of Theorem 1, we use random variables X's and introduce random variables Y's to denote the number of emails sent out from node i to node j during generation periods and deletion periods, respectively. Let $Y_h=K_{ij}[nT+(h-1)T_d,nT+hT_d]$ and $X_{h,k}=K_{ij}[nT+(\Sigma_{f=0}^{h-1}m_f)T_g+(k-1)T_g,nT+(\Sigma_{f=0}^{h-1}m_f)T_g+kT_g]$, where $k\in\{1,2,\ldots,m_h\}$. Let G_h denote the complement of $\{X_{h,1}\leqslant g,X_{h,2}\leqslant g,\ldots,X_{h,m_h}\leqslant g\}$ and $D_h=\{Y_h\leqslant d\}$.

To analyze the embedded Markov chain, we look at each of the r_3 deletion periods. We are interested in $C_{ij}(nT+hT_d)$ for $0\leqslant h\leqslant r_3$. Let $u(h)=C_{ij}(nT+hT_d)$, $0\leqslant h\leqslant r_3$ and u denote the r_3+1 binary row vector. Denote by U the set of all r_3+1 binary row vectors. Relations between these definitions and time instances are shown in Fig. 2. Let F_u denote the union of all events that cause the transitions denoted by u in the tag time interval and E_h^u denote the union of all events that cause transitions denoted by $u(h-1:h)$, the $(h-1)$th and hth elements of u, from time $nT+(h-1)T_d$ to time $nT+hT_d$, i.e., $F_u=\{\cap_{h=1}^{r_3}E_h^u\}$. Actually, we have

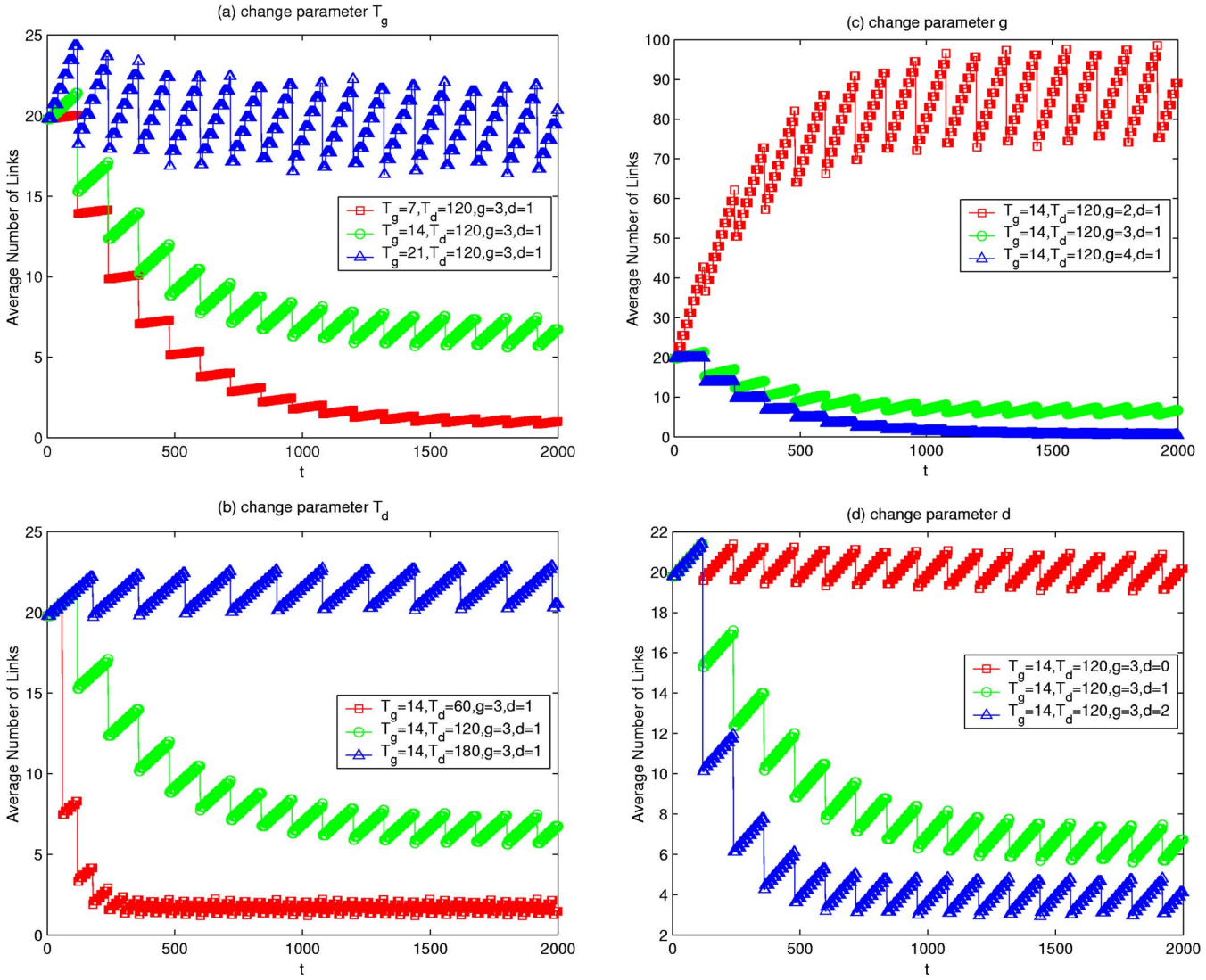

FIG. 3. (Color online) Average number of links of synchronous uniform networks. Both the first generation periods and the deletion periods start at time 0. The traffic rates between all ordered node pairs are $\lambda_{ij}=0.02$. Initially for each node, set generation period $T_g=14$, deletion period $T_d=120$, generation threshold $g=3$, and deletion threshold $d=1$. The average number of links is averaged over 64 simulation runs. Each plot corresponds to the results by varying T_g, T_d, g, and d, respectively.

$$E_h^u = \begin{cases} \bar{G}_h \cup (G_h \cap D_h), & u(h-1{:}h) = 00 \\ G_h \cap \bar{D}_h, & u(h-1{:}h) = 01 \\ D_h, & u(h-1{:}h) = 10 \\ \bar{D}_h, & u(h-1{:}h) = 11. \end{cases}$$

Note that F_u satisfies the following.

(i) For any $u \in U$, $F_u \neq \phi$.

(ii) The set $\{F_u : u \in U\}$ partitions the set of all possible traffic in the tag interval.

Note also that $u(0)=s_{ij}(n)$ and $u(r_3)=s_{ij}(n+1)$. So we have transition probability

$$q\{s_{ij}(n+1)=0|s_{ij}(n)=0\} = \Pr\{\cup_{u:u(r_3)=0,u(0)=0} F_u\} = \sum_{u:u(r_3)=0,u(0)=0} \Pr\{F_u\}.$$

Similarly we have transition probabilities

$$q\{s_{ij}(n+1)=1|s_{ij}(n)=0\} = \sum_{u:u(r_3)=1,u(0)=0} \Pr\{F_u\},$$

$$q\{s_{ij}(n+1)=0|s_{ij}(n)=1\} = \sum_{u:u(r_3)=0,u(0)=1} \Pr\{F_u\},$$

$$q\{s_{ij}(n+1)=1|s_{ij}(n)=1\} = \sum_{u:u(r_3)=1,u(0)=1} \Pr\{F_u\}.$$

Then Eq. (1) is still valid by changing transition matrix Q_{ij}. The Markov chain is also irreducible, recurrent, and aperiodic. So it has a unique steady state distribution. Similar

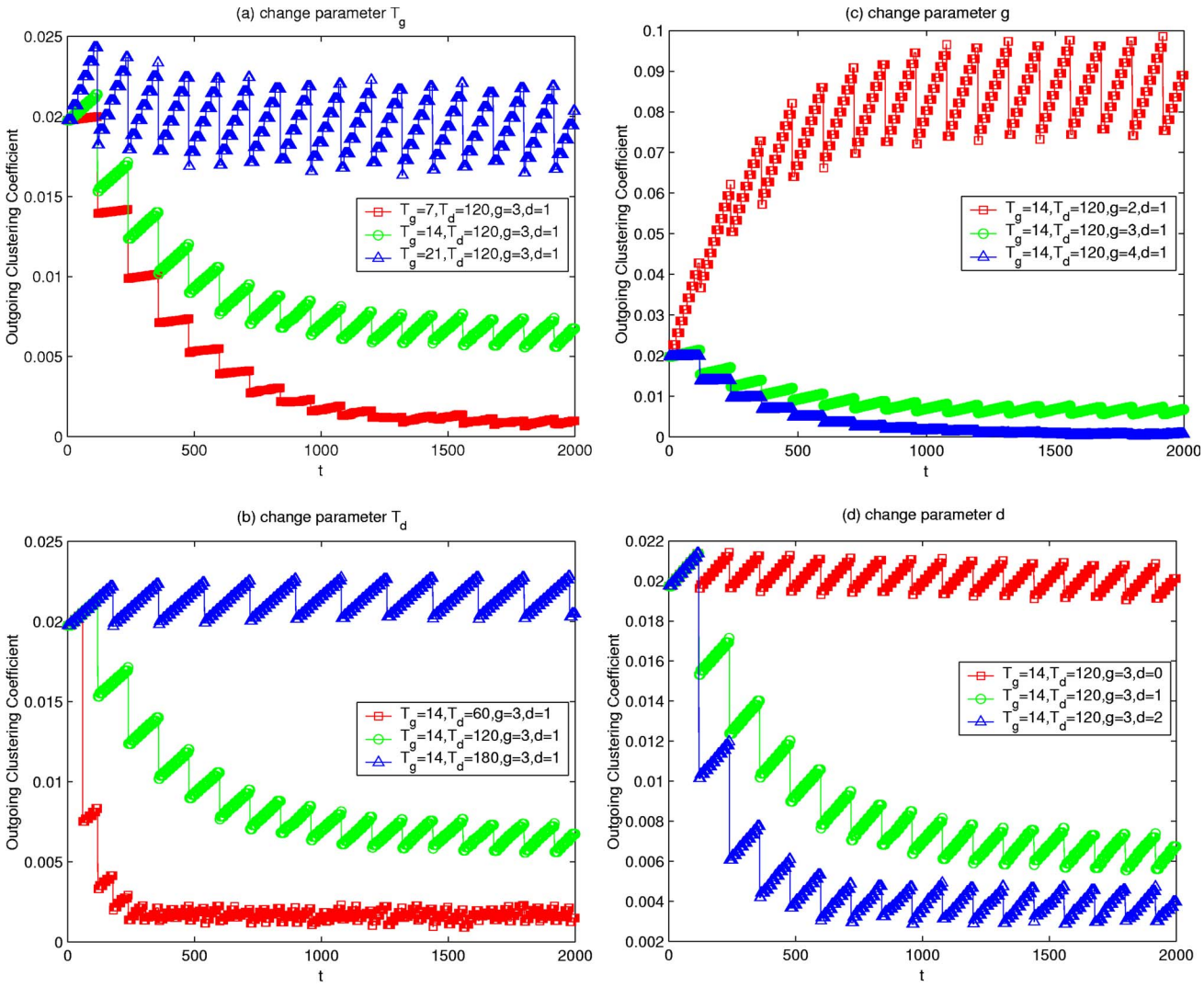

FIG. 4. (Color online) Outgoing clustering coefficients of synchronous uniform networks. Both the first generation periods and the deletion periods start at time 0. The traffic rates between all ordered node pairs are $\lambda_{ij}=0.02$. Initially for each node, set generation period $T_g=14$, deletion period $T_d=120$, generation threshold $g=3$, and deletion threshold $d=1$. The outgoing clustering coefficients are averaged over 64 simulation runs. Each plot corresponds to the results by varying T_g, T_d, g, and d, respectively.

arguments can show that the event of node j in node i's address book has a unique cyclic steady state distribution with period T. ■

Remark 2 (Dependence of E_h^u). Obviously, we have $\Pr(E_h^u \cap E_{h+\delta}^u)=\Pr(E_h^u)\Pr(E_{h+\delta}^u)$ $(|\delta| \in \{\pm 2, \pm 3, \ldots\})$. In general, the events of E_h^u and E_{h+1}^u are dependent so that the calculation of Q_{ij} is more complex. For the case that $g \geqslant d$, we have

$$E_h^u = \begin{cases} \bar{G}_h, & u(h-1:h)=00 \\ G_h, & u(h-1:h)=01 \\ D_h, & u(h-1:h)=10 \\ \bar{D}_h, & u(h-1:h)=11. \end{cases}$$

Hence $\Pr\{E_h^u \cap E_{h+1}^u\}=\Pr\{E_h^u\}\Pr\{E_{h+1}^u\}$ except when $E_h^u=D_h$ and $E_{h+1}^u=G_{h+1}$ or $E_{h+1}^u=\bar{G}_{h+1}$ as

$$\Pr\{D_h \cap G_{h+1}\} = \Pr\{D_h\} - \Pr\{D_h \cap \bar{G}_{h+1}\}.$$

B. All ordered node pairs

Now we can extend Theorem 2 to the whole network by redefining states.

Theorem 3. If T_{max}/T_{min} is a rational number, the topology of the whole network has a unique cyclic steady state distribution with period T.

Proof: Using the same assumptions in the proof of Theorem 2 we denote the state of the whole network by $S(n)=\{s_{ij}(n), 1 \leqslant i,j \leqslant N\}$. To be clear, rewrite $S(n)$ as a N^2 binary row vector. Element h' represents the state of the h'th node pair in the ordered set $\mathcal{L}$ as that in the proof of Theorem 2. $S(n)$ has a total of 2^{N^2} states. Denote the state probability distribution by

$$P(n) = \begin{bmatrix} \Pr\{S(n) = 1111 \cdots 1111\} \\ \Pr\{S(n) = 1111 \cdots 1110\} \\ \Pr\{S(n) = 1111 \cdots 1101\} \\ \Pr\{S(n) = 1111 \cdots 1100\} \\ \vdots \\ \Pr\{S(n) = 0000 \cdots 0000\} \end{bmatrix}.$$

Since each ordered node pair has independent email traffic, we have the following equation:

$$P(n+1) = (\otimes_{(i,j)\in\mathcal{L}} Q_{ij}) P(n), \tag{2}$$

where $\otimes$ is the Kronecker product operator.[2] Similarly, we can argue that the Markov chain is irreducible, recurrent, and aperiodic. Therefore it has a unique steady state distribution $\Pi = \lim_{n\to\infty} P(n)$ such that

$$\Pi = (\otimes_{(i,j)\in\mathcal{L}} Q_{ij})\Pi.$$

Similar arguments show that the topology of the whole network has a unique cyclic steady state distribution with period T. ■

Remark 3 (Comments about asynchronous cases). We can also remove an earlier assumption that the first generation and deletion periods start at the same time. In this asynchronous case, we still can define embedded Markov chains at the end of these deletion periods as above. Obviously these Markov chains still have stationary transition matrices and are irreducible, recurrent, and aperiodic. Then all the above theorems hold though the cyclic steady state distributions will change.

In addition, assume that each ordered node pair has the same network traffic rate, i.e., $\lambda_{ij} = \lambda$. Then in steady state, the whole network can be viewed as a random directed graph with connecting probability $p(t)$, which has period T. There are three robust measures, degree distribution, clustering coefficient, and average path length,[3] to describe random undirected networks. For random directed networks, Zhou [17] extended these definitions to in-degree and/or out-degree distribution and incoming and/or outgoing clustering coefficient. From the random graph theory [9], we have the following result:

Corollary 1. If T_{max}/T_{min} is a rational number and each ordered node pair has same network traffic rate:

(i) both in-degree and out-degree distributions of the whole network have unique cyclic steady state distributions with period T;

(ii) both incoming clustering coefficient and outgoing clustering coefficient of the whole network have unique cyclic steady state values with period T.

We conjecture that the average path length of a connected network has a unique cyclic steady state value with period T. Simulations in the next section justify this observation. The following simple example also illustrates these results.

Example 2. With the same parameter setting for each ordered node pair as that of example 1, asymptotically at the end of every twenty days, the in-degree and out-degree distributions of the network are binomial distributions with parameter 0.77 and both the incoming and outgoing clustering coefficients are 0.77 ten days after this time instance, both in-degree and out-degree distributions of the network are binomial distributions with parameter 0.83 and both the incoming and outgoing clustering coefficients are 0.83.

In reality, a network may have a lot of domains. Email users in the same domain communicate much more than those in different domains. For instance, consider a hierarchy network with two traffic levels. The in-domain and out-domain email traffic has intensity λ_{in} and λ_{out}, respectively, with $\lambda_{in} \gg \lambda_{out}$. The theorems and corollary that we stated in this section are still valid.

Since each ordered node pair evolves independently, all the above results still hold even when each ordered node pair has a different Poisson traffic rate. Also note that these results can be extended to the case where each node has different values for the generation period, deletion period, generation threshold, and deletion threshold.

IV. SIMULATION

In this section simulations are used to verify the analytical results obtained in Sec. III. Here we connect our model to a traditional random graph network model and a small world network model. We start with uniform networks where each ordered node pair has the same Poisson traffic rate and then move to hierarchy networks, in which the Poisson traffic rate of each ordered node pair depends on which hierarchy level it falls in. For both kinds of networks, we investigate three topology measures of random networks: degree distribution, clustering coefficient, and av-

[2]The Kronecker product of matrix $A_{pq} = [a_{ij}]$ and matrix B_{mn} is

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1q}B \\ a_{21}B & a_{22}B & \cdots & a_{2q}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1}B & a_{p2}B & \cdots & a_{pq}B \end{bmatrix}.$$

Please refer to [15,16] for details.

[3]For any node of a random network, its degree is a random variable. The collective distribution of degrees of all nodes is defined as the degree distribution of the graph. For any node that has the neighborhood, the clustering coefficient of this node is the ratio of the number of edges formed by its neighbors to a number of edges of a completely connected neighborhood. Clustering coefficient of a random network is obtained by collectively averaging over all node. The average path length, also known as the characteristic path length, is the collective average of the shortest path length between any node pair. For details, please refer to [6].

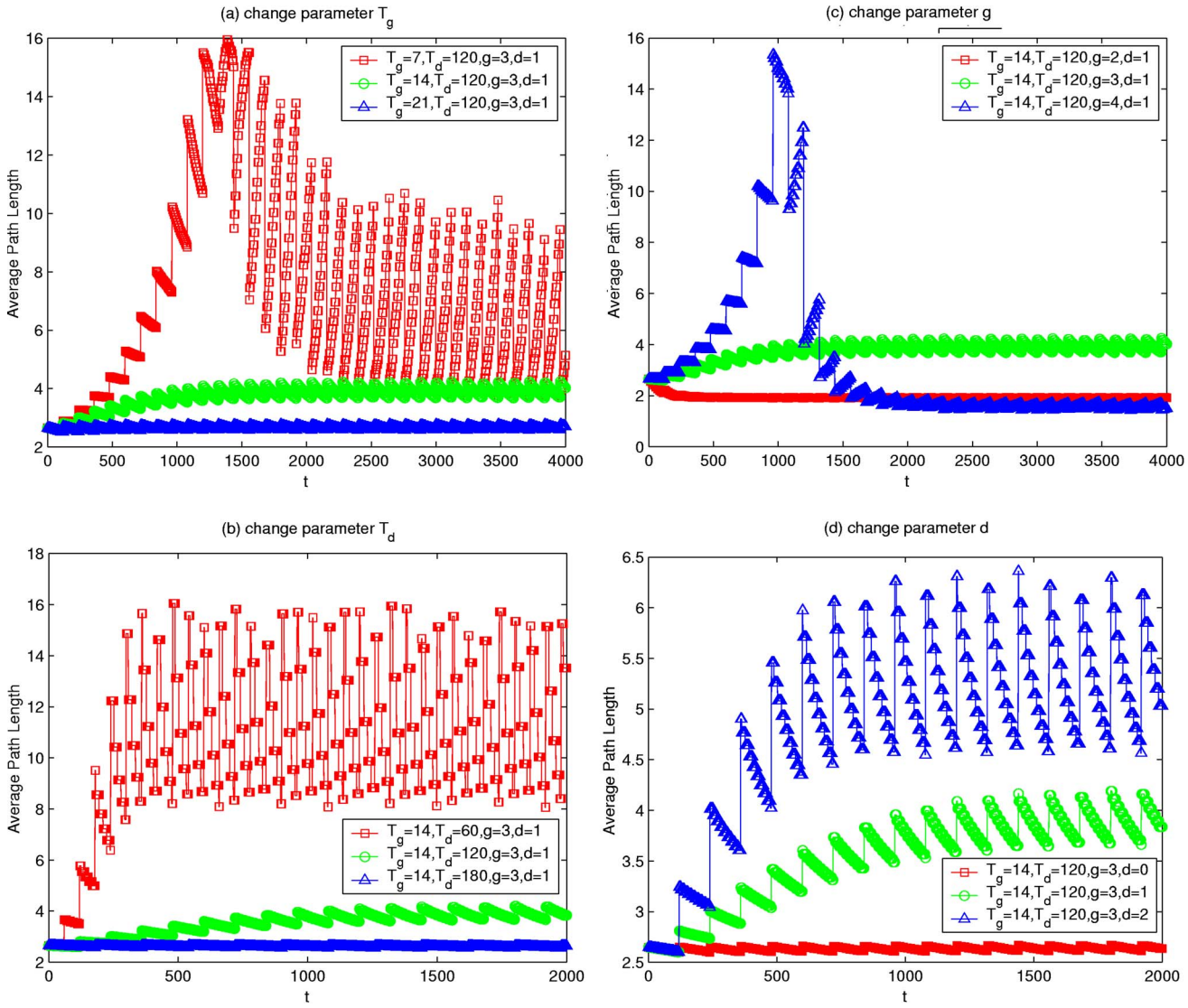

FIG. 5. (Color online) Average path length of synchronous uniform networks. Both the first generation periods and the deletion periods start at time 0. The traffic rates between all ordered node pairs are $\lambda_{ij}=0.02$. Initially for each node, set generation period $T_g=14$, deletion period $T_d=120$, generation threshold $g=3$, and deletion threshold $d=1$. The average path length is averaged over 64 simulation runs. Each plot corresponds to the results by varying T_g, T_d, g, and d, respectively.

erage path length. We also examine the phase transition phenomena.[4]

Following [2] we consider networks with 1000 ($N=1000$) nodes. Since, in steady state, different initial settings give us the same results, we choose a simple random initial setting as follows. For any $(i,j)\in\mathcal{L}$, Poisson email traffic is generated starting at time -1. If the number of emails sent from node i to node j up to time 0, [given by $K_{ij}(-1,0)$] is greater than or equal to 1, $C_{ij}(0)=1$;
otherwise $C_{ij}(0)=0$.

[4]For a large uniform network we examine how a network transitions from being not connected to fully connected as we slowly increase the Poisson arrival rate λ. Related background knowledge is in[8].

A. Synchronous uniform networks

In the synchronous case, both generation and deletion operations are synchronized to start from time 0. In other words, the first generation period and the first deletion period start at time 0. In this subsection, we set parameters the same as those in [2]. For each ordered node pair, $\lambda_{ij}=0.02$; for each node, $T_g=14$, $T_d=120$, $g=3$, and $d=1$. During each run of simulation only one parameter changes while all other parameters are fixed. In the following we investigate the three topology measures and the phase transition phenomena.

1. Average number of links

In Sec. III both the in-degree distribution and the out-degree distribution of the network have cyclic binomial distribution with period T, the least common multiple of the

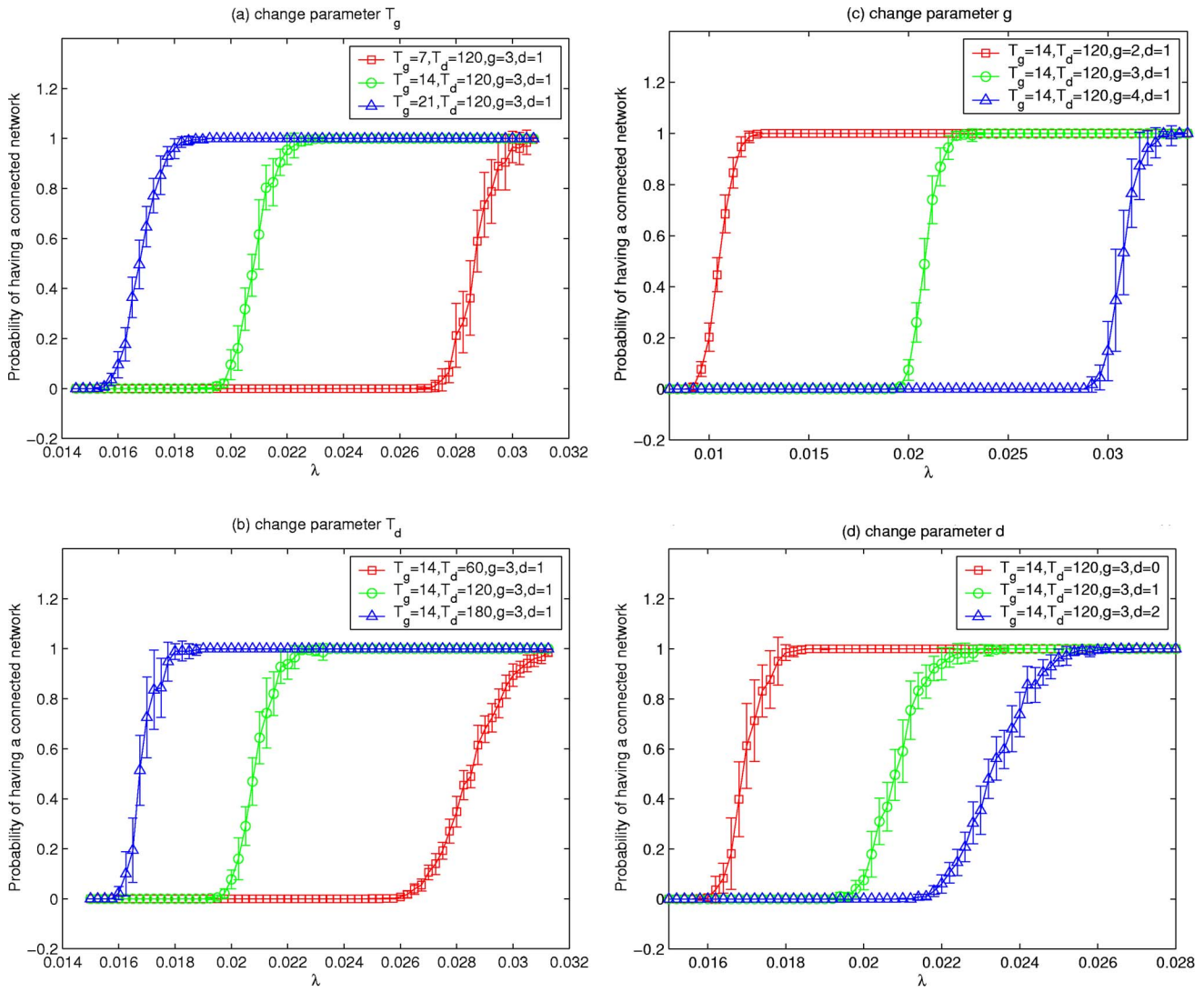

FIG. 6. (Color online) Phase transition phenomena of synchronous uniform networks (probability of having a connected network vs traffic rate). Both the first generation periods and the deletion periods start at time 0. The traffic rates between all ordered node pairs are $\lambda_{ij}=0.02$. Initially for each node, set generation period $T_g=14$, deletion period $T_d=120$, generation threshold $g=3$, and deletion threshold $d=1$. Average and standard deviation of the probabilities are obtained over 20 simulation runs and each run has 1680 steady state networks. Each plot corresponds to the results by varying T_g, T_d, g, and d, respectively.

generation period and the deletion period. Asymptotically, the binomial distribution converges to the Poisson distribution, which has its mode at its average. For simplicity, we investigate the average number of links (or the average degree), $[\Sigma_{(i,j)\in\mathcal{L}}C_{ij}(t)]/N$, instead of the degree distribution.

We ran simulations over 64 runs and averaged the number of links to get the plots in Fig. 3. Results for Poisson traffic are very close to the Bernoulli traffic used in [2]. The analysis in Sec. III implies that the average number of links reaches cyclic steady state values. It is confirmed by the plots in Fig. 3. Then we investigated the effect of changing the generation threshold g. Regardless of what the parameters are, the average number of links has cyclic steady state values with period $T=840$, the least common multiple of T_g and T_d. The smaller the generation threshold, the larger the steady state values.

With the setting of $\lambda_{ij}=0.02$ $[(i,j)\in\mathcal{L}]$, $T_g=14$, $T_d=120$, $g=3$, and $d=1$, by applying the analytical method developed in the above section, we have the generation probability given by

$$P_{ij}^{g} = 1 - e^{-\lambda_{ij}T_g}\left(1 + \lambda_{ij}T_g + \frac{(\lambda_{ij}T_g)^2}{2!} + \frac{(\lambda_{ij}T_g)^3}{3!}\right)$$
$$= 1 - e^{-0.28}\left(1 + 0.28 + \frac{0.28^2}{2!} + \frac{0.28^3}{3!}\right)$$
$$= 0.0002,$$

and the deletion probability given by

$$P_{ij}^{d} = e^{-\lambda_{ij}T_d}(1 + \lambda_{ij}T_d) = e^{-2.4}(1 + 2.4) = 0.31.$$

We also have that

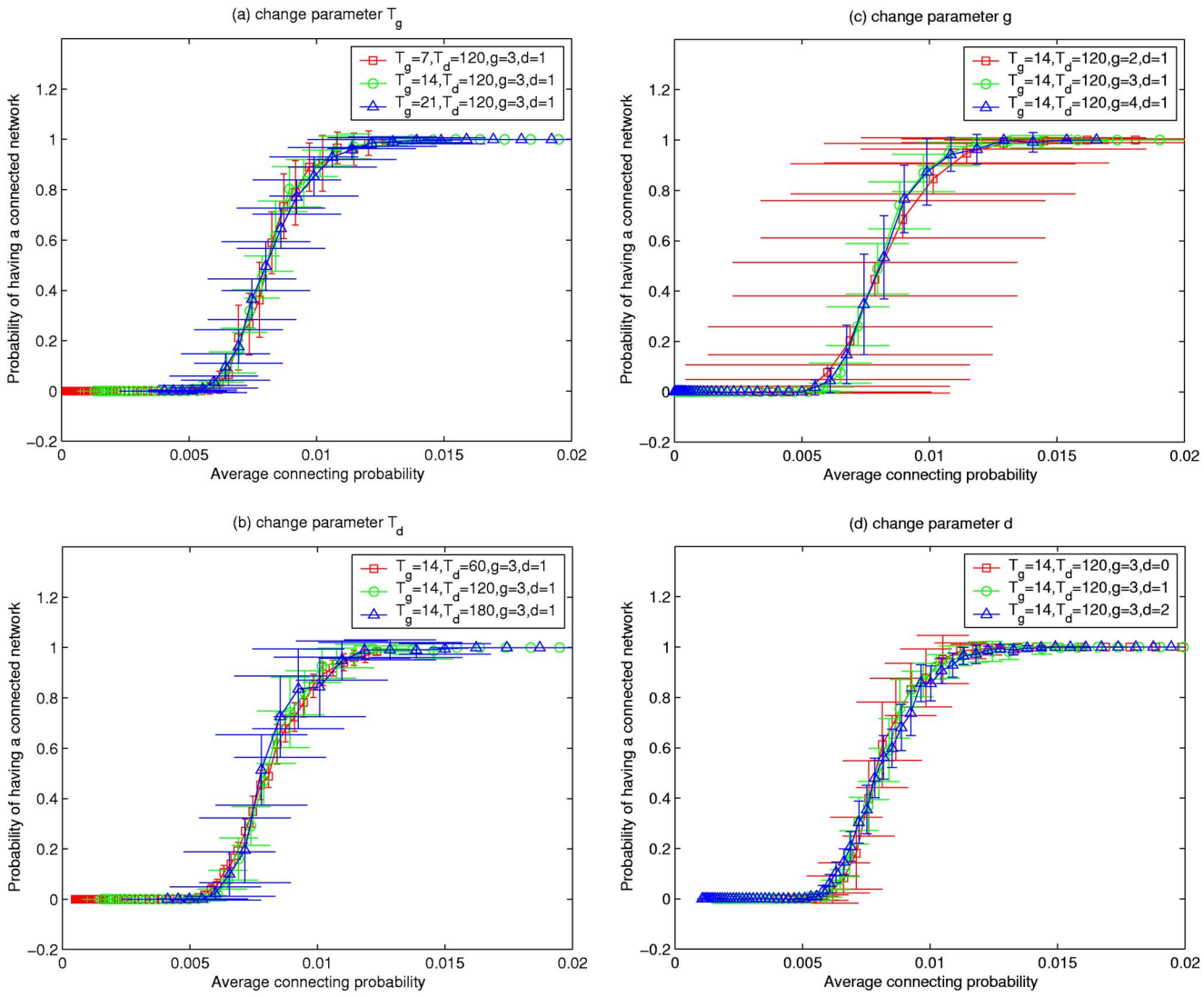

FIG. 7. (Color online) Phase transition phenomena of synchronous uniform networks (probability of having a connected network vs average connecting probability). The traffic rates between all ordered node pairs are $\lambda_{ij}=0.02$. Initially for each node, set generation period $T_g=14$, deletion period $T_d=120$, generation threshold $g=3$, and deletion threshold $d=1$. Average and standard deviation of the probabilities are obtained over 20 simulation runs and each run has 1680 steady state networks. Each plot corresponds to the results by varying T_g, T_d, g, and d, respectively.

$$
\begin{aligned}
&\Pr\{D_h \cap \bar{G}_{h+1}\} \\
&\quad = \Pr\{Y_h \leqslant d, X_{h+1,1} \leqslant g, X_{h+1,2} \leqslant g, \ldots, X_{h+1,m_{h+1}} \leqslant g\} \\
&\quad = \Pr\{Y_h \leqslant 1, X_{h+1,1} \leqslant 3, X_{h+1,2} \leqslant 3, \ldots, X_{h+1,m_{h+1}} \leqslant 3\} \\
&\quad = \Pr\{Y_h = 0, X_{h+1,1b} \leqslant 3, X_{h+1,2} \leqslant 3, \ldots, X_{h+1,m_{h+1}} \leqslant 3\} \\
&\quad + \Pr\{Y_h - X_{h+1,1a} = 1, X_{h+1,1a} = 0, X_{h+1,1b} \leqslant 3, X_{h+1,2} \\
&\quad \leqslant 3, \ldots, X_{h+1,m_{h+1}} \leqslant 3\} + \Pr\{Y_h - X_{h+1,1a} = 0, X_{h+1,1a} \\
&\quad = 1, X_{h+1,1b} \leqslant 2, X_{h+1,2} \leqslant 3, \ldots, X_{h+1,m_{h+1}} \leqslant 3\},
\end{aligned}
$$

where $X_{h+1,1a}$ and $X_{h+1,1b}$ are the fractions of $X_{h+1,1}$ contributing to Y_h and Y_{h+1}, respectively. Then we get

$$
Q_{ij} = \begin{bmatrix} 0.0796 & 0.00536 \\ 0.920 & 0.995 \end{bmatrix}.
$$

Hence, we obtain the steady state distribution $\Pi_{ij} = [0.00578\ 0.994]'$. This shows that the pair is connected with probability 0.005 78 at the end of every 840 time units. So the average number of links of the 1000-node network goes to $0.00578 \times 1000 = 5.78$ at those time instances. This is confirmed by the middle curve in Fig. 3(a). Similar trends are observed by changing deletion threshold d, generation period T_g, and deletion period T_d individually (refer to the plots in Fig. 3).

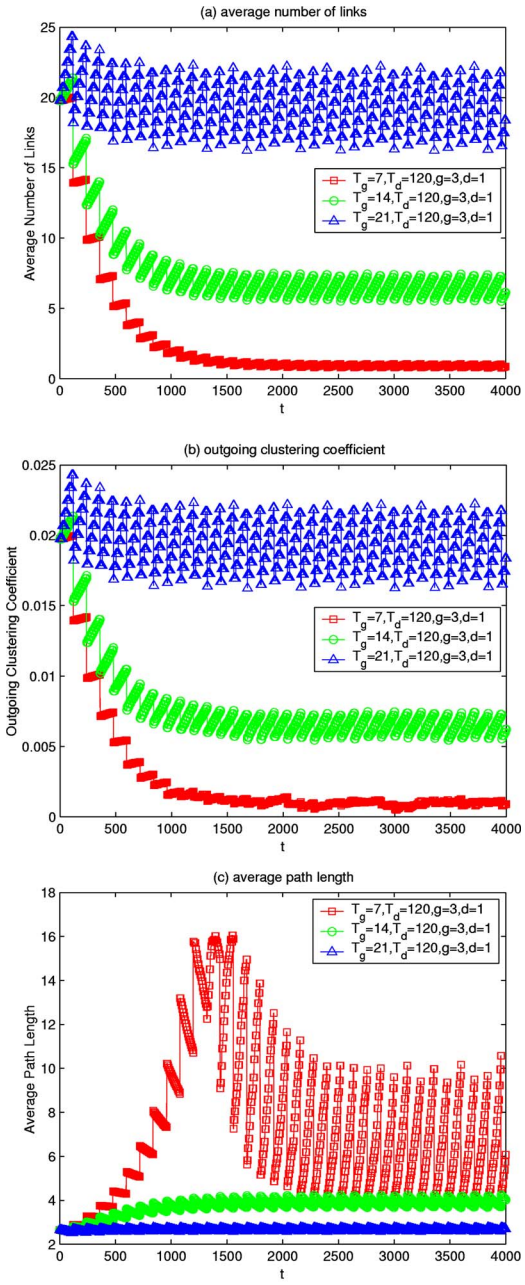

FIG. 8. (Color online) Average number of links, outgoing clustering coefficient, and average path length of asynchronous uniform networks. All first generation periods start at time 1 while all first deletion periods start at time 0. The traffic rates between all ordered node pairs are $\lambda_{ij}=0.02$. Initially for each node, set generation period $T_g=14$, deletion period $T_d=120$, generation threshold $g=3$, and deletion threshold $d=1$. The results are averaged over 64 simulation runs. Each curve in each plot corresponds to different values of T_g.

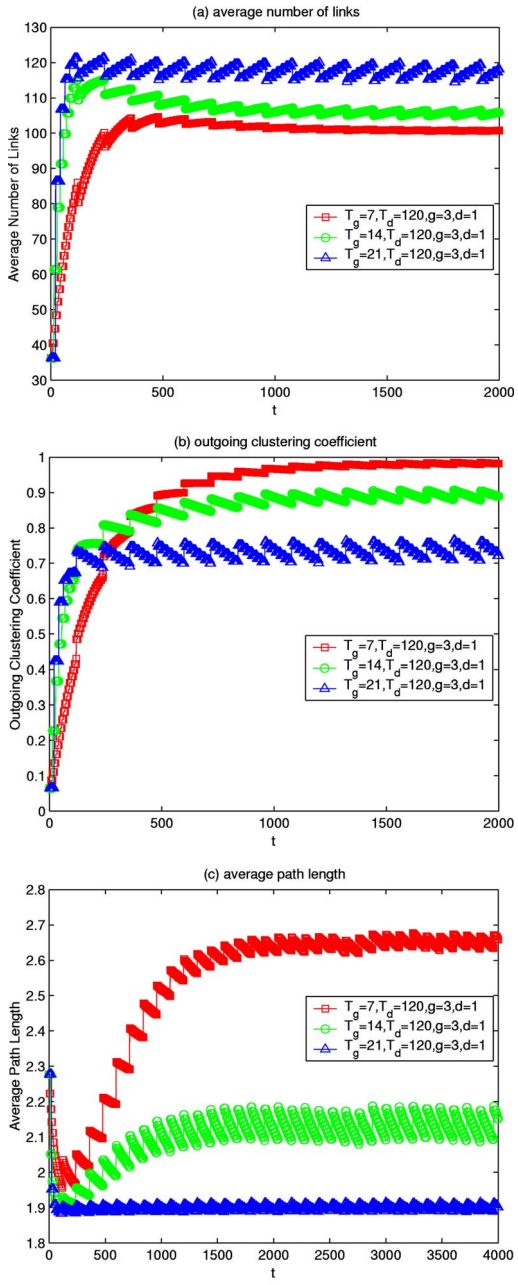

FIG. 9. (Color online) Average number of links, outgoing clustering coefficient, and average path length of synchronous hierarchy networks. Both the first generation periods and the deletion periods start at time 0. The traffic rates between nodes in the same (different) domains are $\lambda_{in}=0.2$ ($\lambda_{out}=0.02$). Initially for each node, set generation period $T_g=14$, deletion period $T_d=120$, generation threshold $g=3$, and deletion threshold $d=1$. The results are averaged over 64 simulation runs. Each curve in each plot corresponds to different values of T_g.

2. Clustering coefficients

As we pointed out in Sec. III, both the incoming and outgoing clustering coefficients of the network are periodic with period T, the least common multiple of the generation and deletion periods. As space is limited, only the simulation results of outgoing clustering coefficients are shown here. By taking the average over 64 runs, we obtain the plots in Fig. 4. These plots confirm the analytical results in Sec. III.

In addition, we noted that there is some quantitative difference between the results in Fig. 4 and those in [2]. This is because that we use different definitions of clustering coefficients. Instead of the conventional clustering coefficients used in [2], the incoming and outgoing clustering coefficient-

sare used to obtain our results as they are defined for random directed graphs like those of email networks while the conventional clustering coefficients are defined for random undirected graphs.

3. Average path length

We are not aware of any closed form results of the average path length of a random graph. However, we conjecture that as long as the network is connected, the average path length of the network is also periodic with period T, the least common multiple of the generation and deletion periods. Simulations are averaged over 64 runs and are presented in Fig. 5 demonstrating the cyclic steady state behavior. In this simulation, all paths are directed and both infinite paths and self-loops are excluded from averaging. We note that two out of three curves in each plot have simulations with infinite paths.

4. Phase transition

The email network evolvement is driven by email traffic. Here we investigate the effect of varying the traffic rate. A network is connected if and only if there is a direct path for each ordered node pair. We inspect the evolvement of the 1000-node network from time 2001 to time 3680 in 20 simulation runs. In each run, the probability of having a connected network is approximated by the fraction of the 1680 $=(3680-2000)$ networks that is connected. We take averages and standard deviations over 20 runs to obtain the plots in Fig. 6. As expected, a phase transition does occur. For instance, for the setting of $T_g=14$, $T_d=120$, $g=3$, and $d=1$, the network is unlikely to be connected if the traffic rate is less than 0.0208 while the network will likely be connected if the traffic rate is larger than this threshold value. The plot in Fig. 6(c) shows that the smaller the generation threshold, the smaller the value of λ needed for a phase transition. Phase transitions are also observed for other different parameter values as shown in the other plots. To make a connection with the random graph theory, we also plot the relationship between probability of having a connected network and the average connecting probability in Fig. 7. In each plot, the average connecting probabilities are averaged over 20 runs, each from time 2001 to time 3680. We observed that all three curves overlap in each plot and the traffic rate threshold of each curve corresponds to the same average connecting probability of 0.0079. In [9], it was stated that the connecting probability threshold of a random undirected graph having N nodes is $\ln(N)/(N-1)$ if N goes to infinity. Substituting $N=1000$, we have the approximate threshold 0.0069. Our observation is consistent with this statement though we are not aware of any similar analytical results for random directed graphs.

B. Asynchronous uniform networks

In addition, we use simulations to verify our claims in Sec. III that all these results hold under the asynchronous assumption. We use same simulation setting as that in Sec. IV A except that the first generation periods and the deletion periods start at time 1 and time 0, respectively. Shown in Fig. 8 are the corresponding results of the average number of links, outgoing clustering coefficients, the and average path length. Three curves in each plot have different generation periods 7, 4, and 21, respectively. These plots show the similar cyclic patterns observed in Sec. IV A. Similar cyclic patterns are observed when the first generation periods start at any time other than 0.

C. Hierarchy networks

In real life, email traffic rates between any two users vary depending on many factors. To mimic this scenario, we consider a simple two-level hierarchy network. We split the whole network into 10 domains, each containing 100 nodes. Let $\lambda_{out}=0.02$ and $\lambda_{in}=0.2$ while keeping other settings the same as above. Averaging over 64 runs, plots in Fig. 9 are obtained. The cyclic properties, of the average number of links, clustering coefficients, and average path lengths are still observed. Each node connects to nodes in the same domain with high probability while connecting to nodes in other domains with much lower probability. In this setting we observed similar cyclic pattern behavior. Consider the setting in Fig. 9 with $T_g=14$. In the steady state, it has an average number of links around 105.3, outgoing clustering coefficient 0.9, and average path length 2.15. Whereas the corresponding random directed graphs with the same average number of links has an outgoing clustering coefficient 0.105 and average path length 1.89. This network shows significant higher clustering coefficients and similar average path length comparing to random directed graphs with the same average number of links. That is the small world phenomenon pointed out in Table 1 of [5]. This is consistent with the definition of the clustering coefficient. It also naturally explains the small world feature of the real email network observed in [3].

In summary, the simulations conducted for uniform and hierarchy networks both confirm the analytical results derived in Sec. III. In addition, an example of a hierarchy network also shows the small world feature as observed in [3].

V. SUMMARY AND FURTHER DIRECTIONS

In this paper we proposed a modified email network model, replacing the Bernoulli traffic assumption in [2] with a Poisson assumption. The mutual independence of the Poisson traffic allows us to analytically study the steady state distribution of the model. Assuming that the ratio of the generation period to deletion period is rational we use tools from Markov chains and random graphs to show that the network has a unique cyclic steady state distribution with a period that is the least common multiple of the generation period and the deletion period. When the traffic rates between any two nodes are the same, the network can be viewed as a random directed graph that has a cyclic steady state distribution and has properties of both traditional random graph models and small world networks. Simulation results confirmed the analytical results of the random di-

rected graph in terms of the average number of links and clustering coefficients. Plots of average path lengths and phase transition phenomena are also consistent with the random graph networks. We also attempted to mimic real email networks with a hierarchy network. This hierarchy model explains the small world feature of real email networks.

There are many further directions for this research. We would like to consider other networks such as sensor and social networks where the Poisson traffic rate is time-varying. In sensor networks the traffic intensity would also be dependent on distance between nodes and power used to transmit messages.

ACKNOWLEDGMENTS

The first two authors acknowledge support from the Imperial College in London and ESPRC Grant No. GR/S83104/01. The first author also thanks Andreas Koga at UH cluster for his help with using cluster machines for simulations.

[1] C. Zhu, A. Kuh, J. Wang, and P. D. Wilde, Proceedings of the 39th Annual Conference on Information Sciences and Systems (CISS 2005), paper no. 70, Baltimore, MD, March 16–18, 2005 (unpublished).
[2] J. Wang and P. D. Wilde, Phys. Rev. E **70**, 066121 (2004).
[3] H. Ebel, L.-I. Mielsch, and S. Bornholdt, Phys. Rev. E **66**, 035103 (2002).
[4] M. E. J. Newman, S. Forrest, and J. Balthrop, Phys. Rev. E **66**, 035101 (2002).
[5] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[6] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[7] C. C. Zou, D. Towsley, and W. Gong, Proceedings of the 13th International Conference on Computer Communications and Networks (ICCCN '04), Chicago, IL, October 11–13 2004, pp. 409–414 (unpublished).
[8] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[9] P. Erdös and A. Rènyi, Publ. Math. (Debrecen) **6**, 290 (1959).
[10] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **401**, 130 (1999).
[11] B. A. Huberman and L. A. Adamic, Nature (London) **401**, 131 (1999).
[12] S. Redner, Eur. Phys. J. B **4**, 131 (1998).
[13] R. Ramanathan and J. Redi, IEEE Commun. Mag. **40**, 20 (2002).
[14] S. Karlin and H. M. Taylor, *A Second Course in Stochastic Processes* (Academic Press, New York, 1981).
[15] J. W. Brewer, IEEE Trans. Circuits Syst. **25**, 772 (1978).
[16] J. W. Brewer, IEEE Trans. Circuits Syst. **26**, 360 (1979).
[17] H. Zhou, Phys. Rev. E **66**, 016125 (2002).