2014 IEEE International Conference on High Performance Computing and Communications (HPCC), 2014 IEEE 6th International Symposium on Cyberspace Safety and Security (CSS) and 2014 IEEE 11th International Conference on Embedded Software and Systems (ICESS)

On Delivery Delay-Constrained Throughput and End-to-end Delay in MANETs

Yujian Fang*, Yuezhi Zhou*, Xiaohong Jiang† and Yaoxue Zhang‡
* Tsinghua University
Email: {fyj12, zhouyz}@{mails, mail}.tsinghua.edu.cn
† Future University Hakodate, Email: jiang@fun.ac.jp
‡ Central South University, Tsinghua University, Email: zyx@csu.edu.cn

***Abstract*—The real achievable performance of mobile ad hoc networks (MANETs), in particular the performance of such networks under practical constraints, is still largely unknown by now. As a step forward in this direction, this paper focuses on a MANET where a maximum allowed delivery delay is imposed to each packet and examines the impact of such delivery delay constraint on its real achievable performance in terms of throughput and packet end-to-end delay. We first determine the throughput capacity of the MANET to reveal the maximum possible throughput the network can support. For a given exogenous rate to each node, we then provide analysis on the achievable throughput, packet delivery ratio and packet end-to-end delay, and show how they are determined by the steady state of relay queues in the network. For the analysis of steady state of relay queues, we further develop a subtle Markov chain model based on the idea of state reduction such that state space explosion problem in the analysis can be avoided. Finally, numerical results are presented to illustrate the performance of the network and the impact brought by the delivery delay constraint.**

Keywords*—Mobile ad hoc networks, Throughput, End-to-end delay, Delivery delay constraint, Performance modeling

I. Introduction

A mobile ad hoc network (MANET) consists of a collection of mobile nodes, each of which communicates with others in a peer-to-peer manner. Because of its distributed nature and the weak dependency on pre-existing infrastructure, MANET has been appealing for many application scenarios, including battlefield networks, metropolitan mesh networks, and vehicular ad hoc networks. Despite the extensive research activities devoted to the study of MANETs, the fundamental theory on the performance of MANET still remains an open problem [1], which greatly stunted the application and commercialization of practical MANETs.

In the seminal work of Gupta and Kumar [2], scaling laws on the network capacity of static ad hoc network were investigated and it was proved that the network throughput there scales as $\Theta(1/\sqrt{n \log n})$ as the number of nodes n increases. Since then there has been a growing interest in the study of MANET capacity scaling laws under various schemes, like applying power control [3], exploring hierarchical cooperation [4], adopting directional antennas [5], *etc*, and under different traffic models, like broadcast [6] and multicast [7]. By fully exploiting the node mobility as a means of message delivery, Grossglauser *et al.* [8] discovered that a non-vanishing throughput is possible, and a two-hop relay algorithm was proposed to achieve the $\Theta(1)$ throughput, but Gamal *et al.* [9] showed that such throughput was achieved at the cost of packet delay that scales as $\Theta(n^{1/2}/v(n))$, where $v(n)$ is the velocity of nodes. Later, plenty of works have then been done to reveal inherent trade-off between capacity and delay (see, for example, [10]–[14]). Although the scaling law results can indicate the general trend of network performance as the number of nodes n increases, they provide little information about the actual performance that can be expected in real network scenarios, which in practice is of great importance and serves as the guideline for network design and implementation.

Some initial works are now available on the exact and real achievable performance of MANETs. Neely *et al.* [11] computed the exact network capacity region under the i.i.d. mobility model, and Urgaonkar *et al.* [15] established the capacity under a more general Markovian mobility. In [16], Liu *et al.* explored the exact throughput capacity and delay under a specific two-hop relay algorithm with packet redundancy; Gao *et al.* [17] extended the work of [16] to allow for transmission range adjustment, and Chen *et al.* [18] further extended the work of [16] to a MANET with directional antennas. It is notable that these works are based on some ideal assumptions, in particular, they all assume that packet delivery delay is unbounded, which is not practical in real networks like vehicular ad hoc networks and battlefield networks, where messages usually have a natural period of validity, exceeding which the message may not be useful anymore. Thus, for the practical performance study of MANETs, the constraint on packet delivery delay should be carefully addressed.

As a step towards the practical performance study for MANETs, this paper considers a MANET where a maximum allowed delivery delay is imposed to each packet. Notice that with such a constraint, “non-fresh” packets exceeding the maximum allowed delivery delay will be dropped so that the precious transmission opportunities can be left to other “fresh” packets, leading to an improvement of overall network performance. It is worth mentioning that although some works with the similar concept of delivery delay constraint are available [19]–[21], they only consider the delivery of a single packet through an empty network and thus cannot reveal the real impact of delivery delay constraint on the performance of a practical network. In a real network, multiple packets coexist and compete for transmission opportunities, and these packets usually need to experience a complex queuing process at relay nodes before reaching their destination nodes.

The main contributions of this paper are summarized as follows:

978-1-4799-6123-8/14 $31.00 © 2014 IEEE
DOI 10.1109/HPCC.2014.77

- For a MANET where a maximum allowed delivery delay is imposed to each packet, we first determine the throughput capacity of the MANET to reveal the maximum possible throughput the network can support.
- Under any exogenous rate to each node, we then provide analysis on the achievable throughput, packet delivery ratio and packet end-to-end delay, and show in closed-form how they are determined by the steady state of relay queues in the network.
- For the analysis of steady state of relay queues, we further develop a subtle Markov chain model based on the idea of state reduction such that state space explosion problem in the analysis can be avoided.
- Finally, extensive simulation and numerical results are provided to validate our theoretical results and to illustrate the impact brought by the delivery delay constraint.

The rest of the paper is organized as follows. The system models are introduced in Section II. We then derive the theoretical performance in Section III, with the aid of a Markov chain proposed in Section IV. The numerical results are studied in Section V for validation and illustration, and finally, we conclude our paper in Section VI.

II. System Models

In this section, we introduce the system models and routing scheme adopted in this study.

A. Network Model

We consider a cell partitioned network where n nodes reside in a network that is divided into $\sqrt{m} \times \sqrt{m}$ homogeneous non-overlapping cells. Assume that time is slotted, and at the beginning of each time slot, a node jumps uniformly and independently into one of the m cells, and stays in the cell till the end of the time slot. In this work, fast mobility is adopted, so that the mobility and the transmission rate are of the same scale. For simplicity, each cell is assumed to support exactly one packet transfer per time slot. Interested readers are referred to [12], [14], [22] for mobilities that are assumed to occur at a much slower time scale than the transmissions.

B. Communication Model

Similar to [14], [23], in this paper we only allow nodes inside the same cell to transmit to each other such that the interference is limited locally and as many as possible simultaneous transmissions can be maintained. Note that interference in MANETs is generally modeled with Protocol Model and Physical Model [2], but we can satisfy the transmission requirements under both models by employing different transmission frequencies among neighboring cells [23], so that simultaneous transmissions in all cells are possible.

C. Traffic Model and Delivery Delay Constraint

The traffic flows in the network are defined in a permutation manner, where each node is the source node for a flow and in the meanwhile the destination of another flow. For each of the n distinct flows, exogenous packets arrive at the source node according to a Bernoulli process with rate λ packets/slot, and a maximum allowed delivery delay is τ imposed to each packet after it leaves the source node. The formal definitions of delivery delay and delivery delay constraint are presented as follows:

Delivery delay: Suppose a packet leaves its source node in time slot t_s, and arrives at its destination node in time slot t_d, the delivery delay of it is defined as $t_d - t_s$. Here we are only interested in the delay caused by the delivery process, so the queuing delay of packet at its relay node rather than at its source node is of concern.

Delivery delay constraint (τ): It is required that the delivery delay $t_d - t_s$ of each packet should not to exceed a given threshold τ, otherwise it is dropped at any intermediate node. Since we can place the remaining delivery time in an extra field of the packet (as shown in Fig. 1), any intermediate node can easily detect overdue packets and drop it upon timeout.

D. Routing Scheme

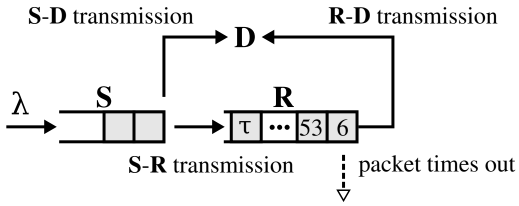

Fig. 1. Illustration of the transmission scheme.

Without loss of generality, we focus on a tagged flow $(\mathbf{S}, \mathbf{D})$ with source node $\mathbf{S}$ and destination node $\mathbf{D}$ in our discussion, and use $\mathbf{R}$ to refer to a general rely node. The routing algorithm considered in this work derives from the well-known two-hop relay algorithm proposed in [8], where a packet leaving $\mathbf{S}$ for $\mathbf{D}$ is either directly transmitted, or through exactly one intermediate node $\mathbf{R}$ (as shown in Fig. 1). The intuition of the algorithm is as follows: we choose a node randomly from each cell to be the transmitter; it first attempts to conduct a source-to-destination ($\mathbf{S}$-$\mathbf{D}$) transmission, and upon failure, it then tries to conduct a source-to-relay ($\mathbf{S}$-$\mathbf{R}$) transmission or a relay-to-destination ($\mathbf{R}$-$\mathbf{D}$) transmission with equal probability. The algorithm to schedule the transmitter and route packets in a given cell C is presented in detail in Fig. 2.

To support the operation of the above routing algorithm, each node in the network is equipped with two kinds of queues, a *local queue* and $n-2$ *relay queues*, both following the first-come first-served (FCFS) principle. A node, as a source node it stores its exogenous packets waiting to be dispatched in its local queue, while as a relay node for up to $n-2$ other flows (except the flows originating from and destined for itself), it uses the $n-2$ relay queues to store packets for these flows (one relay queue per flow). It is interesting to notice that the delivery delay constraint actually imposes a buffer limit on each relay queue, because there will be no more than τ packets in such a queue. It is also noticed that operations on the queues are $\Theta(1)$, making possible real-time transmissions while performing little impact on the whole system in terms of timing guarantee.

```
procedure SCHEDULE-ROUTE($C$)
    $\mathbf{S} \leftarrow$ Randomly chosen node in $C$
    if $\mathbf{D}$ is inside $C$ then
        $\mathbf{S}$ transmits a packet directly to $\mathbf{D}$ from its local
            queue.
    else
        $\mathbf{S}$ randomly chooses one node $\mathbf{N}$ from $C$.
        $p \leftarrow$ Random(1)        ▷ A random number in $[0, 1)$
        if $p < 0.5$ then
            $\mathbf{S}$ transmits a packet to $\mathbf{N}$ (acting as relay node)
                from its local queue.
        else
            $\mathbf{S}$ transmits a packet destined at $\mathbf{N}$ from the
                corresponding relay queue.
        end if
    end if
end procedure
```

Fig. 2. The scheduling and routing algorithm for nodes in cell C.

E. Some Definitions

Throughput capacity: the throughput capacity of a network, denoted as μ, is the maximum exogenous packet rate that can be supported by the network such that the stability of all queues in the network is ensured. By stability we mean that the length of each queue is finite almost surely.

Achievable throughput: for a given exogenous packet rate λ, the achievable throughput of the network is defined as the reception rate at the destination node of each flow. Since packet loss is possible in a delivery delay constrained network, the achievable throughput is less than or equal to λ.

Packet delivery ratio: for a tagged flow $(\mathbf{S}, \mathbf{D})$, the packet delivery ratio is defined as the ratio between the reception rate at $\mathbf{D}$ and the exogenous input rate at $\mathbf{S}$. In a network where stability of all queues is ensured, this metric reveals the possibility for a packet to successfully reach its destination under the delivery delay constraint.

Packet end-to-end delay: the end-to-end delay of a packet is the sum of its queuing delay at all nodes it goes through. Unlike the delivery delay that accounts for only queuing delay at relay(s), end-to-end delay includes also the queuing delay at the source node.

III. Performance Analysis

A. Basic Probabilities

In this subsection, some basic probabilities are presented to support further analysis.

Lemma 1: For a tagged flow $(\mathbf{S}, \mathbf{D})$ in an observed time slot, we denote by p_{sd} the probability that $\mathbf{S}$ acquires the opportunity to conduct a source-to-destination transmission. We then have

$$p_{sd} = \frac{1}{n-1}\left(1 - \frac{m}{n} + \frac{m}{n}\left(1 - \frac{1}{m}\right)^{n}\right). \tag{1}$$

Proof: Suppose there are k nodes other than $\mathbf{S}$ and $\mathbf{D}$ that reside in the same cell as $\mathbf{S}$, for $\mathbf{S}$ to obtain a source-to-destination transmission opportunity, the following conditions should hold: 1) $\mathbf{D}$ is currently inside the same cell as $\mathbf{S}$, with probability $1/m$; 2) among all $k+2$ nodes ($\mathbf{S}$, $\mathbf{D}$ and other k nodes), $\mathbf{S}$ wins the transmission chance, with probability $1/(k+2)$. It follows that

$$p_{sd} = \frac{1}{m}\sum_{k=0}^{n-2}\binom{n-2}{k}\left(\frac{1}{m}\right)^{k}\left(\frac{m-1}{m}\right)^{n-k-2}\frac{1}{k+2}. \tag{2}$$

To further simplify (2), we introduce a more general form of the sum:

$$\sum_{i=0}^{n}\binom{n}{i}a^{i}b^{n-i}\frac{1}{i+2} = \frac{(a+b)^{n+1}(a-b+na)+b^{n+2}}{a^{2}(n+1)(n+2)},$$

which, once applied to (2), finishes our proof for the lemma. ■

Lemma 2: For a tagged node $\mathbf{S}$, let p_{sr} (resp. p_{rd}) represents the probability for $\mathbf{S}$ to acquire the opportunity to conduct a source-to-relay (resp. relay-to-destination) transmission in a certain time slot, we have:

$$p_{sr} = p_{rd} \tag{3}$$

$$= \frac{1}{2}\cdot\frac{m-1}{n-1}\left(1 - \left(\frac{m-1}{m}\right)^{n-1}\right) - \frac{1}{2}\left(\frac{m-1}{m}\right)^{n-1}. \tag{4}$$

Proof: We assume that a positive number of k nodes are in the same cell as $\mathbf{S}$. For $\mathbf{S}$ to perform a source-to-relay (resp. relay-to-destination) transmission, the following conditions should be satisfied: 1) $\mathbf{D}$ is not in the same cell as $\mathbf{S}$, with probability $(m-1)/m$; 2) among all $k+1$ nodes ($\mathbf{S}$ and other k nodes), $\mathbf{S}$ is scheduled as the transmitter with probability $1/(k+1)$; 3) with probability $1/2$, $\mathbf{S}$ chooses to conduct a source-to-relay (resp. relay-to-destination) transmission. Hence

$$\begin{aligned} p_{sr} &= p_{rd} \\ &= \frac{m-1}{m}\sum_{k=1}^{n-2}\binom{n-2}{k}\left(\frac{1}{m}\right)^{k}\left(\frac{m-1}{m}\right)^{n-2-k}\frac{1}{k+1}\cdot\frac{1}{2}. \end{aligned} \tag{5}$$

Similarly, by applying following formula to (5), then (4) follows immediately:

$$\sum_{i=0}^{n}\binom{n}{i}a^{i}b^{n-i}\frac{1}{i+1} = \frac{(a+b)^{n+1}-b^{n+1}}{a(n+1)}.$$

■

B. Throughput Capacity

Theorem 1: For the concerned MANET, its throughput capacity μ is determined as

$$\mu = p_{sd} + p_{sr}. \tag{6}$$

Proof: Recall that there are two types of queues in the network, namely the local queue of $\mathbf{S}$ and the relay queues of $\mathbf{R}$. We now examine the stability for each of them.

Since source-to-relay and source-to-destination transmissions are mutually exclusive transmission events in the same time slot for $\mathbf{S}$, the local queue of $\mathbf{S}$ can be modeled as a Bernoulli/Bernoulli queue with input rate equal to λ and service rate equal to $p_{sd} + p_{sr}$. This suggests that the stability is ensured as long as the input rate satisfy:

$$\lambda \leq p_{sd} + p_{sr}. \tag{7}$$

The queue at $\mathbf{R}$ is a little different, where two matters complicate the problem: firstly, packet arrivals and departures are no longer independent, but rather mutually exclusive; secondly, except for normal transmissions, packets can leave the queue due to delivery delay constraint. To check the stability of the queue, we notice that the delivery delay constraint actually increase the chance of packet departures. Thus, if we denote by μ_r the service rate of the relay queue and by μ_r' the actual output rate of it, according to the symmetry of $n-2$ relay queues, we have

$$\mu_r = \frac{p_{rd}}{n-2} \leq \mu_r'. \tag{8}$$

For the input rate λ_r, it is easy to see that

$$\lambda_r = \lambda \frac{p_{sr}}{p_{sd} + p_{sr}} \frac{1}{n-2}, \tag{9}$$

as each packet leaving the local queue of $\mathbf{S}$ will enter the relay queue of an observed $\mathbf{R}$ with probability $\frac{p_{sr}}{p_{sd}+p_{sr}} \frac{1}{n-2}$. Combining (3), (7), (8) and (9), we have

$$\lambda_r = \lambda \frac{p_{sr}}{p_{sd} + p_{sr}} \frac{1}{n-2} \leq \frac{p_{sr}}{n-2} = \frac{p_{rd}}{n-2} = \mu_r \leq \mu_r',$$

thereby proving the stability for relay queues. ■

Corollary 1: Consider a density-fixed network where $d = n/m = \Theta(1)$. The optimal value of d that maximizes μ as $n \to \infty$ is given as the positive solution of the following equation:

$$d^2 + d + 1 = \exp(d). \tag{10}$$

Proof: It is easy to verify that as $n \to \infty$, $m = n/d \to \infty$, $p_{sd} \to 0$, and

$$p_{sr} \to \frac{1}{2d}(1 - \exp(-d)) - \frac{1}{2}\exp(-d). \tag{11}$$

By checking the zero of the derivative of (11), (10) then follows. ■

C. Achievable Throughput and Delivery Ratio

Till now we have only discovered the throughput capacity by ensuring stability of all queues in the network. Since packet loss may occur in a network with delivery delay constraint, we are more interested in the real achievable throughput as well as the delivery ratio under a given exogenous packet rate λ. In this section, the closed-form expressions for achievable throughput and delivery ratio are presented. Note that some unsolved terms in the expression requires a deep look into the relay queues where timeouts take place, and we will deal with that in Section IV.

Theorem 2: Let $T(\lambda)$ denote the achievable throughput (i.e., the reception rate of $\mathbf{D}$) under a given exogenous packet rate λ, we have

$$T(\lambda) = \lambda \frac{p_{sd}}{p_{sd} + p_{sr}} + \lambda \frac{p_{sr}}{p_{sd} + p_{sr}} \delta(\lambda), \tag{12}$$

in which

$$\delta(\lambda) = 1 - \pi_\tau^{(\lambda)}(1 - \mu_r), \tag{13}$$

where $\pi_t^{(\lambda)}$ refers to the probability that a packet leaving the relay queue (due to either relay-to-destination transmission or timeout) has stayed in the queue for t time slots.

Proof: For a packet leaving the local queue of $\mathbf{S}$, it will reach $\mathbf{D}$ through a direct source-to-destination transmission with probability $p_{sd}/(p_{sd} + p_{sr})$, or be stored in the queue of a relay node $\mathbf{R}$ with probability $p_{sr}/(p_{sd} + p_{sr})$. Let $\delta(\lambda)$ be the probability that a packet entering the relay queue of $\mathbf{R}$ may reach the destination within τ time slots, (12) then comes naturally. To decide $\delta(\lambda)$, we notice that for each packet dropped by $\mathbf{R}$: 1) it has been in the relay queue for τ time slots, with probability $\pi_\tau^{(\lambda)}$; 2) it is the header packet of the relay queue since any prior packets must have left or timed out; 3) in the τth time slot during its stay in $\mathbf{R}$, it fails to be delivered, with probability $1 - \mu_r$. Combining the above facts yields (13) and finishes the proof. ■

Corollary 2: Consider the throughput without delivery delay constraint, we have

$$\delta(\lambda) = 1 - \pi_\infty^{(\lambda)}(1 - \mu_r) \to 1 \tag{14}$$

according to the stability of all queues. Therefore

$$T(\lambda) \to \lambda, \quad \tau \to \infty. \tag{15}$$

Corollary 3: Let $R(\lambda)$ be the delivery ratio under a given exogenous packet rate λ, i.e., the ratio of packets that successfully reach the final destination, then

$$R(\lambda) = \frac{T(\lambda)}{\lambda}. \tag{16}$$

D. End-to-end Delay

For a given exogenous input rate $\lambda < \mu$, this subsection presents analysis on the packet end-to-end delay. Note that unlike the vast works on end-to-end delay analysis in literature (see, for example, [11], [16]), due to the packet loss caused by delivery delay constraint here we count delay *only* for packets that successfully reach their destinations.

Theorem 3: In the concerned MANET with delivery delay constraint and a given exogenous input rate $\lambda < \mu$, the expected end-to-end delay $D(\lambda)$ of a packet that successfully reaches its destination node is given by

$$D(\lambda) = D_s(\lambda) + D_r(\lambda) \frac{p_{sr} \cdot \delta(\lambda)}{p_{sd} + p_{sr} \cdot \delta(\lambda)}, \tag{17}$$

where

$$D_s(\lambda) = \frac{1 - \lambda}{\mu - \lambda}, \tag{18}$$

and

$$D_r(\lambda) = \frac{\left(\sum_{t=1}^{\tau-1} \pi_t^{(\lambda)} t\right) + \pi_\tau^{(\lambda)} \mu_r \tau}{\delta(\lambda)}. \tag{19}$$

Proof: For a packet that successfully reach $\mathbf{D}$, let ω_1 be the event that it has traveled one hop, and let ω_2 be the event that it has traveled two hops. According to the analysis in the proof of Theorem 2, we have:

$$\Pr(\omega_1) = \frac{p_{sd}}{p_{sd} + p_{sr}\delta(\lambda)} \tag{20}$$

and

$$\Pr(\omega_2) = \frac{p_{sr}\delta(\lambda)}{p_{sd} + p_{sr}\delta(\lambda)}. \tag{21}$$

Let D_s denote the queuing time a packet spends in $\mathbf{S}$, and let D_r denote the queuing time it spends in $\mathbf{R}$ (if the packet is delivered through two hops), then

$$\begin{aligned} D(\lambda) &= \Pr(\omega_1)\mathbb{E}[D_s|\omega_1] + \Pr(\omega_2)\mathbb{E}[D_s + D_r|\omega_2] && (22) \\ &= \mathbb{E}[D_s] + \Pr(\omega_2)\mathbb{E}[D_r|\omega_2], && (23) \end{aligned}$$

where (23) is due to the fact that whether a packet is transmitted through a relay node only depends on the time slot it leaves the local queue of $\mathbf{S}$, independent of its queuing time at $\mathbf{S}$. Recall that the local queue of $\mathbf{S}$ is a simple Bernoulli/Bernoulli queue with input rate λ and service rate μ, it follows that $\mathbb{E}[D_s] = D_s(\lambda)$ with its value given in (18). As for $\mathbb{E}[D_r|\omega_2] = D_r(\lambda)$, we note that

$$\Pr(D_r = t|\omega_2) = \begin{cases} \pi_t^{(\lambda)}/\delta(\lambda) & t < \tau \\ \pi_t^{(\lambda)}\mu_r/\delta(\lambda) & t = \tau \end{cases}, \tag{24}$$

and (19) can be obtained by summing up $\Pr(D_r = t|\omega_2) \cdot t$. ■

Corollary 4: Consider the end-to-end delay without delivery delay constraint where the relay queue of $\mathbf{R}$ degenerates to a birth-death chain, it follows that

$$D_r(\lambda) \to \frac{1}{\mu_r - \lambda_r}, \quad \tau \to \infty, \tag{25}$$

and therefore

$$D(\lambda) \to \frac{1 - \lambda + (n-2)\mu}{\mu - \lambda}, \quad \tau \to \infty. \tag{26}$$

IV. $\pi^{(\lambda)}$: The Markov Chain Approach

As a complement to the unsolved terms $\pi^{(\lambda)}$ in Theorem 2 and Theorem 3, this section digs into the queuing process at $\mathbf{R}$ and models each queue there with a subtle Markov chain, which not only provides us with desired terms essential to the theorems but also effectively prevents the problem of state space explosion.

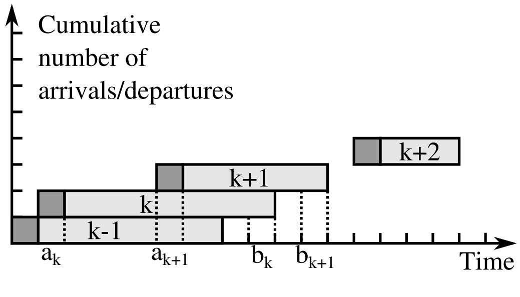

Fig. 3. Illustration of the Markov variable.

As is illustrated in Fig. 3, for a potential relay queue of the tagged flow, we denote by a_k (dark-shaded) the arrival time slot of the kth packet at the queue, and denote by b_k the corresponding departure time slot of it. By monitoring on each packet, the stay duration of each packet $b_k - a_k$ (light-shaded) is guaranteed not to exceed τ. To formalize the constraints on a_k and b_k, we have:

$$\begin{aligned} &\forall k, \quad a_k < a_{k+1}, && (27a) \\ &\forall k, \quad b_k < b_{k+1}, && (27b) \\ &\forall k, \quad 1 \le b_k - a_k \le \tau. && (27c) \end{aligned}$$

To serve our purpose, we now focus on the queuing time $c_k = b_k - a_k$ of the kth packet, and develop the Markov chain to get the stationary distribution of it. We start with two preliminary lemmas.

Lemma 3: Given the queuing time c_k of the last packet, the distribution of the arrival time difference $a_{k+1} - a_k$ can be presented as:

$$\begin{aligned} &\Pr(a_{k+1} - a_k = a|c_k = i) \\ &= \begin{cases} \left(1 - \frac{\lambda_r}{1-\mu_r}\right)^{a-1} \cdot \frac{\lambda_r}{1-\mu_r} & 1 \le a < i \\ \left(1 - \frac{\lambda_r}{1-\mu_r}\right)^{a-1} \cdot \lambda_r \cdot \mathbf{1}_{i=\tau} & a = i \\ \left(1 - \frac{\lambda_r}{1-\mu_r}\right)^{i-1} (1 - \mathbf{1}_{i=\tau}\lambda_r) \cdot \\ (1-\lambda_r)^{a-i-1} \cdot \lambda_r & a > i \end{cases}, \end{aligned} \tag{28}$$

where $\mathbf{1}_{i=\tau}$ is the standard indicative variable indicating $i = \tau$.

Proof: A key observation to prove the lemma is that there is competence between the transmission of the kth packet and the arrival of the $(k+1)$th packet, since $\mathbf{R}$ cannot receive and transmit a packet simultaneously. We now prove the lemma by considering the following three cases of a:

For $1 \le a < i$, we know that since time slot a_k, the relay queue can receive one packet each time slot with conditional probability $\lambda_r/(1 - \mu_r)$, given that the kth packet is not transmitted.

For $a = i$, we notice that the $(k+1)$th packet cannot arrive at $\mathbf{R}$ in time slot b_k if $i < \tau$, because it is used to transmit the kth packet to its destination; however, the situation is a little different when $i = \tau$, where with probability $1 - \mu_r$ the kth packet is dropped instead of being delivered to $\mathbf{D}$, making it possible for the $(k+1)$th packet to arrive at $\mathbf{R}$.

For $a > i$, again, we treat $i = \tau$ as a special case where with probability $1 - \mu_r$ the kth packet times out, and with probability $(1 - \mu_r)(1 - \frac{\lambda_r}{1-\mu_r}) + \mu_r = 1 - \lambda_r$, the $(k+1)$th does not arrive in time slot b_k.

Combining all three cases, the conditional probability is acquired, considering that the $(k+1)$th packet fails to arrive at $\mathbf{R}$ during the $a - 1$ time slots after a_k, but succeeds in time slot $a_k + a$. ■

Lemma 4: Given $c_k = i$, $a_{k+1} - a_k = a$, the conditional transition probability $\Pr(c_{k+1} = j|c_k = i, a_{k+1} - a_k = a)$ is determined as

$$\begin{aligned} &\Pr(c_{k+1} = j|c_k = i, a_{k+1} - a_k = a) \\ &= \begin{cases} 0 & 1 \le j < (i-a)^+ + 1 \\ (1-\mu_r)^{j-(i-a)^+-1} \cdot \mu_r & (i-a)^+ + 1 \le j < \tau \\ (1-\mu_r)^{j-(i-a)^+-1} & j = \tau \end{cases} \\ &= \begin{cases} 0 & 1 \le a < (i-j)^+ + 1 \\ (1 \;-\; \mu_r)^{j-(i-a)^+-1} \;\cdot \\ (\mathbf{1}_{j=\tau}(1-\mu_r) + \mu_r) & a \ge (i-j)^+ + 1 \end{cases}, \end{aligned} \tag{29}$$

where $(\cdot)^+$ refers to the function $\max\{\cdot, 0\}$.

Proof: We prove the lemma case by case.

For $1 \leq j < (i-a)^+ + 1$, it is easy to verify that constraint (27b) is not satisfied, therefore the corresponding probability is always zero.

For $j \geq (i-a)^+ + 1$, recall that the packets in the relay queue are served in a FCFS manner, which means conditionally the $(k+1)$th packet will not get the opportunity to be delivered during its first $(i-a)^+$ time slots in $\mathbf{R}$. For the packet to be transmitted in the jth time slot in its life, it will not be transmitted for the following $j-(i-a)^+-1$ time slots, and eventually be delivered with probability μ_r.

The case of $j = \tau$ is similar to that of $j \geq (i-a)^+ + 1$, except that the packet is sure to leave the queue in the τth time slot, either transmitted or dropped due to delivery delay constraint. ■

We are now ready to present the transition probability for the Markov chain $\{c_k\}$, with the help of Lemma 3 and Lemma 4.

Theorem 4: Let $\mathbf{P} = (P_{ij})_{\tau\times\tau}$ be the transition matrix for the Markov chain $\{c_k\}$, we have

$$\begin{aligned} P_{ij} &= \Pr(c_{k+1} = j | c_k = i) \\ &= \left(\frac{\lambda_r \left((1-\mu_r-\lambda_r)^{(i-j)^+} - (1-\mu_r-\lambda_r)^{i-1} \right)}{(1-\mu_r)^{1-j+i}(\mu_r+\lambda_r)} \right. \\ &\quad \left. + \left(1 - \frac{\lambda_r}{1-\mu_r}\right)^{i-1} (1-\mu_r)^{j-1} \right) (\mathbf{1}_{j=\tau}(1-\mu_r)+\mu_r). \end{aligned} \tag{30}$$

Proof: We calculate P_{ij} by conditioning on $a_{k+1} - a_k$ and summing over all possible value of it:

$$\begin{aligned} P_{ij} &= \Pr(c_{k+1} = j | c_k = i) \\ &= \sum_{a=1}^{\infty} \Pr(a_{k+1} - a_k = a | c_k = i) \cdot \\ &\qquad \Pr(c_{k+1} = j | c_k = i, a_{k+1} - a_k = a), \end{aligned} \tag{31}$$

with terms therein provided by Lemma 3 and Lemma 4 respectively. Note that both terms are presented as piecewise functions regarding different a ranges, and that $(i-j)^+ + 1 \leq (i-1)^+ + 1 \leq i$, we can acquire the sum over pieces respectively:

$$\begin{aligned} &\sum_{a=1}^{(i-j)^+} \Pr(a_{k+1} - a_k = a | c_k = i) \cdot \\ &\qquad \Pr(c_{k+1} = j | c_k = i, a_{k+1} - a_k = a) \\ &= \sum_{a=1}^{(i-j)^+} 0 = 0; \end{aligned} \tag{32}$$

$$\begin{aligned} &\sum_{a=(i-j)^+ + 1}^{i-1} \Pr(a_{k+1} - a_k = a | c_k = i) \cdot \\ &\qquad \Pr(c_{k+1} = j | c_k = i, a_{k+1} - a_k = a) \\ &= \frac{\lambda_r \left((1-\mu_r-\lambda_r)^{(i-j)^+} - (1-\mu_r-\lambda_r)^{i-1} \right)}{(1-\mu_r)^{1-j+i}(\mu_r+\lambda_r)} \cdot \\ &\qquad (\mathbf{1}_{j=\tau}(1-\mu_r)+\mu_r); \end{aligned} \tag{33}$$

$$\begin{aligned} &\sum_{a=i}^{i} \Pr(a_{k+1} - a_k = a | c_k = i) \cdot \\ &\qquad \Pr(c_{k+1} = j | c_k = i, a_{k+1} - a_k = a) \\ &= \left(1 - \frac{\lambda_r}{1-\mu_r}\right)^{i-1} (1-\mu_r)^{j-1} \cdot \lambda_r \mathbf{1}_{i=\tau} \cdot \\ &\qquad (\mathbf{1}_{j=\tau}(1-\mu_r)+\mu_r); \end{aligned} \tag{34}$$

$$\begin{aligned} &\sum_{a=i+1}^{\infty} \Pr(a_{k+1} - a_k = a | c_k = i) \cdot \\ &\qquad \Pr(c_{k+1} = j | c_k = i, a_{k+1} - a_k = a) \\ &= \left(1 - \frac{\lambda_r}{1-\mu_r}\right)^{i-1} (1-\mu_r)^{j-1} \cdot (1 - \lambda_r \mathbf{1}_{i=\tau}) \cdot \\ &\qquad (\mathbf{1}_{j=\tau}(1-\mu_r)+\mu_r); \end{aligned} \tag{35}$$

and combining all the sums yields the result presented in (30). ■

Since an irreducible Markov chain with finite state space is always positive recurrent, we can get the stationary distribution through $\pi^{(\lambda)}\mathbf{P} = \pi^{(\lambda)}$. Intuitively, $\pi^{(\lambda)}$ presents the time distribution of how long a packet may stay in the relay queue of $\mathbf{R}$, which is sufficient to provide various performance metrics, as shown in the previous section.

Remark 1: To model a Markovian queue, the most common way would be to discuss the backlog of it. However, each packet in our system has its own remaining valid time (see $\mathbf{R}$ in Fig 1), so this approach is no longer sufficient. An intuitive workaround is to involve packet remaining valid time into the state space, so that a vector containing remaining valid times of all packets in the queue is used to describe the queuing state. It is trivial to see that this workaround creates a vast state space, making the problem almost insoluble. Our approach, by monitoring only the features of a leaving packet, successfully presents the desired metrics of the queuing process while avoiding the problem of state space explosion.

V. Numerical Results

This section presents extensive simulation and theoretical results to validate our theoretical models and explore the impact brought by the delivery delay constraint.

A. Simulation Setting

We build a specific JAVA-based simulator, with our system model and routing algorithm implemented. Each simulation task corresponds to a set of network parameters and runs over a period of 1.0×10^7 time slots. For accuracy considerations, we only collect date from the last 80% of time slots of simulation to ensure that the system is in its steady state. The parameter settings adopted in our simulation are summarized in Table I.

TABLE I. SIMULATION PARAMETER RANGES

Parameter	Value	Comments
m	fixed to 200	Number of cells
n	fixed to 359[1]	Number of nodes
τ	$[500, 5000]$, step increment 500	Delivery delay constraint
λ	$[0.005, 0.150]$[2] step increment 0.005	Exogenous input rate

[1] The proportion of n and m is set to match the optimal node density proposed in Corollary 1, which is approximately 1.793.

[2] According to Theorem 1, we have $\mu = 0.1501$, therefore only the range $\lambda < \mu$ is discussed.

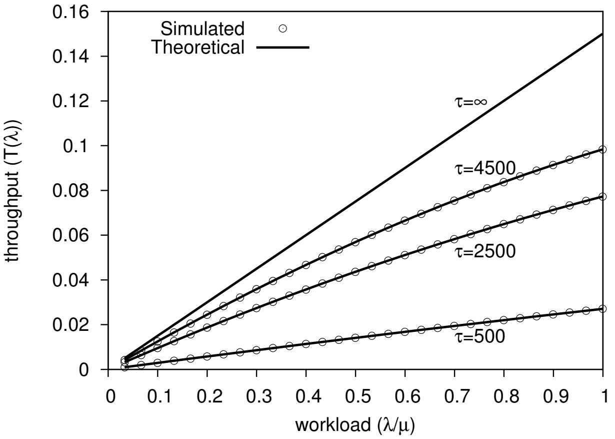

Fig. 4. Theoretical and simulated achievable throughput.

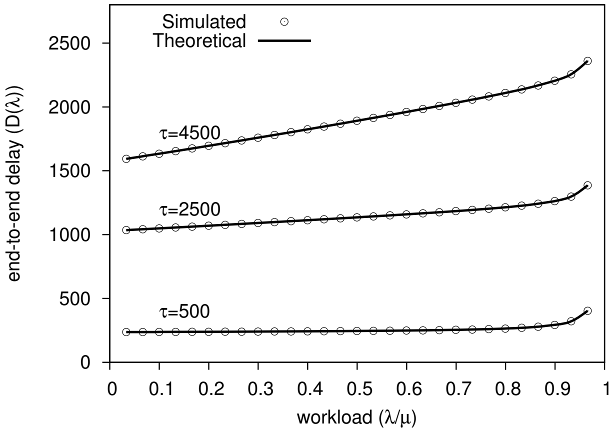

Fig. 5. Theoretical and simulated end-to-end packet delay.

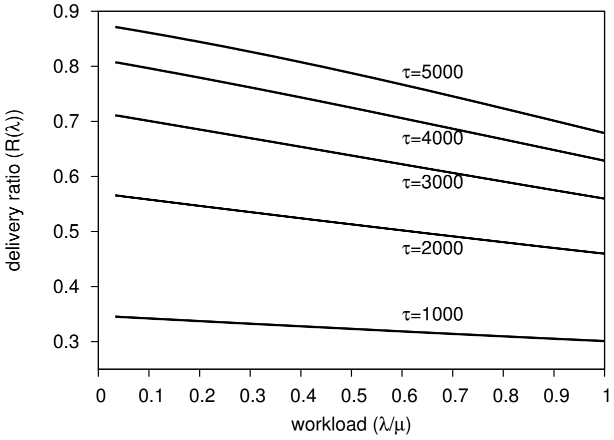

Fig. 6. Impact of delivery delay constraint on the delivery ratio.

B. Model Validation

In Fig. 4, we compare the simulated achievable throughput to the theoretical one obtained from Theorem 2 under different settings of τ. We can see that the theoretical result coincide with the simulated data precisely, indicating that our theoretical model is accurate in depicting the exact throughput under delivery delay constraint. For comparison, we also include in Fig. 4 the workload-throughput curve $T(\lambda) = \lambda$ for the special case of $\tau = \infty$. It is interesting to notice that under any given workload λ/μ, the packet loss rate just corresponds to the difference between the achievable throughput and the corresponding point on the line of $T(\lambda) = \lambda$. We can see that as the delivery delay constraint becomes less stringent (i.e., when τ is larger), packet loss rate tends to decrease and the achievable throughput curve will eventually approach to λ.

Fig. 5 presents both the simulated end-to-end packet delay and the theoretical one from Theorem 3, which shows clearly that our theoretical model for the end-to-end packet delay analysis is also very accurate. An interesting phenomenon we can see from Fig. 5 is that under a certain delivery delay constraint, the end-to-end delay grows slowly and almost linearly when the workload is not heavy and then approach to infinity as $\lambda \to \mu$ because of the queuing process at $\mathbf{S}$. From Corollary 4 we know that as $\tau \to \infty$, the curves will gradually approach to the one shown in (26).

C. Discussions

To further explore the impact brought by the delivery delay constraint, we first examine in Fig. 6 how the delivery ratio (i.e., the ratio of achievable throughput to exogenous packet rate) varies with workload under different settings of τ. We can see that for a given delivery delay constraint τ, the delivery ratio decreases slowly and almost linearly as the workload increases. This is actually surprising and inspiring, because packet loss due to the constraint τ is not very sensitive to the variation of workload and will not dramatically degrade even as the workload approaches unity. Another observation from Fig. 6 is that although the delivery ratio can be increased in the scenario where a less stringent delivery delay constraint is imposed, the delivery ratio tends to be more sensitive to the variation of workload under such scenario. Thus, the results in this figure indicate that a desirable trade-off can be initiated between delivery delay constraint and delivery ratio in a delivery delay-constrained MANET.

We further examine in Fig. 7 the impact of delivery delay constraint on the end-to-end packet delay, another key metric of network performance. As is revealed in the figure that when a more stringent delivery delay constraint is imposed (i.e., when τ is small), the packet delay becomes less sensitive to the variation of workload, which matches the observation from Fig. 5. It can be also observed from the figure that an increase in τ would incur almost the same fraction of increase in the end-to-end delay. This is because that according to (26) where $\tau \to \infty$, the packet delay introduced by waiting time at relay

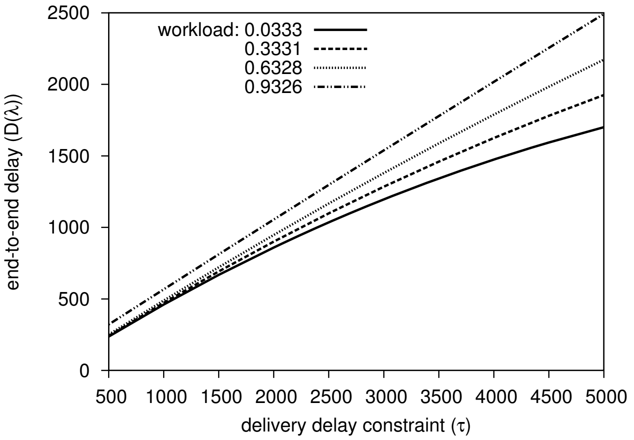

Fig. 7. Impact of delivery delay constraint on the end-to-end packet delay.

queues scales as n, which overwhelms the queuing delay at source $\mathbf{S}$ and dominates the packet end-to-end delay. Thus, we can acquire a desired end-to-end delay by applying an appropriate constraint τ on the delivery delay.

VI. Conclusions

As a step towards practical performance evaluation of MANETs, this paper examined a MANET with the constraint of maximum allowed delivery delay to each packet and investigated the impact brought by the constraint to the network performance in terms of throughput and packet end-to-end delay. The results in this paper indicate that delivery delay constraint may affect network performance in a complicated way. On one hand, a less stringent delivery delay constraint leads to a higher throughput and also a higher delivery ratio, but the delivery ratio there becomes more sensitive to the variation of workload and tends to decrease more rapidly as workload increases. On the other hand, a less stringent delivery delay constraint usually incurs a larger end-to-end delay, which in general increases as workload increase but tends to be more sensitive to the variation of workload. Thus, a graceful tradeoff between throughput and packet delay in a MANET can be initiated by applying an appropriate constraint on delivery delay in such network. It is noted that this paper mainly focuses on the practical constraint of delivery delay, while the constraint on end-to-end delay still remains an open problem, and is left as future work. It is also expected that this work will provide further inspiration for practical performance modeling and analysis for other network scenarios as well.

Acknowledgment

This work is supported by the International Science and Technology Cooperation Program of China under Grant No. 2013DFB10070.

References

[1] A. Goldsmith, M. Effros, R. Koetter, M. Médard, A. Ozdaglar, and L. Zheng, "Beyond shannon: the quest for fundamental performance limits of wireless ad hoc networks," *IEEE Commun. Mag.*, vol. 49, no. 5, pp. 195–205, 2011.

[2] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 388–404, 2000.

[3] M. Franceschetti, O. Dousse, D. N. Tse, and P. Thiran, "Closing the gap in the capacity of wireless networks via percolation theory," *IEEE Trans. Inf. Theory*, vol. 53, no. 3, pp. 1009–1018, 2007.

[4] A. Ozgur, O. Lévêque, and D. N. Tse, "Hierarchical cooperation achieves optimal capacity scaling in ad hoc networks," *IEEE Trans. Inf. Theory*, vol. 53, no. 10, pp. 3549–3572, 2007.

[5] C. Peraki and S. D. Servetto, "On the maximum stable throughput problem in random networks with directional antennas," in *Proc. 4th ACM Int. Symp. Mobile Ad Hoc Networking & Computing*. ACM, 2003, pp. 76–87.

[6] A. Keshavarz-Haddad, V. Ribeiro, and R. Riedi, "Broadcast capacity in multihop wireless networks," in *Proc. 12th Annu. Int. Conf. Mobile Computing and Networking*. ACM, 2006, pp. 239–250.

[7] X.-Y. Li, "Multicast capacity of wireless ad hoc networks," *IEEE/ACM Trans. Netw.*, vol. 17, no. 3, pp. 950–961, 2009.

[8] M. Grossglauser and D. Tse, "Mobility increases the capacity of ad-hoc wireless networks," in *INFOCOM 2001. 20th Annu. Joint Conf. IEEE Computer and Communications Societies. Proc. IEEE*, vol. 3. IEEE, 2001, pp. 1360–1369.

[9] A. E. Gamal, J. Mammen, B. Prabhakar, and D. Shah, "Throughput-delay trade-off in wireless networks," in *INFOCOM 2004. 23rd Annu. Joint Conf. IEEE Computer and Communications Societies*, vol. 1. IEEE, 2004.

[10] X. Lin, G. Sharma, R. R. Mazumdar, and N. B. Shroff, "Degenerate delay-capacity tradeoffs in ad-hoc networks with brownian mobility," *IEEE/ACM Trans. Netw.*, vol. 14, no. SI, pp. 2777–2784, 2006.

[11] M. J. Neely and E. Modiano, "Capacity and delay tradeoffs for ad hoc mobile networks," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 1917–1937, 2005.

[12] Y. Wang, X. Chu, X. Wang, and Y. Cheng, "Optimal multicast capacity and delay tradeoffs in manets: A global perspective," in *INFOCOM, 2011 Proc. IEEE*. IEEE, 2011, pp. 640–648.

[13] G. Sharma, R. Mazumdar, and N. B. Shroff, "Delay and capacity tradeoffs in mobile ad hoc networks: A global perspective," *IEEE/ACM Trans. Netw.*, vol. 15, no. 5, pp. 981–992, 2007.

[14] L. Ying, S. Yang, and R. Srikant, "Optimal delay–throughput tradeoffs in mobile ad hoc networks," *IEEE Trans. Inf. Theory*, vol. 54, no. 9, pp. 4119–4143, 2008.

[15] R. Urgaonkar and M. J. Neely, "Network capacity region and minimum energy function for a delay-tolerant mobile ad hoc network," *IEEE/ACM Trans. Netw.*, vol. 19, no. 4, pp. 1137–1150, 2011.

[16] J. Liu, X. Jiang, H. Nishiyama, and N. Kato, "Delay and capacity in ad hoc mobile networks with f-cast relay algorithms," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 2738–2751, 2011.

[17] G. Juntao, L. Jiajia, X. Jiang, O. Takahashi, and N. Shiratori, "Throughput capacity of manets with group-based scheduling and general transmission range," *IEICE Trans. Commun.*, vol. 96, no. 7, pp. 1791–1802, 2013.

[18] Y. Chen, J. Liu, X. Jiang, and O. Takahashi, "Throughput analysis in mobile ad hoc networks with directional antennas," *Ad Hoc Networks*, vol. 11, no. 3, pp. 1122–1135, 2013.

[19] E. Perevalov and R. S. Blum, "Delay-limited throughput of ad hoc networks," *IEEE Trans. Commun.*, vol. 52, no. 11, pp. 1957–1968, 2004.

[20] A. A. Hanbali, P. Nain, and E. Altman, "Performance of ad hoc networks with two-hop relay routing and limited packet lifetime," in *Proc. 1st Int. Conf. Performance Evaluation Methodolgies and Tools*. ACM, 2006, p. 49.

[21] S. Zhou and L. Ying, "On delay constrained multicast capacity of large-scale mobile ad-hoc networks," in *INFOCOM, 2010 Proc. IEEE*. IEEE, 2010, pp. 1–5.

[22] S. Toumpis and A. J. Goldsmith, "Large wireless networks under fading, mobility, and delay constraints," in *INFOCOM 2004. 23rd Annu. Joint Conf. IEEE Computer and Communications Societies*, vol. 1. IEEE, 2004.

[23] X. Wang, W. Huang, S. Wang, J. Zhang, and C. Hu, "Delay and capacity tradeoff analysis for motioncast," *IEEE/ACM Trans. Netw.*, vol. 19, no. 5, pp. 1354–1367, 2011.