SCIENTIFIC REPORTS

OPEN

Percolation-theoretic bounds on the cache size of nodes in mobile opportunistic networks

Peiyan Yuan [1], Honghai Wu[2], Xiaoyan Zhao[3] & Zhengnan Dong[1]

The node buffer size has a large influence on the performance of Mobile Opportunistic Networks (MONs). This is mainly because each node should temporarily cache packets to deal with the intermittently connected links. In this paper, we study fundamental bounds on node buffer size below which the network system can not achieve the expected performance such as the transmission delay and packet delivery ratio. Given the condition that each link has the same probability p to be active in the next time slot when the link is inactive and q to be inactive when the link is active, there exists a critical value p_c from a percolation perspective. If $p > p_c$, the network is in the supercritical case, where we found that there is an achievable upper bound on the buffer size of nodes, independent of the inactive probability q. When $p < p_c$, the network is in the subcritical case, and there exists a closed-form solution for buffer occupation, which is independent of the size of the network.

Current portable/wearable devices have been integrated with many sensors and wireless functions, such as WiFi and Bluetooth, which confers on them powerful capabilities, especially in sensing, computing and communication. The more powerful these devices become, the more likely it is that they share contents locally, leading to the emergence of mobile opportunistic networks (MONs)[1]. MONs use node mobility to provide occasional contact opportunities for mobile devices to transmit packets. Such a new networking paradigm employs device-to-device communication, which is rarely considered in traditional infrastructure-based networks, to alleviate the over-loading criticality and enhance the coverage areas. On the other hand, in infrastructure-less areas, MONs may be the only communication solution. These benefits enable numerous applications, including rural communications[2], wild monitoring[3], crowd sensing[4] and metropolitan awareness of issues[5].

Compared to traditional wireless networks, one of the features of MONs is that the intermittent connectivity is mainly caused by broken links due to some external factors varying in different scenarios. For example, in an ad hoc mobile vehicle network[6], two vehicles are connected when they enter the communication range of each other, and the link between them is disconnected when they move outside the transmission range. In[7], S. Wang, J. Zhao and L. Tong studied a cognitive radio network where the activity of primary users is driven by an ON-OFF process such that the secondary users should keep inactive until the primary users switch from the ON to OFF mode.

The factors mentioned above are external in some sense, which implies that link disconnection and inactivity can not be eliminated by improving internal conditions such as enhancing the network capacity or node process speed. With external constraints, node buffer size can not approach zero even if the network capacity approaches infinity, since each node should temporarily cache packets to deal with the broken links. Therefore, there exists a performance limit on node buffer occupation. On the other hand, current research communities mainly focus on the data forwarding schemes of MONs (e.g., PROPHET[8], Bubble[9], Spray-Wait[10], SMART[11] and Hotent[12]). Some analytical results with respect to flooding time[13], network diameter[14], and delay-capacity tradeoff[15] have also been obtained, providing valuable insights into data forwarding efficiency and system performance. However, such insights may only reflect one aspect of MONs without considering the influence of node buffer size.

In this paper, we study node buffer occupation in MONs, where broken links are caused by node mobility or a channel access scheme. We assume that the system capacity can be regarded as infinite, compared to the low packet generation rate. By using this assumption, we try to illustrate that even if the system capacity is unlimited,

[1]Henan Normal University, School of Computer and Information Engineering, Xinxiang, 453000, China. [2]Henan University of Science and Technology, Information Engineering College, Luoyang, 471000, China. [3]National Digital Switching System Engineering and Technological Research Center, 450000, Zhengzhou, China. Honghai Wu, Xiaoyan Zhao and Zhengnan Dong contributed equally to this work. Correspondence and requests for materials should be addressed to P.Y. (email: peiyan@htu.cn)

there still exists a lower bound on node buffer size below which the network system cannot achieve the suggested performance, such as the transmission delay and packet delivery ratio. We use the edge-Markovian dynamic graph to model MONs. This model can reproduce the power law + exponential tail distribution of the pairwise node inter-contact time, which has been observed in several real data traces of MONs[16]. Starting from an arbitrary initial probability distribution of links, at every time slot, links change their states (active or inactive) based on a two-state Markovian process with probabilities p (active) and q (inactive). If a link is inactive at slot t, it will be active at slot $t+1$ with probability p. If, instead, the link is active at slot t, it switches to the inactive state at the next slot with probability q. From the perspective of percolation theory, there is a critical value p_c for p. If $p > p_c$, the network is in the supercritical case, and there exists a connected giant cluster, a unique infinite connected cluster at any time, with a high probability when the number of nodes becomes infinity. In contrast, when $p < p_c$, the network is in the subcritical case, and the giant cluster disappears almost certainly when the number of nodes goes to infinity. Fundamental limits on node buffer size are quite different in the two cases. In the supercritical case, there is an upper bound on the buffer size of nodes, independent of the inactive probability q, while in the subcritical case, there exists a closed-form solution for buffer occupation that is independent of the size of the network.

The rest of this paper are organized as follows. We first review related works and introduce the preliminaries. Then, we analyze the node buffer occupation in the supercritical case and present the results in the subcritical case. After that, we conduct the experiment and analyze the results, followed by a short conclusion.

Related Works

MONs is a natural evolution from traditional mobile ad hoc networks. In MONs, the links are intermittently connected due to node mobility and power on/off, mobile nodes communicate with each other opportunistically and route packets in a store-carry-and-forward style. In the past several years, much effort has been expended to improve the performance of opportunistic routing algorithms in terms of reducing the forwarding delay or increasing the packet delivery ratio. Some valuable results have been achieved that provide theoretical guidance for performance optimization. We introduce them in detail.

Cache-aware Opportunistic Routing Algorithms.

Considering the limited buffer size of portable devices, cache-aware solutions become very important in MONs. A. Balasubramanian *et al.* take MONs routing as a resource allocation problem and turn the forwarding metric into per-packet utilities that incorporate two factors: one is the expected contribution of a packet if it were replicated, and the other is the packet size. The utility, then, is the ratio of the former factor over the latter, which determines how a packet should be replicated in the system[17]. To deal with the short contacts and fragmented bundles, M. J. Pitkanen and J. Ott[18] integrated application level erasure coding on top of existing protocols. They used Reed Solomon codes to divide single bundles into multiple blocks and observed that the block redundancy increased the catch hit rate and reduced the response latency.

S. Kaveevivitchai and H. Esaki proposed a message deletion strategy for a multi-copy routing scheme[19]. They employed extra nodes deployed at the system's hot regions to relay the acknowledgement (ACK) messages, and copies matching the ID of ACK messages are dropped from the buffer. A. T. Prodhan *et al.* proposed TBR[20], which ranks messages with their TTL, hop count and number of copies. A node will delete the copy of a message if it receives a higher priority message and its buffer is full. Recently, D. Pan *et al.*[21] developed a comprehensive cache schedule algorithm that integrates different aspects of storage managements including queue strategy, buffer replacement and redundancy deletion.

Performance Analysis of Cache-aware Opportunistic Routing Algorithms.

Some analytical results mainly focus on metrics such as the flooding time[13, 22, 23], network diameter[14] and delay-capacity tradeoff[15], in which the buffer size of nodes is usually assumed to be unlimited. Several works discuss the congestion issue with the epidemic algorithm. For example, A. Krifa *et al.*[24] proposed an efficient buffer management policy by modeling the relationship between the number of copies and the mean delivery delay/rate. When a new packet copy arrives at a node and the node finds its buffer full, it drops the packets with the minimal marginal utility value. G. Zhang and Y. Liu employed revenue management and dynamic programming to study the congestion management strategy of MONs[25]. Given a class of utility functions, they showed that one arrived packet should be accepted and that the solution is optimal if and only if the value of the benefit function is greater than that of the cost function.

The authors of[26] evaluated the impact of buffer size on the efficiency of four kinds of routing algorithms. They observed that these protocols reacted differently to the increase of buffer size in mobile vehicle networks. Generally speaking, both the Epidemic and MaxProp benefit from the increased buffer size on all nodes (i.e., the mobile and terminal nodes). PROPHET and SprayWait instead have no significant improvement when only the buffer size of terminal nodes increases. X. Zhuo *et al.*[27] explored the influence of contact duration on data forwarding performance. To maximize the delivery rate, they modeled the data replication problem with mixed integer programming technology, subject to the storage constraint.

Preliminaries

External Factors and Broken Links.

In this paper, we study the general influence of external factors on the network connectivity rather than focusing on a specific type of external constraint. At each time slot, each link switches between the inactive state and active state. Two endpoints of a link can transmit or receive packets only if the link is active.

We model the external factors in the network with the edge-Markovian dynamic graph (EMDG), which implies that: (1) States of each link vary from one time slot to another, and are i.i.d. among time slot $t+1$ and the

$$A = \begin{pmatrix} & 0 & 1 \\ \hline 0 & 1-p & p \\ 1 & q & 1-q \end{pmatrix}$$

Figure 1. Transition matrix.

time slots before t. (2) The probabilities of being active or inactive are two constants p and q for all links, respectively. Mathematically, let n denote the number of nodes and $0 \leq p, q \leq 1$. The EMDG can be denoted by $\mathcal{G}(n, p, q, L(t))$ if

$$L(t) = \left\{ l \in \binom{n}{2} : X_t(l) = 1 \right\}$$

where $\left\{ X_t(l) : l \in \binom{n}{2} \right\}$ are independent Markov chains with the transition matrix and l is the edge of EMDG. As shown in Fig. 1, the transition matrix indicates the transition of the link state in the next time slot, where 0 and 1 denote that the link is inactive or active, respectively. For example, the element a_{12} means the link is inactive and will become active in the next time slot with probability p.

Percolation of Connected Links. Considering the possibility of link activity, we can not guarantee a connected path for each source-destination pair at all times. However, since the states of links are i.i.d, there exists a critical probability $p_c = c\frac{\log n}{n}$ such that (please refer to property 3.5 of ref. [28]):

Supercritical case. If $p \geq c\frac{\log n}{n}(c > 1)$, then w.h.p. $\mathcal{G}(n, p, q, L(t))$ is in the supercritical case and there exists a connected giant cluster.

Subcritical case. There is a constant $c \leq 1/2$ such that for $p \leq c\frac{\log n}{n}$, $\mathcal{G}(n, p, q, L(t))$ is in the subcritical case and the giant cluster disappears.

Bounds on Node Buffer Size in the Supercritical Case. In this section, we study the bounds on buffer size of nodes in $\mathcal{G}(n, p, q, L(t))$ when $p \geq c\frac{\log n}{n}$. The main result is that the expected buffer size of nodes can be bounded within a certain range as stated in Theorem 1.

Theorem 1 *For a randomly selected node u of $\mathcal{G}(n, p, q, L(t))$ if $p \geq c\frac{\log n}{n}$, at the end of a time slot t,*

$$E(B_u(t)) \geq rT_d\frac{p+q}{q} \tag{1}$$

And in some forwarding schemes, it is

$$E(B_u(t)) \leq c_1 rT_d\frac{\log n}{pM} \tag{2}$$

where the $E(B_u(t))$ is a random variable that denotes the expectation of the node buffer size, r is the packet generation rate per node, T_d is the duration of one time slot, c_1 is a constant and M is the number of nodes in the giant cluster.

Proof of Inequality (1). Here, we take the node u as an example. Suppose it belongs to the giant cluster, it does not need to buffer packets when the links connecting it are active. This is mainly because there exists a connected path with a high probability between the node u and other nodes; the transmission delay hence tends to zero. Instead, it at least needs to store packets generated by itself when links are inactive. The expected waiting time before links become active is $1/\pi_1$, where π_1 belongs to the stationary distribution π, which is a special distribution for a Markov chain such that if the chain starts with its stationary distribution, the marginal distribution of all states at any time will always be the stationary distribution. The variables i and j denote the rows and columns in the transition matrix above, so as to compute the transition probability of the link stage in the next time slot. π_j refers to π_1 and π_2, which denotes the value of the stationary distribution of a Markov chain. With the transition matrix A, we have:
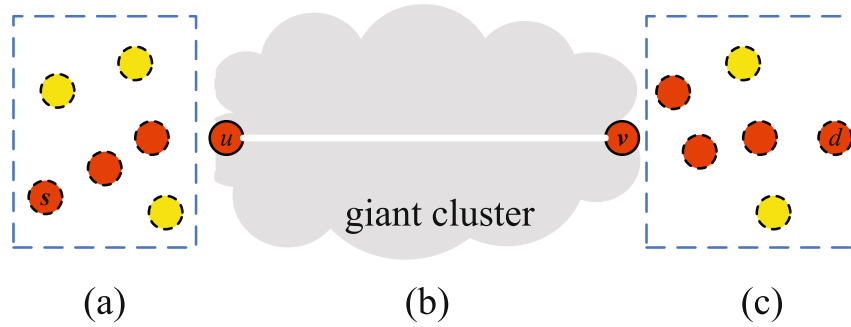
Figure 2. Three stages in the optimal forwarding scheme.

Figure 3. Four types of nodes.

$$\begin{cases} \pi_j = \sum_{i=1}^{2} \pi_i a_{ij}, \, j = 1, \, 2 \\ \pi_j > 0, \, \sum_{j=1}^{2} \pi_j = 1 \end{cases}$$

where a_{ij} denotes elements of A. We obtain $\pi_1 = \frac{q}{p+q}$. Applying Little's Law,

$$E(B_u(t)) \geq rT_d \frac{p+q}{q}$$

According to Little's Law, the long-term average number of customers in a stable system is equal to the long-term average effective arrival rate multiplied by the average time a customer spends in the system. Note that $\frac{p+q}{q}$ is the expected number of time slots before links become active, $T_d\frac{p+q}{q}$ is the average time a customer spends in the system, and r is the arrival rate. Hence, we obtain the number of packets generated by u with equation (1).

In the following subsections, we first design an optimal forwarding scheme, and then prove that with this scheme, the buffer occupation specified in equation (2) can be achieved.

Optimal Forwarding Scheme (OFS). Nodes in the supercritical case can be classified into two types based on whether they belong to the giant cluster. One type is connected nodes, which are in the giant cluster, and the other type is disconnected nodes. The former constitutes a connected component and the packets can be transmitted fast. In contrast, the disconnected nodes need larger buffer sizes than the connected nodes since they have to store packets before the links turn active. The worst case is that both the source s and the destination d are disconnected nodes and are just separated by the giant cluster, as shown in Fig. 2. There exist three stages in the optimal forwarding scheme: the source flooding stage, the shortest path stage and the destination flooding stage. We first introduce the source flooding stage.

(1) The source flooding stage. In general, the flooding scheme achieves the highest packet delivery ratio and the lowest transmission delay, but it requires the largest storage space, which is necessary for our study, since we focus on the fundamental bounds on node buffer size especially in the worst case. Considering that source s and its neighbors are disconnected nodes, it needs to flood packets generated by itself and received from its neighbors. In the flooding scheme, if there is a packet at node u at time slot $t-1$, node u will send the packet to all the nodes that are connected to u at time slot t if the link between them is active. To characterize the flooding process, we use an infectious disease-diffusion algorithm[29] and further classify nodes into a susceptible state and an infected state. We call a node infected, if it carries a packet and the node is susceptible if it does not carry the packet. Then, there exist four kinds of nodes: disconnected and susceptible (*DS*), disconnected and infected (*DI*), connected and susceptible (*CS*), and connected and infected (*CI*), as shown in Fig. 3.

With the disease-diffusion algorithm, node s first sends copies of a packet m to its neighbors. The newly infected neighbors then repeat the same process until any node that belongs to the giant cluster receives m, as shown in Fig. 2(a). At the end of this stage, the infected nodes (i.e., several disconnected nodes and one connected node) constitute the source expanding tree, *SET*.

(2) The shortest path stage. Since there exists a connected path for each source-destination pair with a high probability in the giant cluster, we use the shortest path to transmit packets in this stage. Suppose a node u in the *SET* belongs to the giant cluster. It sends m to node v along the shortest path as shown in Fig. 2(b). In other words, if there is a packet at time slot $t-1$ at node u, at time slot t node u instantaneously sends the packet to the next hop in the shortest path if the link is active.

(3) The destination flooding stage. After the node v receives the message m, it diffuses m among the disconnected nodes until m reaches the destination node d. Finally, nodes having a copy of m form the destination expanding tree, *DET*.

With OFS, it can be intuitively inferred that the buffer occupation of disconnected nodes is limited. This is mainly because in each time slot, the giant cluster exists, and the size of *SET* and *DET* are finite. Therefore, the disconnected nodes only need to buffer messages coming from nearby sources or going to nearby destinations.

Contact Probability between disconnected and connected nodes. Let $C(n, p, q, L(t))$ denote the set of connected nodes and $D(n, p, q, L(t))$ the set of disconnected nodes (i.e., a node is a connected node if it belongs to the giant component; otherwise, it is a disconnected node). Let $w_1 \in D(n, p, q, L(t))$, $w_2 \in C(n, p, q, L(t))$ and $l_{w_1 w_2} \in L$. For any t_0, t ($t > t_0$), define the set of nodes in D that have been connected to C in at least one time step $i \in \{t_0 + 1, \ldots, t_0 + t\}$

$$H_{t_0 t} = \left\{ w_1 \in D\backslash C : X_i(l_{w_1 w_2}) = 1, w_2 \in C \right\}$$

where $l_{w_1 w_2}(i) = 1$ denotes that the link between w_1 and w_2 is active at moment i.

The following lemma is a key component in proving equation (2). Roughly speaking, it states that the probability that one disconnected node does not have contact with one connected node during $[0, t]$ (i.e., the link between them remains inactive within time step t) decreases exponentially in t.

Lemma 1 Let $w_1 \in D\backslash C$, for any $t_0 \geq 0$ and $t \geq 1$. It holds that

$$P(w_1 \notin H_{t_0 t}) \leq e^{-ptM} \tag{3}$$

where M is the size of C.

Proof: For any node $w_2 \in C$, the event $w_1 \notin H_{t_0 t}$ indicates that there is no active link between w_1 and w_2 during $[0, t]$. From the EMDG, the link inactivity probability is $(1 - p)$, and we obtain $P(l_{w_1 w_2}(i) = 0 | l_{w_1 w_2}(i-1) = 0) = 1 - p$.

Therefore,

$$P\left(w_1 \notin H_{t_0 t} \right) = P\left(\bigcap_{w_2 \in C} \bigcap_{i=t_0+1}^{t_0+t} X_i(l_{w_1 w_2}) = 0 \right)$$

$$= P\left(\bigcap_{w_2 \in C} (1 - p)^t \right)$$

Since Markov chains of different links are independent, it holds that

$$P\left(w_1 \notin H_{t_0 t} \right) = (1 - p)^{tM} \leq e^{-ptM}$$

Finite Transmission Time.

Lemma 2 Let $t_{s \to u}$ denote the transmission time from node s to node u, which represents the time when the package is first transmitted from the source node into nodes belonging to the giant cluster. It satisfies

$$P\left(t_{s \to u} \geq \frac{\log n}{pM} \right) \leq \frac{1}{n} \tag{4}$$

Proof: Recall that in stage 1 of the OFS, the flood process will stop if one *DI* node encounters any *CS* node. Assume there are k *DI* nodes and all of them will encounter the *CS* node u. The elapsed times are $t_{1,u}, \ldots, t_{k,u}$. Hence, $t_{s \to u} = \min(t_{1,u}, \ldots, t_{k,u})$. On the other hand, from lemma 1, we know that there exists one node $j \in D(n, p, q, L(t))$. It will contact the node u after $\frac{\log n}{pM}$ steps with at least probability $1 - O(1/n)$. This is mainly because

$$P\left(j \in H_{t_0 t} \right) \geq 1 - e^{-ptM} \geq 1 - e^{-\log n} = 1 - 1/n$$

when $t \geq \frac{\log n}{pM}$.

Since $t_{s \to u}$ is the minimum of k elapsed time, $t_{s,u} \leq \frac{\log n}{pM}$. We then obtain:
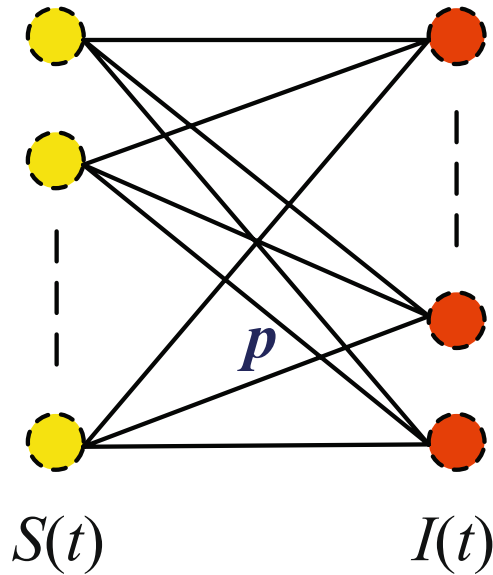
Figure 4. Number of active links.

$$P\left(t_{s,u} \geq \frac{\log n}{pM}\right) = 1 - P\left(t_{s,u} \leq \frac{\log n}{pM}\right) \leq \frac{1}{n}$$

Proof of Inequality (2).   Let $E(B_s(t))$ denote the expectation of buffer occupation in stage 1. We have:

$$E(B_s(t)) \leq rT_d S_n \frac{\log n}{pM}$$

For the buffer occupation in stage 3, we correspondingly have

$$E(B_d(t)) \leq rT_d D_n \frac{\log n}{pM}$$

where S_n is the size of SET and D_n represents that of DET. Finally, the buffer occupation of node u with the OFS is the sum of the above two parts, $E(B_u(t)) \leq rT_d(S_n + D_n)\frac{\log n}{pM} = c_1 rT_d \frac{\log n}{pM}$, where c_1 is determined by S_n and D_n. We will discuss the value of S_n and D_n in the evaluation section.

Bound on Node Buffer Size in the Subcritical Case.   If $p \leq c\frac{\log n}{n}(c \leq 1/2)$, the network is in the subcritical case, and the giant cluster does not exist. The main result is that there is a closed form for buffer occupation of nodes as shown in Theorem 2.

Theorem 2   *For a randomly selected node u of $\mathcal{G}(n, p, q, L(t))$ if $p \leq c\frac{\log n}{n}(c \leq 1/2)$, at the end of a time slot t,*

$$E(B_u(t)) = \frac{rT_d}{2p} \tag{5}$$

Similar to stages (1) and (3) of the OFS, we still use the infectious disease-diffusion algorithm to distribute copies, but this time the shortest path stage disappears. The flooding scheme therefore dominates the forwarding process.

The Number of Average Copies of a Packet When It Is Delivered.   Let $S(t)$ denote the number of susceptible nodes at a moment t and $I(t)$ denote the number of infected nodes. From the infectious disease-diffusion process, we obtain the varying rate of $I(t)$:

$$I'(t) = (n - 1)pSIp \tag{6}$$

Proof: Recall that with EMDG, the active probability of links is p, which implies that a node can contact $(n-1)p$ other nodes in one time slot. Since there are $I(t)$ infected and $S(t)$ susceptible nodes, the number of active links among themis $S(t) \times I(t) \times p$ as shown in Fig. 4. Hence, the number of new infected nodes in one time slot is $(n-1)pSIp$.

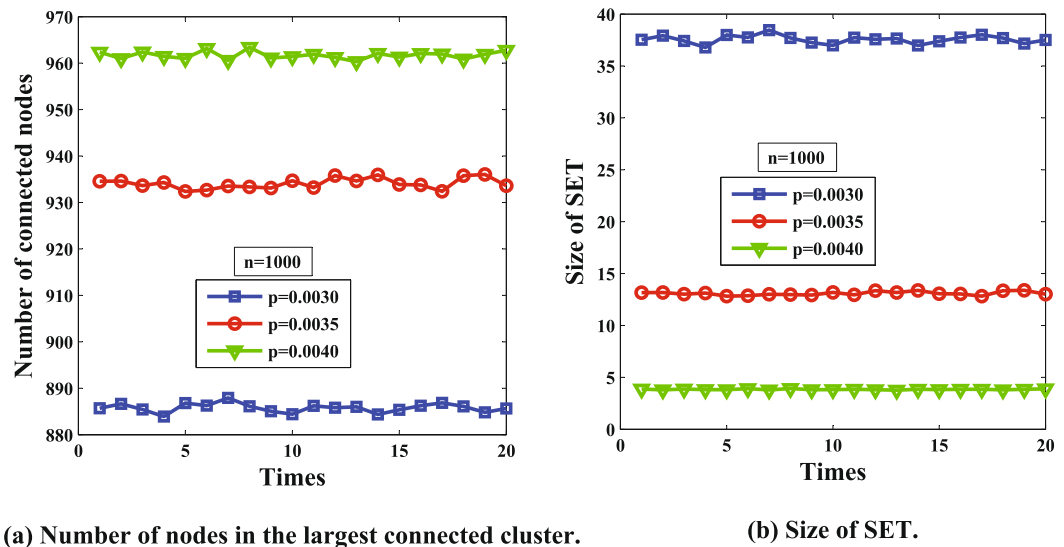With the condition $I(0) = 1$ and $I(t) + S(t) = n$, we obtain:

(a) Number of nodes in the largest connected cluster.

(b) Size of SET.

Figure 5. Number of nodes in the largest connected cluster and SET.

(a) Buffer size in different moments
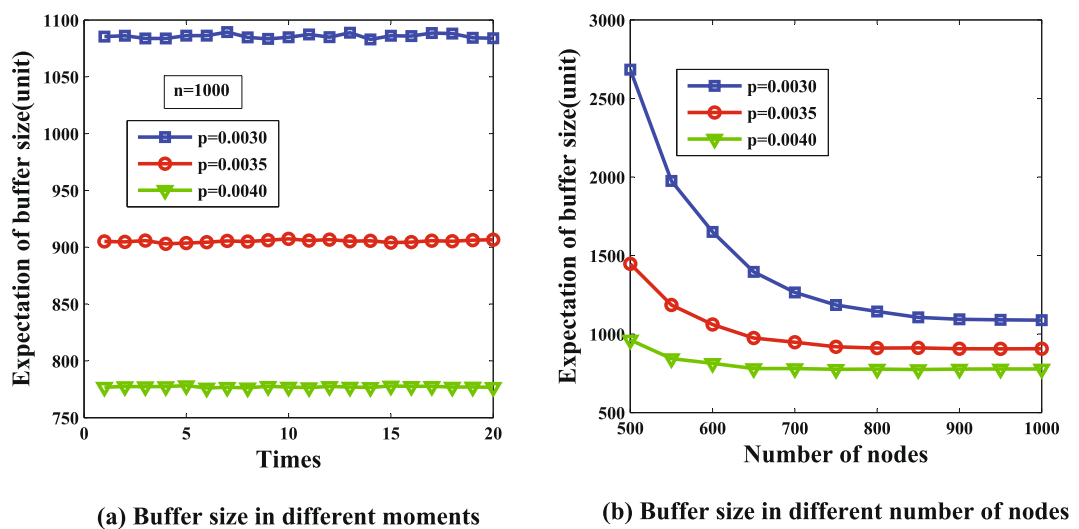
(b) Buffer size in different number of nodes

Figure 6. Buffer size in supercritical case.

$$I(t) = \frac{n}{1 + (n-1)e^{-n(n-1)tp^2}}$$

(7)

Correspondingly, the cumulative delay distribution $\mathcal{P}(t)$ until the first copy of a packet is delivered to its destination can be formulated as $\mathcal{P}'(t) = (n-1)pIp(1 - \mathcal{P})$. After some algebras, we obtain:

$$\mathcal{P}(t) = 1 - \left(\frac{n}{(n-1) + e^{n(n-1)tp^2}} \right)$$

(8)

The average number of copies of a packet when it is delivered to the destination, $E(c)$ is hence derived:

$$E(c) = \int_0^\infty I(t)\mathcal{P}'(t)dt - 1 = (n-1)/2$$

(9)

Proof of equation (5). We now prove equation (5).

Assume that each node is the source of one flow and the destination for another flow; then, there are n flows in the network. Given the $(n-1)/2$ copies of a packet and the packet generation rate r, the number of packets generated by nodes is nrT_d and the total number of packets (the packets generated by nodes and their copies) in
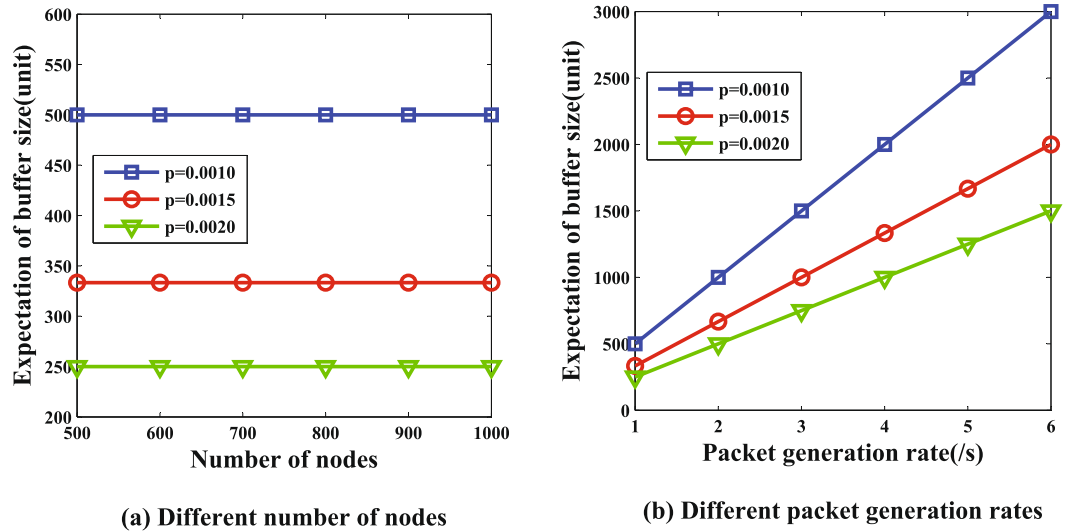
(a) Different number of nodes

(b) Different packet generation rates

Figure 7. Buffer size in subcritical case.

the system is $nrT_d(n-1)/2$, which is equally shared among the n nodes. The arrival rate of relay packets for each node is therefore $(n-1)rT_d/2$. Since a copy is deleted if the node encounters the destination, the contact rate is just the service rate $(n-1)p$. According to the queuing theory, the average buffer size of nodes is the ratio of the arrival rate and service rate. Hence, we obtain $E(B_u(t)) = \frac{rT_d}{2p}$.

Compute S_n, D_n and M. We demonstrate how to compute the values of S_n, D_n and M in this section.

Recall that S_n and D_n denote the size of SET and DET, respectively. Nodes in SET and DET are disconnected and do ont belong to the giant cluster M. Hence, the message diffusion in SET/DET also takes flooding in the same way as that in the subcritical case, and the number of infected nodes in SET/DET can be calculated by equation (7), where the parameters n and p are set in advance and the only issue is how to ascertain the value of t. Fortunately, from lemma 2, we know that the transmission time $t_{s \rightarrow u}$ can not exceed $\frac{\log n}{pM}$ with at least probability $1 - O(1/n)$. This means that the maximum value of t is $\frac{\log n}{pM}$, and we can use it to calculate the bound of the buffer size in equation (7). Using a similar process, we can get the value of D_n.

For the number of connected nodes in the giant cluster (i.e., M), it can be obtained by the MATLAB function largestcomponent().

Numerical Results

In this section, we present the simulation results to show the expectation of the buffer size in different conditions and other key components such as the number of nodes in the largest connected cluster.

For the supercritical case, we initially deploy 1000 nodes in the network. The value of c is 1.1 and $r = T_d = 1$. After that, we vary the parameter of p from 0.003 to 0.004 with a step 0.0005 to evaluate the network connectivity performance under different active probabilities. In addition, the simulation results are the average over 100 runs for statistical confidence.

Figure 5(a) illustrates the number of connected nodes. It is obvious that the number of connected nodes ranges from 880 to 970 as the active probability increases from 0.0030 to 0.0040, which means there indeed exists a giant connected cluster in the network. The two parameters are positively related. That is, the smaller the value of p is, the smaller the value of M should be.

Figure 5(b) shows the size of SET. We note that in the supercritical case, the source flooding stage is very short and the number of nodes in SET is smaller compared to those in the giant cluster. For example, the size of SET is smaller than 40 in $p = 0.003$, and under the same condition, M tends to 890.

After calculating the values of S_n, D_n and M, we can compute the expectation of node buffer size by equation (2). The results are shown in Fig. 6(a,b). Figure 6(a) shows the expected buffer size at different moments, and there is a slight fluctuation. Figure 6(b) further shows the influence of the number of nodes on the expectation of buffer size. We observe that the expectation curve is declining sharply when the number of nodes increases from 500 to 700 and levels out at the end, which demonstrates that only bits of buffer size are needed in this scenario.

For the subcritical case, we calculate the expectation of the node buffer size according to equation (5) and plot the results in Fig. 7, where the value of p is 0.001, 0.0015 and 0.002, respectively, and all of them satisfy the subcritical condition. Figure 7(a) shows the influence of p on buffer size, and Fig. 7(b) shows the role of packet generation r. The value of the buffer size ranges from 250 units to 500 with different active probability p and shows no relationship with the number of nodes in Fig. 7(a). In Fig. 7(b), the value of buffer size has a positive correlation with the parameter r. Generally speaking, the node buffer size is independent of the number of nodes in the subcritical case, which instead relies on the active probability p.

Conclusion

In this paper, we study the cache potentiality of mobile opportunistic networks. We model the MONs with EMDG and find that there exists a critical probability p_c from the perspective of percolation theory. If the active probability $p \geq p_c$, the network is in the supercritical case and there exist two bounds on the buffer size of nodes: the lower bound is dependent on both the active probability p and the inactive probability q, while the upper bound is independent of q. If, instead, $p \leq p_c$, there is a closed-form solution for buffer occupation that, is independent of the size of the network. In the future, extensive simulations under various scenarios will be run to evaluate the results, and to explore the influence of different buffer management metrics.

References

1. Yuan, P. & Liu, P. Data fusion prolongs the lifetime of mobile sensing networks. *Journal of Network and Computer Applications* **49**, 51–59 (2015).
2. Pentland, A., Fletcher, R. & Hasson, A. Daknet: Rethinking connectivity in developing nations. *Computer* **37**, 78–83 (2004).
3. Haas, Z. J. & Small, T. A new networking model for biological applications of ad hoc sensor networks. *IEEE/ACM Transactions on Networking* **14**, 27–40 (2006).
4. Ma, H., Zhao, D. & Yuan, P. Opportunities in mobile crowd sensing. *IEEE Communications Magazine* **52**, 29–35 (2014).
5. Zhao, D., Ma, H., Liu, L. & Zhao, J. On opportunistic coverage for urban sensing. In *Proceedings of 2013 IEEE 10th International Conference on Mobile Ad-Hoc and Sensor Systems, Hangzhou, China*. New York, USA:IEEE 231–239, doi:10.1109/MASS.2013.66 (December, 2013).
6. Wu, H., Ma, H. & Zhao, D. Videocent: a quality-oriented incentive mechanism for video delivery in opportunistic networks. *Wireless Networks* **21**, 769–781 (2015).
7. Wang, S., Zhang, J. & Tong, L. Delay analysis for cognitive radio networks with random access: A fluid queue view. In *Proceedings of INFOCOM 2010, San Diego, CA, USA*. New York, USA:IEEE. 1–9, doi:10.1109/INFCOM.2010.5461943 (May, 2010).
8. Lindgren, A., Doria, A. & Schelén, O. Probabilistic routing in intermittently connected networks. *ACM SIGMOBILE mobile computing and communications review* **7**, 19–20 (2003).
9. Hui, P., Crowcroft, J. & Yoneki, E. Bubble rap: Social-based forwarding in delay-tolerant networks. *IEEE Transactions on Mobile Computing* **10**, 1576–1589 (2011).
10. Spyropoulos, T., Psounis, K. & Raghavendra, C. S. Efficient routing in intermittently connected mobile networks: the multiple-copy case. *IEEE/ACM transactions on networking* **16**, 77–90 (2008).
11. Chen, K. & Shen, H. Smart: Lightweight distributed social map based routing in delay tolerant networks. In *Proceedings of 2012 20th IEEE International Conference on Network Protocols (ICNP), Austin, TX, USA*. New York, USA:IEEE. 1–10, doi:10.1109/ICNP.2012.6459982 (November, 2012).
12. Yuan, P., Ma, H. & Fu, H. Hotspot-entropy based data forwarding in opportunistic social networks. *Pervasive and Mobile Computing* **16**, 136–154 (2015).
13. De Abreu, C. S. & Salles, R. M. Modeling message diffusion in epidemical dtn. *Ad Hoc Networks* **16**, 197–209 (2014).
14. Chaintreau, A., Mtibaa, A., Massoulie, L. & Diot, C. The diameter of opportunistic mobile networks. In *Proceedings of the 2007 ACM CoNEXT conference, New York, USA*. New York, USA: ACM. 12, doi:10.1145/1364654.1364670 (December, 2007).
15. Bogo, F. & Peserico, E. Optimal throughput and delay in delay-tolerant networks with ballistic mobility. In *Proceedings of the 19th annual international conference on Mobile computing & networking, Miami, Florida, USA*. New York, USA: ACM. 303–314, doi:10.1145/2500423.2500432 (October, 2013).
16. Becchetti, L. *et al.* Flooding time in opportunistic networks under power law and exponential intercontact times. *IEEE Transactions on Parallel and Distributed Systems* **25**, 2297–2306 (2014).
17. Balasubramanian, A., Levine, B. & Venkataramani, A. Dtn routing as a resource allocation problem. In *ACM SIGCOMM Computer Communication Review* **37**, 373–384 (ACM, 2007).
18. Pitkänen, M. J. & Ott, J. Redundancy and distributed caching in mobile dtns. In *Proceedings of 2nd ACM/IEEE international workshop on Mobility in the evolving internet architecture, Kyoto, Japan*. New York, USA: ACM. **8**, doi:10.1145/1366919.1366930 (August, 2007).
19. Kaveevivitchai, S. & Esaki, H. Independent dtns message deletion mechanism for multi-copy routing scheme. In *Proceedings of the Sixth Asian Internet Engineering Conference, Bangkok, Thailand*. New York, USA: ACM. 48–55, doi:10.1145/1930286.1930293 (November, 2010).
20. Prodhan, A. T., Das, R., Kabir, H. & Shoja, G. C. Ttl based routing in opportunistic networks. *Journal of Network and Computer Applications* **34**, 1660–1670 (2011).
21. Pan, D., Ruan, Z., Zhou, N., Liu, X. & Song, Z. A comprehensive-integrated buffer management strategy for opportunistic networks. *EURASIP Journal on Wireless Communications and Networking* **2013**, 1–10 (2013).
22. Rocha, L. E. & Masuda, N. Individual-based approach to epidemic processes on arbitrary dynamic contact networks. *arXiv preprint arXiv:1510.09179* (2015).
23. Stegehuis, C., van der Hofstad, R. & van Leeuwaarden, J. S. Epidemic spreading on complex networks with community structures. *Scientific Reports* **6** (2016).
24. Krifa, A., Barakat, C. & Spyropoulos, T. Optimal buffer management policies for delay tolerant networks. In *Proceedings of 5th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks, San Francisco, California, USA*. New York, USA: IEEE. 260–268, doi:10.1109/SAHCN.2008.40 (2008).
25. Zhang, G., Wang, J. & Liu, Y. Congestion management in delay tolerant networks. In *Proceedings of the 4th Annual international conference on wireless internet, Maui, Hawaii, USA*. Brussels, Belgium: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering). **65** (2008).
26. Soares, V. N., Farahmand, F. & Rodrigues, J. J. Evaluating the impact of storage capacity constraints on vehicular delay-tolerant networks. In *Proceedings of Communication Theory, Reliability, and Quality of Service, 2009. CTRQ'09. Second International Conference on, Colmar, France*. New York, USA: IEEE. 75–80, doi:10.1109/CTRQ.2009.11 (2009).
27. Zhuo, X., Li, Q., Gao, W., Cao, G. & Dai, Y. Contact duration aware data replication in delay tolerant networks. In *Proceedings of 19th IEEE International Conference on Network Protocols, Vancouver, BC, Canada*. New York, USA: IEEE. 236–245, doi:10.1109/ICNP.2011.6089057 (2011).
28. Andrea, C. *et al.* Flooding time in edge-Markovian dynamic graphs. In *Proceedings of Twenty-Seventh ACM Symposium on Principles of Distributed Computing, PODC 2008, Toronto, Canada*. New York, USA: ACM. 213–222, doi:10.1145/1400751.1400781 (August 2008).
29. Brauer, F., Castillo-Chavez, C. & Castillo-Chavez, C. *Mathematical models in population biology and epidemiology, Texts in Applied Mathematics* vol. 40 (Springer, 2001).

Author Contributions

Peiyan Yuan asked the questions, suggested the answers and wrote the manuscript; Honghai Wu gave some valuable advices; Xiaoyan Zhao conducted the experiment; Zhengnan Dong analyzed the results. All authors reviewed the manuscript.

Additional Information

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.