Routing strategy on a two-dimensional small-world network model

Ming Li,[1] Feng Liu,[1,*] and Feng-Yuan Ren[2]
[1]*School of Electronics and Information Engineering, Beihang University, Beijing 100083, People's Republic of China*
[2]*Department of Electronics Engineering, Tsinghua University, Beijing 100084, People's Republic of China*
(Received 8 November 2006; revised manuscript received 10 March 2007; published 28 June 2007)

Based on a two-dimensional small-world network model, we propose an efficient routing strategy that enhances the network capacity while keeping the average packet travel time low. We deterministically increase the weight of the links attached to the "congestible nodes" and compute the effective distance of a path by summing up the weight of the links belong to that path. The routing cost of a node is a linear combination of the minimum effective distance from the node to the target and its queue length. The weight assignment reduces the maximum load of the network, while the incorporation of dynamic information further balances the traffic on the network. Simulation results show that the network capacity is much improved compared with the reference strategies, while the average packet travel time is relatively small.

DOI: 10.1103/PhysRevE.75.066115 PACS number(s): 89.75.Hc, 89.75.Fb

I. INTRODUCTION

In recent years, small-world networks (SWNs) have attracted wide research interest. Watts and Strogatz (WS) found that the small-world networks possess two properties: small vertex-vertex separation and large clustering coefficient [1]. Since many real-world networks have small-world characteristics, such as metabolic networks [2], transportation networks [3], the World Wide Web [4], and the Internet [5], the topological properties and dynamical processes of SWNs have been studied intensively [6–8].

In small-world networks, the speed of information propagation is much higher than that in regular networks [3,8,9]. In particular, the small-world property is important in obtaining high efficiency of traffic delivery in transportation and communication networks [3,9]. These networks are often characterized by the existence of a few long-range links. In the study of the navigation and searching process on small-world networks [10–12], some "greedy algorithm," in which messages are merely forwarded through the neighbors nearest to the destinations, is shown to be efficient in finding short paths and the long-range links are responsible for the high efficiency [10].

However, in traffic dynamics where many transmission processes take place simultaneously, if communication speed is the only factor taken into account, the end nodes of random links (congestible nodes) would attract more traffic than other nodes and not be able to handle them at a time due to the finite processing capability, finally leading to congestion [13]. So finding an efficient routing strategy that avoids congestion while keeping the time delay of communication as low as possible is extremely important. While most previous works on traffic dynamics are based on scale-free network models [13–21], we investigate the same issue on a classical small-world network model.

As a matter of fact, the network topology is crucial to the network function and performance [13,19,22–25]. When designing routing strategies, better performance can be achieved if the characteristics of network topology are considered. Here we adopt the slightly modified two-dimensional (2D) version [26] of the WS model. In the WS SWN model series, the long-range links between randomly selected nodes—say, shortcuts—characterize the network topology. On the one hand, they reduce the distance between vertices; on the other hand, they cause congestion. So it is very important for the shortcuts to function efficiently. For the same SWN model we study, Fukś *et al.* [27,28] propose a partial routing algorithm that allows packets to move randomly when they are far away from the destinations, but otherwise follow the shortest paths. By doing so, the utilization of shortcuts is restricted to local scale. Though the network capacity is enhanced compared with the pure shortest path routing, the time delay increases again. So introducing randomness into the routing table may not be a good choice to utilize the shortcuts. Some other previous studies focus on enhancing the message-processing capability of the congestible nodes, and the network capacity can be significantly increased [29,30]; in fact, when the capabilities of the nodes are all equal, an equivalent strategy is to reduce the packet arrival rate of the congestible nodes by increasing the cost (weight) of the corresponding links.

In this paper, we find that there exists a simple routing strategy that exploits the characteristics of the specific network topology. Aiming at reducing the maximum load, we set the weight of the links attached to the congestible nodes as proportional to the network length $\bar{l}_0$ with all the congestible nodes being treated equally. The total weights of the paths between pairs of vertices are computed as the effective distances. In order to further balance the load and enhance the network capacity, we linearly combine the minimum effective distance with local queue length to form a routing cost function, and packets are routed according to the minimum cost. The new strategy is called the efficient routing strategy (ERS). Simulations show that for the specific SWN model, the ERS performs much better than several previous strategies.

The paper is organized as follows. In Sec. II, the network model is defined. In Sec. III, the details of the ERS are presented, the selection of parameters is explained, and the congestion-relieving mechanism is elaborated. In Sec. IV, re-

*Corresponding author. liuf@buaa.edu.cn

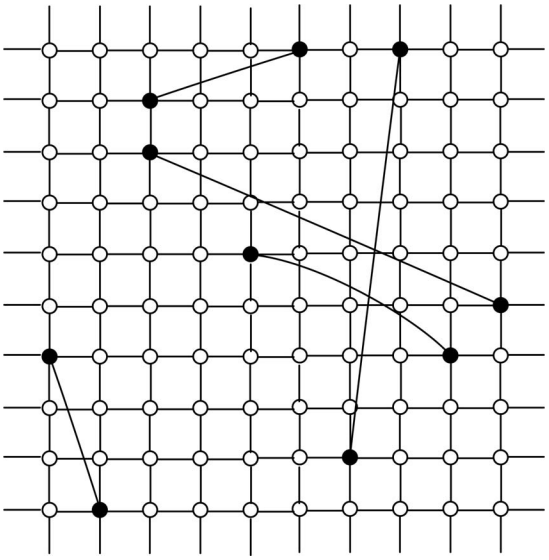

FIG. 1. An example of the 2D small-world model used in this paper.

sults and further explanations are given. We conclude the paper in Sec. V.

II. MODEL AND DEFINITIONS

In the real world, computer hosts are distributed on the surface of our planet, so we employ a two-dimensional regular lattice as the underlying structure. Another reason for this is that the characteristics of realistic Internet traffic can be reproduced by 2D lattices [31]. Long-range links exist between randomly chosen sites, for the geometrical position of the sites connecting distant areas are usually random. One may think of using the original WS small-world model [1]. But as pointed out by Newman and Watts [26], one of the serious problems with random rewiring is that the average distances between pairs of vertices diverge since there is a finite probability of a portion of the lattice becoming detached from the rest in the model. So we adopt a slightly modified version [26,27] here. The model is defined as follows.

For a two-dimensional square lattice with periodical boundary condition which has $N=L\times L$ nodes, each node is only connected to its direct neighbors and every node is a host that generates, forwards, and receives packets. Choose two vertices on the above lattice at random and add a shortcut between them; repeat this for pN times, where p is the rate of shortcuts. Multiple links are not allowed. Since the structure of the underlying lattice is not changed, the local property of the lattice is maintained. Note that for convenience, here p is a rate and not a probability, which is slightly different from the usual small-world models; nevertheless, in the statistical sense, the topological characteristics of the network should be close to that of the one taking p as a probability. Figure 1 shows one realization of our model for $L=10$ and $p=0.05$.

Now we give the definitions of some graph-related quantities. The end nodes of the random links are defined as *congestible nodes*, since the random links usually "attract" a large amount of traffic. The *shortest path* between nodes s and t is the path with the minimum number of links. The *shortest path length* is the number of links on a shortest path and is denoted by l_0. The *length* of the network is the average of l_0 over all pairs of vertices, denoted by $\bar{l}_0$.

Corresponding to the path length, the *effective distance* is defined. For any path between nodes s and t as $P\{s\rightarrow t\}:=s\equiv i_0,i_1,\ldots,i_{n-1},i_n\equiv t$, the effective distance is defined as

$$d_{eff}(P\{s\rightarrow t\})=\sum_{k=1}^{n} w_{i_{k-1},i_k}, \tag{1}$$

where w_{i_{k-1},i_k} is the link weight. The shortest effective path between nodes s and t is the path that has the minimum effective distance $d_{eff}^{min}(s,t)$. When all the link's weights equal 1, the minimum effective distance coincides with the shortest path length. Corresponding to the shortest path length, we define the *shortest effective path length* as the number of links traversed by a shortest effective path, denoted by l, and the *effective length* of the network is the average of l over all pairs of vertices, denoted by $\bar{l}$. Note that Eq. (1) is a general definition, and we will present the detailed method to construct the effective distance in the next section.

To represent the geographical separation between sites, the *manhattan distance* between two nodes n_1 and n_2 is defined as

$$d_M(n_1,n_2)=L-\left|\left|i_1-i_2\right|-\frac{L}{2}\right|-\left|\left|j_1-j_2\right|-\frac{L}{2}\right|, \tag{2}$$

where i_1,j_1 and i_2,j_2 are the x and y coordinates for n_1 and n_2, respectively. The *manhattan route* is the shortest path on the underlying lattice in the absence of any shortcuts.

III. EFFECTIVE DISTANCE AND ROUTING STRATEGY

The effective distance represents a static estimation of the communication cost of routes and is of great importance to network efficiency. When the routing cost is solely determined by the effective distance—i.e., under the static routing protocol (SRP) [32]—network performance can be improved by an optimization on the effective distance. Usually the network capacity is characterized by the critical packet generation rate at the transition from a free-flow to a congestion state of traffic and is inversely proportional to the maximum betweenness of the nodes [13,22], where the betweenness of a node i is defined as

$$B_i=\sum_{s\neq t}\sum_{t\neq i}\frac{\sigma_{st}(i)}{\sigma_{st}}, \tag{3}$$

where $\frac{\sigma_{st}(i)}{\sigma_{st}}$ is the fraction of the the shortest effective paths passing through node i that originate from node s and end at node t. The fraction of the shortest effective paths is calculated by the same method of computing the number of the shortest paths in [33]. To characterize the importance of the nodes, we use the effective betweenness centrality (EBC)

$$\eta_i = \frac{1}{N(N-1)} B_i. \tag{4}$$

The concept of betweenness centrality is first used in describing the importance of people in social networks [34].

Under the SRP, maximizing the network capacity is to minimize the maximum EBC: η_{max}. Danlia *et al.* [15] proposed a deterministic optimization algorithm to find the optimal SRP by adding weight 1 to all the links of those nodes with maximum EBC iteratively. However, finding the optimal d_{eff} that yields the minimum η_{max} is time consuming and uneconomical when the system size is large, and the globally optimized solution is almost impossible to be obtained.

In order to avoid such optimization process, we first develop a simple method of assigning link weight to lower the maximum EBC of the network, then linearly combine the resulting effective distance with nodes' queue length to form the routing cost. The resulting ERS can be regard as a kind of dynamic routing protocol (DRP). The dynamical routing cost enables packets to detour from the original routes when heavy congestion is encountered and thus balances the traffic in the network. Since the mean packet arrival rate of a node is proportional to its EBC [22] under the SRP, if the maximum EBC of the network is lowered and the distribution of EBC becomes more homogeneous, traffic will be easier to distribute more uniformly in the network. So a good static effective distance is also expected to perform well under the DRP, and the aim of our method of assigning link weight is to reduce the η_{max} as much as possible.

The details of the strategy and the dynamical processes are listed as follows.

(i) Weight assignment.

(a) Initialization: assign weight 1 to all the edges; calculate $\overline{l_0}$.

(b) Weight updating. For each edge of each congestible node, if the other end of the link is not a congestible node, then assign new weight a to this link; if the other end of the link is also a congestible node, then the link's weight is updated to $2a-1$. Calculate the effective distance, and find out the minimum effective distance between all pairs of vertices. The form of a is

$$a = c\overline{l_0}, \tag{5}$$

where parameter c is a constant and its value will be given later.

(ii) At each time step, each node has the same probability (or packet generation rate, denoted by λ) to create a new packet with a randomly chosen destination. Every node maintains an unlimited queue which is FIFO (first-in first-out), and newly generated packets are appended to the tail of the queue.

(iii)At each time step, a node s picks one packet in front of its queue that should be delivered to node t. If t is a neighbor of s, then send the packet to t; otherwise, compute the cost C_i for neighbor i of node s:

$$C_i = (1-\beta) d_{eff}^{min}(i,t) + \beta q_i, \tag{6}$$

where q_i is the queue length of node i and the coefficient β is a constant. Then find the neighbor node with the minimum C_i; if there is more than one qualified node, pick one of them at random as the next hop node; add the current packet to its queue and remove the packet in s. The forwarding capability of all nodes is set to 1 for simplicity.

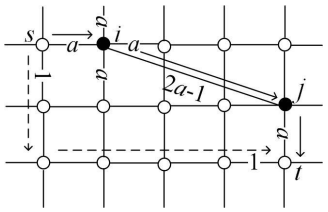

FIG. 2. Illustration of the weight assigning scheme and congestion-relieving mechanism of our strategy. Solid arrows stand for the shortest path, and dotted arrows represent one of the manhattan routes.

(iv) The newly generated or arrived packets, if their destination is exactly the current node, are received by this node and removed from the network immediately.

(v) Steps (ii)–(iv) are updated parallel in time for all the nodes in the network.

The aim of weight assignment is to reduce the importance of the congestible nodes by increasing the weight of the links attached to them. Since the random links greatly reduce the distance between vertices and concentrate a large portion of shortest paths, congestible nodes are the most susceptible to jamming. As shown in Fig. 2, assume a packet at node s should be delivered to node t. There are two kinds of possible routes: one is the shortest path and the other is the manhattan route. Consider light traffic in which the queue length information can be neglected. If step (i) is not applied, the packet will definitely take the shortest path; otherwise, if the value of a makes the total weight of the shortest path larger than $d_M(s,t)$, the packet will be diverted to a manhattan route. For the whole network, given a properly selected value of a, a considerable portion of packets would change their choice from the shortest paths to manhattan routes. In this way the importance of the congestible nodes is decreased, and when traffic grows, they are not susceptible to jamming any more.

In the weight assignment scheme above, all the congestible nodes are treated equally, which can be regarded as a zeroth-order approximation to their importance. The importance of the congestible nodes is closely related to the geographical range of the shortcuts connecting them, which is defined as the manhattan distance between the end nods of a shortcut. If the geographical range of a shortcut is large, the number of shortest paths that pass through it is also large and so does the importance of the congestible nodes at the end of the shortcut. Since shortcuts are randomly distributed on the network and their geographical ranges are also random, we neglect the details of shortcuts for simplicity. It is straightforward that the weight of the link that connects two congestible nodes equals $2a-1$: consider the present model as a weighted network. Assume the other nodes' weights all equal to 1, and it is usual to adopt the following relation between the node weight and link weight: $w_{l_{ij}} = \frac{w_i + w_j}{2}$ (∗), where i and j denote two nodes and l_{ij} is the link between them. Particularly, in Fig. 2, we let i and j represent two congestible

nodes. Since $w_s=w_t=1$ and $w_{l_{si}}=w_{l_{jt}}=a$, in order to satisfy (*), w_i and w_j have to be $2a-1$. Then using (*) again, we get $w_{l_{ij}}=2a-1$. Figure 2 shows a general scene of weight attribution.

Now we explain the chosen form of parameter a in Eq. (5). Since our motivation is to reduce the maximum EBC of the network, the following explanations are based on a static viewpoint; i.e., the queue length is not taken into account. First, consider the determinants of a. As mentioned above, the function of a is to reduce the importance of the congestible nodes and increase that of the low-EBC nodes by avoiding the shortest paths. Most of the shortest paths pass through congestible nodes, and a few of them pass through the low-EBC nodes. So at the point which η_{max} reaches minimum, the portion of avoided shortest paths can be roughly estimated to be $1/2$ and is independent of the topology. So a is exclusively determined by the network topology—i.e., L and p. Second, consider the trend of a with respect to L and p. If p is fixed and L is increased, since the average geographical range of shortcuts (equals $L/2$) becomes larger, shortcuts concentrate more shortest paths and the average importance of congestible nodes increases. So the value of a needs to be increased to counteract the impact of shortcuts on the network. Otherwise, if L is fixed and p is increased, the average geographical range of shortcuts remains unchanged, but since the number of shortcuts (so does the number of congestible nodes) increases, the average importance of one congestible node decreases. So the value of a needs to be decreased. Third, the above tendency should be viewed integrally for the whole network. For our SWN model, the overall impact of shortcuts on the topology of a given lattice can be reflected by the length $\bar{l}_0$ of the network and the dependence of $\bar{l}_0$ on L and p is as follows [26]:

$$\bar{l}_0(L,p) \sim \begin{cases} \dfrac{L}{2}, & \sqrt{p}L \ll 1, \quad (7a) \\ \dfrac{\ln(\sqrt{p}L)}{\sqrt{p}}, & \sqrt{p}L \gg 1. \quad (7b) \end{cases}$$

Since a is a quantity that adapts to the impact of shortcuts and, when $\sqrt{p}L \gg 1$, the above tendency coincides with that of $\bar{l}_0$ to L and p, we adopt the proportional form of a to $\bar{l}_0$ for simplicity—i.e., $a=c\bar{l}_0$.

As to the coefficient c, we determine it by seeking the point at which η_{max} reaches the minimum. We simulate the trend of η_{max} changing with c for different L and p presented in Fig. 3; to be explicit, we only show the envelope of the curves. The average curve for the data sets is shown; η_{max} is normalized by $1/(2L)$, which is the EBC of an arbitrary node of the underlying network in the absence of shortcuts. For all the L and p presented, the optimal value of c for the extremal low points of η_{max} is 0.4 ± 0.1. Moreover, $c=0.4$ is also very close to the minimum point of the average curve. So for simplicity we choose $c=0.4$ for the ERS. Note that $a>1$ when $\bar{l}_0>2.5$ which holds in most cases.

To gain insight into the effect of the weight assignment, we demonstrate the change of nodes' EBC before and after applying step (i). Figures 4(a) and 4(c) are the gray maps of

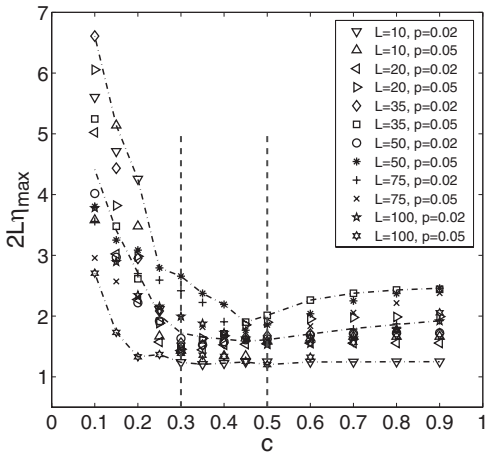

FIG. 3. Finding optimal c under different L and p. Upper and lower curves: the envelope for the curves of the trend of η_{max} changing with c under different L and p. Middle curve: the average of all the data sets.

EBC for all nodes of the original networks; Figs. 4(b) and 4(d) stand for the networks after applying the effective distance. Each gray square represents a node; the bright color represents the nodes with large EBC (mainly congestible nodes), while dark means the opposite. When the effective distance is applied, η_{max} is reduced by several times, so is σ_η, the standard variance of EBC. This also indicates the distribution of EBC becomes more homogeneous.

The dynamical part of the routing cost is also crucial to the performance of the network. The parameter β in Eq. (6) determines how much dynamical information is incorporated in the routing process. In general, a small portion of dynamical information is enough. A similar strategy, the *determinis-*

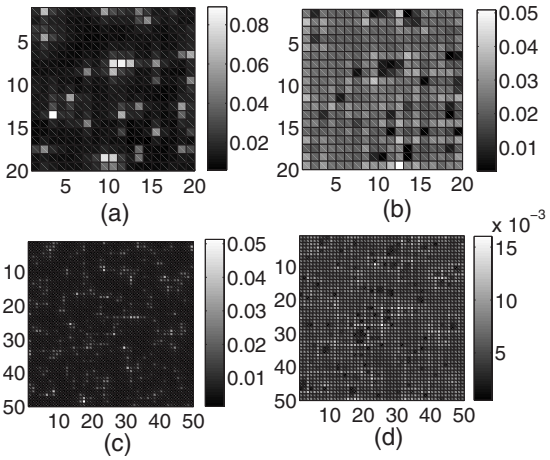

FIG. 4. (a),(c) The gray map of each node's EBC value before applying step (i); (b),(d) the gray map of each node's EBC value after applying step (i). The upper figures are for $L=20$, $p=0.05$ and the lower ones are for $L=50$, $p=0.05$. For (a), (b), (c), and (d), respectively, the values of η_{max} are 0.09, 0.051, 0.052, and 0.016; η_{avg} are 0.018, 0.023, 0.0044, and 0.0064; the average EBC of congestible nodes are 0.054, 0.007, 0.016, and 0.004; σ_η are 0.014, 0.007, 0.0046, and 0.002. The average distance $\bar{l}_0$ is 7.14 and 10.95 for (a) and (c); the effective length $\bar{l}$ is 9.2 and 16 for (b) and (d).

tic protocol, has been proposed by Echenique *et al.* in [17]. The routing cost is $d_{eff}^{i}=C_i=h_d d_i+(1-h_d)c_i$, where d_i is the shortest path length between node i and destination and $c_i=q_i$. It is reported that the optimal value of h_d is 0.75. Correspondingly, $\beta=0.25$. Following their results, we fix $\beta=0.25$ for the ERS throughout this paper.

The ERS helps the network to function more efficiently. First, under light traffic, the queue length can be neglected and the mean packet arrival rate of a node is nearly proportional to its EBC. Since the maximum EBC is decreased, the network capacity is increased. Second, under heavy traffic, congestion can be further alleviated due to the queue length information. When packets accumulate in the large-EBC nodes, the succeeding packets have smaller probability to flow into these nodes. In this way the mean packet arrival rates of large-EBC nodes decrease and those of the low-EBC nodes increase. In this case, the weight assignment is very helpful for traffic to distribute uniformly, because the difference of the initial mean packet arrival rate between nodes is reduced due to compression of the EBC distribution. So the capacity of the network increases again. Finally, from Figs. 4(c) and 4(d), the effective length $\bar{l}$ has not increased much compared with $\bar{l}_0$ of the original graph, and $\bar{l}$ is also smaller than $L/2$. That is because in the ERS, only part of the packets makes a detour to the manhattan route. Since under the SRP the average travel time of packets is proportional to $\bar{l}$ when λ approaches zero [22], the time delay performance of packets is expected to be low under ERS.

IV. SIMULATIONS AND EXPLANATIONS

In this section we investigate the performance of the ERS. The efficiency of the network is reflected by both the network capacity and the average travel time of packets. The network capacity is represented by the critical packet generation rate λ_c, which is the transition point from a free-flow traffic state to a congested state. The phase transition is characterized by the order parameter presented in [23]:

$$\xi(\lambda)=\lim_{t\to\infty}\frac{1}{\lambda N}\frac{\langle\Delta W\rangle}{\Delta t}, \tag{8}$$

where $W(t)$ is defined as the number of total accumulated packets in the network at time t. $\Delta W=W(t+\Delta t)-W(t)$, and $\langle\cdots\rangle$ indicates an average over time windows of width Δt. This quantity represents the ratio between undelivered and generated packets over long enough time periods. When $\lambda<\lambda_c$, ξ tends to zero; when $\lambda>\lambda_c$, the network enters a congestion state, the number of accumulated packets grows linearly in time, and ξ equals a finite value larger than 0 and smaller than 1. We set the terminating time step as 50 000 for all the following simulations and discard the data in the first 30 000 steps. The decision threshold of $\xi(\lambda)$ for congestion is set to be 0.001.

For comparison with previous strategies, we use the Manhattan routing strategy (MRS) as the first reference. It is defined as follows [27]: for a packet at node s whose destination is node t, it first selects any node i that has the minimum manhattan distance $d_M(i,t)$ from the neighbor set $\mathcal{N}$ and constructs a new set $\mathcal{A}$; then, from $\mathcal{A}$ it chooses any node that has the minimum queue length as set $\mathcal{B}$. The next hop node is selected randomly from $\mathcal{B}$. Since the MRS considers the queue length, it is also a kind of DRP.

The *deterministic protocol* [17] (DP) is used as the second reference. The routing cost of a node i linearly combines the shortest path length and queue length: $C_i=0.75d(i,t)+0.25q_i$, where t is the destination. If the dynamical part is abandoned, it is restored to the shortest path (SP) routing. For convenience, we name the routing strategy in which the routing cost linearly combines a distance metric and the node queue length as the *dynamic counterpart* of the one in which the routing cost is only the distance metric; and conversely, the latter is the *static counterpart* of the former.

Another reference strategy is based on the "efficient path" (EP), recently proposed by Yan *et al.* [16]. In this routing strategy, the routing cost for a path $P\{i\to j\}:=i\equiv x_0,x_1,\ldots,x_{n-1},x_n\equiv j$ between nodes i and j is $L(P(i\to j):\alpha)=\sum_{i=0}^{n-1}k(x_i)^{\alpha}$, where $k(x_i)$ is the degree of node x_i. The efficient path between i and j is the path that minimizes the above cost. It is reported that the optimal value of α equals 1 for scale-free Barabasi-Albert (BA) networks. This is a static strategy; in order to facilitate comparison, we modify the routing cost to be dynamical in the same way of this paper: $C_i=0.75L_{min}^{\alpha=1}(i,t)+0.25q_i$. Since the optimal α in SWNs is unknown, we adopt $\alpha=1$ and name this strategy the *dynamical efficient path* (DEP). Because the only difference between the routing cost of the ERS, DP, and DEP is the static part, comparing with the other two can explicate the advantage of the weight assignment scheme in ERS.

Since it is of great importance to know the limit of our effort to improve the capacity, we also study the superior bound of λ_c. Under any SRP, λ_c reaches a maximum when η_{max} achieves a minimum [22]:

$$\lambda_c=\frac{1}{N\eta_{max}}. \tag{9}$$

Taking into account that $\eta_{max}\geqslant\eta_{avg}$ and the absolute lower bound of η_{avg} is the average EBC of the network under shortest path routing (because any changes from the shortest paths will result in longer $\bar{l}$ and $\eta_{avg}\equiv\frac{\sum_{i=1}^{N}B_i}{N^2(N-1)}=\frac{\bar{l}}{N}$ [13]), we get

$$\lambda_c\leqslant\frac{1}{\sum_i\eta_i}=\frac{1}{\bar{l}}. \tag{10}$$

Noting that $\bar{l}>\bar{l}_0$, we get the superior bound of network capacity (λ_{c_1}) for a given network under the SRP:

$$\sup\lambda_{c_1}=\frac{1}{\bar{l}_0}. \tag{11}$$

Now we generalize this to an arbitrary network and routing strategy. Inequality (10) can be rewritten as $\lambda_c\frac{\sum_i B_i}{N-1}=\sum_i\frac{B_i}{N-1}\leqslant N$, where the left part corresponds to the sum of the packet arrival rates of all nodes and the right part is the sum of the outflowing rates (equal to 1) of all nodes. Regarding

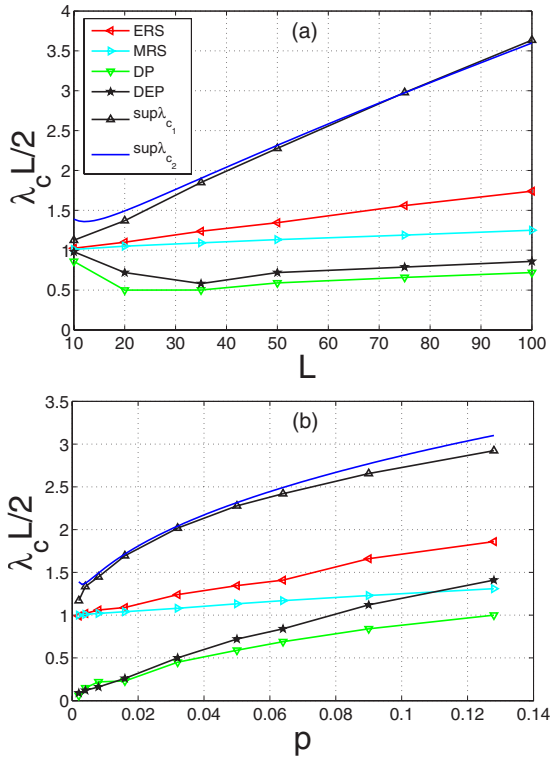

FIG. 5. (Color online) Comparison results of the normalized critical packet generation rates. (a) $p=0.05$, $L=10,20,35,50,75,100$; (b) $L=50$, $p=0.002-0.128$. Legend is the same as in (a); $\sup \lambda_{c_1}$ is calculated from simulation data of $\bar{l}_0$.

the network as a unit queuing system, from queuing theory we know the input rate must be less than output rate for a stable queue; so for any network that enters the stationary state and is in a free-flow phase, this relation holds. Now that Eq. (10) is universal for all network and routing strategies, Eq. (11) is also universal. So we use Eq. (11) as a reference for the ERS.

The above superior bound only corresponds to a given network topology; in order to give a theoretical approximation of the superior bound of network capacity (λ_{c_2}) for the ensemble of the network model in the small-world regime, we combine Eq. (11) and Eq. (7b), which yields the third reference

$$\sup \lambda_{c_2} = \frac{\sqrt{p}}{\ln(\sqrt{p}L)}. \tag{12}$$

Figure 5 summarizes the results of the simulated network capacity of the ERS, MRS, DP, and DEP and the superior bounds. We give the results of λ_c for different L and p and the upper bounds of λ_c. We use the λ_{c_0} for shortest path routing on the underlying lattice as a normalization factor. In this case, the network is completely homogeneous; every node's EBC is the same and $\bar{l}=\bar{l}_0$. Since $\bar{l}_0=L/2$, we have $\lambda_{c_0}=\sup \lambda_{c_1}=2/L$. From Fig. 5, it is clear that the network capacity of the ERS is larger than that of the other strategies for almost all L and p presented.

The performance of the ERS on network capacity can be attributed to two factors: the weight assignment scheme and the integration of queue length information. The former reduces the maximum EBC and homogenizes the distribution of EBC, and the latter further balances the dynamic traffic load on the network.

To understand the effects of the above factors on the performance of the ERS, we must compare the traffic load between the ERS and its static counterpart (set $\beta=0$ and we denote it as SERS). In the static routing literature, betweenness centrality is often termed as load [35]; however, when the routing strategy is dynamical, BC has no definition. So we introduce *dynamic load* (l_d) to characterize the traffic load of nodes under a dynamical routing protocol and call BC the static load (l_s) here to avoid confusion. Notice that when the network is in a steady state ($\lambda<\lambda_c$) and the arrival and delivery processes are Poisson, the relationship between the average queue length Q_i of a node i and its mean packet arrival rate μ_i is [22] $Q_i=\frac{\mu_i}{1-\mu_i}$. Under any SRP, the mean packet arrival rate is proportional to the node's betweenness centrality: $\mu_i=\lambda \eta_i N$. But under the DRP, the above relationship does not hold and μ_i can be derived from Q_i, which is a measurable quantity: $\mu_i=\frac{Q_i}{1+Q_i}$. Normalized by λN (the average number of packets generated in the whole network in unit time step), we get

$$l_{d_i}(\lambda) = \frac{\mu_i}{\lambda N} = \frac{Q_i}{\lambda N(1+Q_i)}, \quad \lambda < \lambda_c, \tag{13}$$

where l_{d_i} represents the number of packets that flow into the queue of node i on average for each packet generated in the network. For a given network topology and routing strategy, this quantity is solely determined by λ. As $\lambda \to 0$, queue length can be neglected; $\mu_i \to \lambda \eta_i N$, and this quantity is restored to the static load: $l_{d_i} \to l_{s_i}$. Also, for the static counterpart strategy $l_{d_i}=\eta_i=l_{s_i}$. Thus we are able to compare the "load" between static routing strategies and dynamical ones under various λ.

The comparison result of static load and dynamic load between the SP, DEP, and ERS is presented in Fig. 6. As one can see, for shortest path routing (SP) the static load distribution shows the combined form of two Poisson-type decays, resulting from short-range and long-range links [35]. For the SERS, the contribution from long-range links disappears and the horizontal range of static load distribution is compressed, which are the outcomes of the weight assignment. For the ERS, when the traffic generation rate is low ($\lambda=0.1\lambda_c$) the dynamic load coincides with the static load of the SERS, which means most of the packets still adopt the shortest effective paths; under heavy traffic ($\lambda=0.95\lambda_c$), the occurrence of an intermediate dynamic load is dramatically increased with the high load part missing. This indicates that the queue length information in the routing cost makes the dynamic load become more homogeneous.

Though the queue length information is very important, the following shows that the weight assignment is comparatively more pivotal to the high network capacity of the ERS, which is closely related to the degree of homogeneity of the

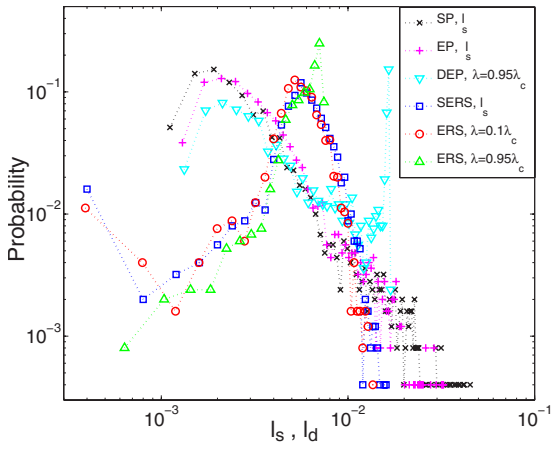

FIG. 6. (Color online) The distribution of static load and dynamic load for different strategies and various packet generation rates. $L=50$, $p=0.05$.

static load distribution of its static counterpart. In general, if the latter is more homogeneous, then it is easier for the dynamic load to distribute more uniformly and to gain larger capacity. For example, in Fig. 6, both the load distributions of the ERS and that of the SERS under heavy traffic are much less heterogeneous than those of the DEP and EP, and the small-load part of the ERS and DEP resembles that of the SERS and EP. This can be seen as the inertia of the packets: the redirection of traffic to the small-load nodes only happens when it is necessary under extremely heavy traffic, or else the small-load nodes do not bear much more load. So if few nodes bear small static load under a SRP, they are more likely to be fully utilized under heavy traffic for the corresponding DRP. Obviously this is the case for the ERS.

Note that the capacity performances for the DEP and DP are close. That is because for the "efficient path" strategy, the weight of a node is proportional to its degree. In our SWN model, the distribution of node degree follows an exponential decay which means that the degrees of congestible nodes are small and are near to those of the other nodes. Therefore the weights of the links that are attached to congestible nodes are close to that of the other links, while in the SP all the link weights equal 1. Since the performance of a DRP is largely determined by the weight assignment of its static counterpart, it is natural that DEP and DP performs similarly.

Figure 5 also shows the predominance of the ERS over the MRS. The relatively high network capacity for the MRS is mainly due to the homogeneity of the underlying regular lattice. If shortcuts are absent, the static load of every node is the same; now that the shortcuts are randomly distributed on the lattice and thus the underlying graph does not become more heterogeneous, the sites other than congestible nodes are equally utilized in the MRS. However, the MRS merely finds the node on the manhattan path that has the smallest queue and there is no possibility for packets to take a roundabout path. In contrast, the ERS enables packets to go around the jamming nodes.

It is worthy to note that the ERS performs relatively better when the network size N and p grow; i.e., it has good scalability. In Figs. 5(a) and 5(b), when pN is small, the ERS

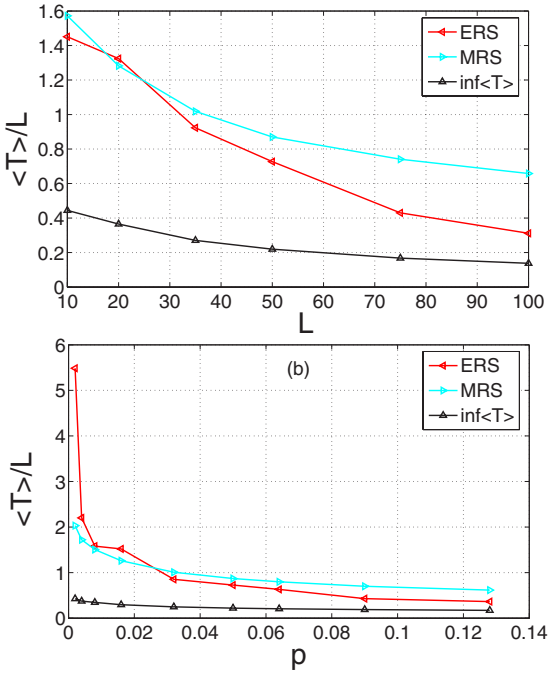

FIG. 7. (Color online) Normalized average travel time of packets for different L and p and the inferior bound, $\lambda=2/L$. (a) $p=0.05$, $L=10,20,35,50,75,100$; (b) $L=50$, $p=0.002-0.128$.

does not perform much better than the MRS, or even a little worse when $pN=5$, but when $pN \gg 1$, as p or L increases, the difference between the network capacity of the ERS and that of the MRS grows larger. In fact, the ERS works well when the SWN model in this paper is indeed a small world, since the basic form of the assigned link weight of shortcuts, $\bar{l}_0$ [Eq. (7b)] holds under the small-world condition $\sqrt{pN} \gg 1$. The network is indeed a small world when $\sqrt{pN} \gg 1$; this is confirmed by the coincidence of the theoretic upper bounds 1 and 2 in Figs. 5(a) and 5(b).

Now we consider another performance indicator: the average travel time of packets, $\langle T \rangle$. We compare the average travel time of packets under the same packet generation rate $\lambda=2/L$ between the ERS and MRS. The inferior bound to $\langle T \rangle$ is also evaluated numerically, $\inf\langle T \rangle = \bar{l}_0$, since no path can be shorter than the shortest path. Results are shown in Fig. 7. In Figs. 7(a) and 7(b), when L and p are small, the difference between the time delays is also small, but when L and p grow large, the ERS begins to show its efficiency. That is because with the increase of L and p, λ_c increases which makes $\lambda=2/L$ fall far below the critical value and the linear relationship of $\langle T \rangle$ to $\bar{l}$ begins to take effect, as stated in Sec. III. Under the MRS, packets have little chance to use the shortcuts and most of them follow manhattan routes on the underlying lattice, so the average number of hops from source to destination (equivalent to the effective length) is close to $L/2$. But in the ERS, only about half of the packets deviate from the shortest paths, so $\bar{l}$ is much smaller than $L/2$. The sharp increase of $\langle T \rangle$ when $p \to 0$ is because the network enters a congested state under the ERS at $\lambda=2/L$. In addition, the difference between $\langle T \rangle$ of the ERS and the inferior bound is decreased as L and p increases, which indi-

cates that the ERS is time efficient in the small-world regime.

V. CONCLUSION AND DISCUSSION

Small-world network models have great impact on the research of the structure and dynamics of complex networks [36]. Based on a two-dimensional SWN model, we have proposed an efficient routing strategy that incorporates global static effective distance and local queue length information. The efficiency of the ERS lies in both enhancing the network capacity and lowering the average travel time of packets. The effective distance, which is the most important for the proper function of the ERS, is constructed from a proper assignment of the weight of the links attached to the congestible nodes. The static load distribution of the network is homogenized due to the weight assignment and thus provides strong basis for the dynamic load to distribute uniformly. We have tested the ERS with different network sizes up to $N=10\,000$ and various p; it is shown that the ERS is more efficient than the reference strategies when the network is indeed a small-world one.

The weight assignment scheme in the ERS is simple; it takes into account the characteristics of network topology and only considers a few congestible nodes that are the most "influential" on the network performance. Though the ERS is applicable only to the SWN in this paper, we believe that the idea here might inspire new routing strategies on other complex network models, like scale-free networks. For example, corresponding to the congestible nodes here, one may identify a few "hub nodes" that have largest degrees and assign weights according to the topological properties of them. A noteworthy strategy on this route is called *hub avoidance* (HA) [32] for scale-free networks. It first removes a few hub nodes of the network and uses the shortest path routing for the nodes in each remaining connected cluster, then puts these hub nodes back along with their edges and assigns routes using SP for the pairs of vertices that are disconnected when those hub nodes are absent. In terms of weight assignment, this is equivalent to assigning infinite weight to those hub nodes: while they appear to be nonexistent for the vertex pairs in the first round of SP routing, the hub nodes are on the unique paths for the remaining unrouted pairs of vertices. HA is proved to perform well in the scale-free network model [21,32]. One may wonder whether the network performance may improve further when the hub nodes in scale-free networks are given finite weights.

In addition, the ERS can be easily implemented because the main computing cost is the calculation of network length and minimum effective distances before applying the routing table. There are many efficient algorithms for the shortest path problem [37], and it can be done in a distributive manner [38]. Moreover, the additional form of routing cost is simple enough to be computed.

To provide an outlook, the maximum capacity of the ERS is still far from the theoretical limit, which indicates that there is still some work left to do. In our paper, all congestible nodes are treated equally; to find more efficient ways of enhancing the network efficiency, future work may involve the dependence of link weight on the importance of the shortcuts or congestible nodes.

ACKNOWLEDGMENTS

The authors wish to thank the CNS/ATM laboratory, CAAC, for providing the experimenting devices and research environment. Useful discussions with Han Zhao are appreciated. This work was supported by the National Natural Sciences Foundation of China under Grant No. 60502017, the National Basic Research Program of China under Grant No. 2006CB303000, and a Blue-Sky New Star Grant of Beihang University (2004).

[1] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[2] H. Jeong, B. Tombor, R. Albert, Z. N. Oltavi, and A.-L. Barabási, Nature (London) **407**, 651 (2000).
[3] V. Latora and M. Marchiori, Phys. Rev. Lett. **87**, 198701 (2001).
[4] L. A. Adamic, Lect. Notes Comput. Sci. **1696**, 443 (1999).
[5] M. Faloutsos, P. Faloutsos, and C. Faloutsos, Comput. Commun. Rev. **29**, 251 (1999).
[6] S. N. Dorogovtsev, J. F. F. Mendes, and C. Faloutsos, Adv. Phys. **51**, 1079 (2002).
[7] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[8] M. E. J. Newman, J. Stat. Phys. **101**, 819 (2000).
[9] I. Vragović, E. Louis, and A. Díaz-Guilera, Phys. Rev. E **71**, 036122 (2005).
[10] J. M. Kleinberg, Nature (London) **406**, 845 (2000).
[11] H. Zhu and Z.-X. Huang, Phys. Rev. E **70**, 036117 (2004).
[12] A. P. S. de Moura, A. E. Motter, and C. Grebogi, Phys. Rev. E **68**, 036106 (2003).
[13] L. Zhao, Y.-C. Lai, K. Park, and N. Ye, Phys. Rev. E **71**, 026125 (2005).
[14] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **401**, 130 (1999).
[15] B. Danila, Y. Yu, J. A. Marsh, and K. E. Bassler, Phys. Rev. E **74**, 046106 (2006).
[16] G. Yan, T. Zhou, B. Hu, Z.-Q. Fu, and B.-H. Wang, Phys. Rev. E **73**, 046108 (2006).
[17] P. Echenique, J. Gómez-Gardeñes, and Y. Moreno, Phys. Rev. E **70**, 056105 (2004).
[18] P. Echenique, J. Gómez-Gardeñes, and Y. Moreno, Europhys. Lett. **71**, 325 (2005).
[19] Z. Y. Chen and X. F. Wang, Phys. Rev. E **73**, 036107 (2006).
[20] W.-X. Wang, C.-Y. Yin, G. Yan, and B.-H. Wang, Phys. Rev. E **74**, 016101 (2006).
[21] B. Danila, Y. Yu, S. Earl, J. A. Marsh, Z. Roroczkai, and K. E. Bassler, e-print arXiv:cond-mat/0603861.
[22] R. Guimerà, A. Díaz-Guilera, F. Vega-Redondo, A. Cabrales, and A. Arenas, Phys. Rev. Lett. **89**, 248701 (2002).

[23] A. Arenas, A. Díaz-Guilera, and R. Guimerà, Phys. Rev. Lett. **86**, 3196 (2001).
[24] D. J. Ashton, T. C. Jarrett, and N. F. Johnson, Phys. Rev. Lett. **94**, 058701 (2005).
[25] T. C. Jarrett, D. J. Ashton, M. Fricker, and N. F. Johnson, Phys. Rev. E **74**, 026116 (2006).
[26] M. E. J. Newman and D. J. Watts, Phys. Rev. E **60**, 7332 (1999).
[27] H. Fukś and A. T. Lawniczak, Math. Comput. Simul. **51**, 103 (1999).
[28] H. Fukś, A. T. Lawniczak, and S. Volkov, ACM Trans. Model. Comput. Simul. **11**, 233 (2001).
[29] B. K. Singh and N. Gupte, Phys. Rev. E **71**, 055103(R) (2005).
[30] Z.-H. Liu, W.-C. Ma, H. Zhang, Y. Sun, and P. M. Hui, Physica A **370**, 843 (2006).
[31] R. V. Solé and S. Valverde, Physica A **289**, 595 (2001).
[32] S. Sreenivasan, R. Cohen, E. López, Z. Toroczkai, and H. E. Stanley, Phys. Rev. E **75**, 036105 (2007).
[33] M. E. J. Newman, Phys. Rev. E **64**, 016132 (2001).
[34] L. C. Freeman, Sociometry **40**, 35 (1977).
[35] K.-I. Goh, B. Kahng, and D. Kim, Phys. Rev. Lett. **87**, 278701 (2001).
[36] M. E. J. Newman, SIAM J. Appl. Math. **45**(2), 167 (2003).
[37] U. Brandes, J. Math. Sociol. **25**(2), 163 (2001).
[38] K. A. Lehmann and M. Kaufmann (unpublished).