End-to-End Delay Modeling in Buffer-Limited MANETs: A General Theoretical Framework

Jia Liu, Min Sheng, *Member, IEEE*, Yang Xu, Jiandong Li, *Senior Member, IEEE*, and Xiaohong Jiang, *Senior Member, IEEE*

***Abstract*—This paper focuses on a class of important two-hop relay mobile ad hoc networks (MANETs) with limited-buffer constraint and any mobility model that leads to the uniform distribution of the locations of nodes in steady state, and develops a general theoretical framework for the end-to-end (E2E) delay modeling there. We first combine the theories of fixed-point (FP), quasi-birth-and-death process, and embedded Markov chain to model the limiting distribution of the occupancy states of a relay buffer, and then apply the absorbing Markov chain theory to characterize the packet delivery process, such that a complete theoretical framework is developed for the E2E delay analysis. With the help of this framework, we derive a general and exact expression for the E2E delay based on the modeling of both packet queuing delay and delivery delay. To demonstrate the application of our framework, case studies are further provided under two network scenarios with different MAC protocols to show how the E2E delay can be analytically determined for a given network scenario. Finally, we present extensive simulation and numerical results to illustrate the efficiency of our delay analysis as well as the impacts of network parameters on delay performance.**

***Index Terms*—Mobile ad hoc networks (MANETs), limited buffer, end-to-end delay, performance modeling.**

I. Introduction

With the development of wireless communication technologies, mobile ad hoc networks (MANETs) have become an appealing candidate for many critical applications, such as emergency rescue, disaster relief, coverage extension for cellular networks, etc. [1]–[5]. Although lots of work has been done to facilitate the commercialization of MANETs, understanding their fundamental delay performance has been a critical research issue for them to support various applications with different quality of service (QoS) requirements [6], [7].

Manuscript received April 11, 2015; revised June 29, 2015; accepted August 19, 2015. Date of publication September 1, 2015; date of current version January 7, 2016. This work was supported in part by Japan JSPS under Grant 15H02692, in part by China NSFC under Grant 61571352, Grant 61231008, Grant 61172079, and Grant 91338114, and in part by China 111 Project Grant B08038. The associate editor coordinating the review of this paper and approving it for publication was Prof. Jianwei Huang. *(Corresponding author: Min Sheng.)*

J. Liu is with the School of Systems Information Science, Future University Hakodate, Hakodate 041-8655, Japan, and also with the State Key Laboratory of ISN, Xidian University, Xi'an 710071, China (e-mail: jliu871219@gmail.com; liujia@mail.xidian.edu.cn).

X. Jiang is with the School of Systems Information Science, Future University Hakodate, Hakodate 041-8655, Japan (e-mail: jiang@fun.ac.jp).

M. Sheng and J. Li are with the State Key Laboratory of ISN, Xidian University, Xi'an 710071, China (e-mail: mshengxd@gmail.com; jdli@ieee.org).

Y. Xu is with the School of Economics and Management, Xidian University, Xian 710071, China (e-mail: yxu@xidian.edu.cn).

Digital Object Identifier 10.1109/TWC.2015.2475258

End-to-end (E2E) delay, the time that a packet takes to reach its destination after it is generated by its source, serves as the most fundamental delay metric. The available theoretical studies on E2E delay of MANETs mainly focus on deriving its upper bound or approximation. Regarding the delay upper bound of MANETs, Neely *et al.* [8] derived some useful results for a cell-partitioned MANET with the two-hop relay (2HR) routing scheme and i.i.d mobility model. Later, Gamal *et al.* [9] and Sharma *et al.* [10] extended the results of [8] to the continuous network model and general mobility model, respectively. Inspired by these works, extensive research activities have been devoted to the study of delay upper bound for MANETs under various network scenarios, such as under the motioncast in [11], under the cognitive networks in [12], under the packet redundancy in [13], under the multi-hop back-pressure routing in [14], and under the power control in [15]. Regarding the delay approximation, Jindal *et al.* [16] explored recently the E2E delay approximation for MANETs with multi-hop relay routing, and Liu *et al.* studied the E2E delay approximation for MANETs with probing-based 2HR routing [17] and limited packet redundancy [18].

In addition to the studies on delay upper bound or approximation for MANETs, Neely *et al.* [8] also applied the queueing theory to derive the exact expression for E2E delay. Following this line, recently some results have been reported on the modeling of really achievable E2E delay in MANETs. Chen *et al.* [19] explored the MANETs with Aloha MAC protocol and determined the corresponding exact E2E delay there under the continuous network model. For a cell-partitioned MANET with broadcast-based routing scheme, Gao *et al.* [20] proposed a new theoretical framework for the analysis of its exact E2E delay based on the theory of Quasi-Birth-and-Death process.

It is notable that the common limitation of above studies is that to simplify their analysis of E2E delay, they assume the relay buffer of a node, which is used for temporarily storing packets of other nodes, has an infinite buffer size. In a practical MANET, however, the buffer size of a mobile node is usually limited due to both its storage space limitation and computing capability limitation. Thus, for the practical delay performance study of MANETs, the constraint on buffer space should be carefully addressed. Notice that the E2E delay modeling with practical limited-buffer constraint still remains a technical challenge. This is mainly due to the lack of a general theoretical framework to efficiently characterize the highly dynamic behaviors in such networks, like the complicated buffer occupancy states of a relay buffer, as well as the highly dynamic queuing process and delivery process of a packet.

As a step towards the modeling of real achievable E2E delay for the practical MANETs with buffer constraint, we focus on a class of important 2HR MANETs with limited shared relay buffer and propose a general theoretical framework for the E2E delay modeling there. The main contributions of this paper are summarized as follows.

- For the concerned MANET, we first combine the theories of Fixed-Point (FP), Quasi-Birth-and-Death (QBD) process and embedded Markov chain (EMC) to construct an analytical model to fully depict the complicated occupancy behaviors of a relay buffer with limited buffer size.
- Based on the above modeling of relay buffer occupancy behaviors, we then apply the absorbing Markov chain (AMC) theory to characterize the packet delivery process, such that a complete theoretical framework is developed for the E2E delay modeling in the concerned buffer-limited MANETs. This framework is general in the sense that it can be applied to conduct E2E delay analysis for a 2HR MANET with any mobility model that leads to the uniform distribution of the locations of nodes, such as the i.i.d mobility model [8], the random walk model [9], the random way-point model [21], etc..
- To demonstrate the application of the proposed framework, case studies are further provided under two network scenarios, i.e., the cell partitioned networks with local scheduling-based MAC protocol (LS-MAC) [8] and Equivalent-Class based MAC protocol (EC-MAC) [22], to show how the E2E delay can be analytically determined for a given network scenario by applying our framework. Finally, extensive simulation and numerical results are provided to validate the efficiency of the proposed E2E delay model and also to illustrate the impacts of network parameters on delay performance.

The remainder of this paper is organized as follows. Section II introduces preliminaries involved in this paper. The complicated relay buffer occupancy behaviors are analyzed in Section III. We derive the queuing delay, delivery delay and E2E delay in Section IV, and conduct case studies in Section V. The simulation results and corresponding discussions are provided in Section VI. Finally, Section VII concludes this paper.

II. Preliminaries

In this section, we first present some basic assumptions and the buffer constraint, and then introduce the routing scheme and some critical definitions involved in this study.

A. Basic Assumptions

We consider the following minimal set of assumptions:

(A.i) The ad hoc network is time-slotted and consists of n mobile nodes.

(A.ii) The packet generating process in each source node is independent and assumed to be a Bernoulli process, where a packet is generated by its source node with probability λ in a time slot.

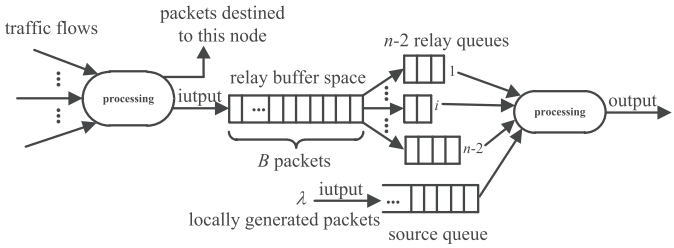


Fig. 1. Illustration of buffer structure of a node.

(A.iii) The widely-used permutation traffic model [8], [23], [24] is adopted. With this traffic model, there are n unicast traffic flows in the network, each node is the source of one traffic flow and also the destination of another traffic flow. We denote by $\varphi(i)$ the destination node of the traffic flow originated from node i, then the source-destination pairs are matched in a way that the sequence $\{\varphi(1), \varphi(2), \cdots, \varphi(n)\}$ is just a derangement of the set of nodes $\{1, 2, \cdots, n\}$.

(A.iv) During a time slot the total amount of data that can be transmitted from a transmitter to its corresponding receiver is fixed and normalized to one packet.

(A.v) We consider the mobility model that leads to the uniform distribution of the locations of nodes in steady state, which covers many typical mobility models such as the i.i.d mobility model, the random walk model, the random way-point model, etc.. More formally, we denote by $X_i(t)$ the location of ith node at time slot t and assume the process $\{X_i(\cdot)\}$ is stationary and ergodic with stationary distribution uniform on the network area; moreover, the trajectories of different nodes are independent and identically distributed.

B. Buffer Constraint

As illustrated in Fig. 1, each node in the MANET maintains $n-1$ individual queues, one source queue for storing the packets that are locally generated at this node, and $n-2$ parallel relay queues for storing packets of other flows (one queue per flow). All these queues follow the FIFO (first-in-first-out) discipline.

Similar to the available studies on buffer-limited wireless networks [25], [26], we consider the following practical buffer constraint that all the $n-2$ relay queues of a node share a common relay buffer with the limited buffer size of B packets, while the buffer size of source queue is unlimited. We adopt this buffer constraint here mainly due to the following reasons. First, the mathematical tractability of this assumption allows us to gain important insights into the structure of E2E delay analysis. Second, the analysis under this assumption provides a meaningful theoretical result in the limit of infinite source buffer. Third, in a practical wireless network, each node usually prefers to reserve a much larger buffer space for storing its own packets than that for storing packets of other flows.

Also, even in the case that the buffer space of source queue is not enough when bursty traffic comes, the congestion control in the upper layer can be executed to avoid the loss of locally generated packets [25].

C. Handshake-Based 2HR Scheme

Regarding the routing scheme, we focus on the 2HR scheme, because it is simple yet efficient and thus serves as a class of attractive routing protocols for MANETs [8], [23]. To avoid unnecessary packet loss and support the efficient operation of the concerned buffer-limited MANETs, we introduce a handshake mechanism with negligible overhead[1] into the 2HR scheme such that the packet dropping will not happen even in the case of relay buffer overflow. Once a node (say **S**) gets access to the wireless channel in a time slot, it executes the new handshake-based 2HR (H2HR for short) routing scheme summarized in Algorithm 1.

Algorithm 1 H2HR algorithm

1: **if** The destination **D** is within the transmission range of **S then**
2: **S** executes **Procedure** 1.
3: **else if** There exist other nodes within the transmission range of **S then**
4: With equal probability, **S** selects one node as the receiver.
5: **S** executes **Procedure** 2 or **Procedure** 3 equally with the receiver.
6: **end if**

Procedure 1 Source-to-destination (**S-D**) transmission

1: **if S** has packets in its source queue **then**
2: **S** transmits the head-of-line (HoL) packet in its source queue to **D**.
3: **S** removes the HoL packet from its source queue.
4: **S** moves ahead the remaining packets in its source queue.
5: **else**
6: **S** remains idle.
7: **end if**

Procedure 2 Source-to-relay (**S-R**) transmission

1: **if S** has packets in its source queue **then**
2: **S initiates a handshake with the receiver to check whether the relay buffer of receiver is full or not.**
3: **if** The relay buffer of receiver does not overflow **then**
4: The receiver dynamically allocates a new buffer space to the end of the corresponding relay queue.
5: **S** transmits the HoL packet in its source queue to the receiver.
6: **S** removes the HoL packet from its source queue.
7: **S** moves ahead the remaining packets in its source queue.
8: **end if**
9: **else**
10: **S** remains idle.
11: **end if**

Procedure 3 Relay-to-destination (**R-D**) transmission

1: **if S** has packets destined to the receiver **then**
2: **S** transmits the HoL packet in its corresponding relay queue to the receiver.
3: **S** removes the HoL packet from this relay queue.
4: **S** moves ahead the remaining packets in this relay queue.
5: This relay queue releases one buffer space to the common relay buffer of **S**.
6: **else**
7: **S** remains idle.
8: **end if**

D. Definitions

Here we introduce some important definitions involved in this study.

Relay-buffer Overflowing Probability (ROP): For the concerned MANET with a given packet generating rate λ in each node, the relay-buffer overflowing probability $p_o(\lambda)$ of a node is defined as the probability that the relay buffer of this node overflows (i.e, the relay buffer is full).

Queuing Delay: The queuing delay is defined as the time it takes a packet to move to HoL in the source queue (i.e., the source node starts to deliver it) after it is generated by its source.

Delivery Delay: The delivery delay is defined as the time it takes a packet to reach its destination after its source starts to deliver it.

End-to-end Delay: The end-to-end delay is defined as the time it takes a packet to reach its destination after it is generated by its source, which is the sum of its queuing delay and delivery delay.

III. Relay Buffer Analysis

In this section, we first introduce three basic probabilities. Based on these probabilities, we then apply the QBD process modeling and EMC technique to depict the occupancy behaviors of a relay buffer. Finally, we construct a self-mapping function for the ROP $p_o(\lambda)$ (i.e., $p_o(\lambda)$ is the fixed-point of this function) to determine the limiting distribution of the occupancy states, which will help us conduct delay analysis in Section IV.

Due to the symmetry of nodes and traffic flows, we only focus on one node **S** in the following analysis. We denote by p_{sd}, p_{sr} and p_{rd} the probabilities that in a time slot **S** gets access to the wireless channel and decides to execute **S-D**, **S-R** and **R-D** transmission respectively[2]. These probabilities can be

[1]The handshake mechanism can be easily implemented by sending only one indicator bit from the receiver to the transmitter (e.g., bit 0 when the relay buffer is full, and bit 1 otherwise), so the impact of this overhead can be neglected in our analysis.

[2]It is notable that $p_{sr} = p_{rd}$, and executing a transmission doesn't mean that **S** will successfully transmit a packet in this time slot.

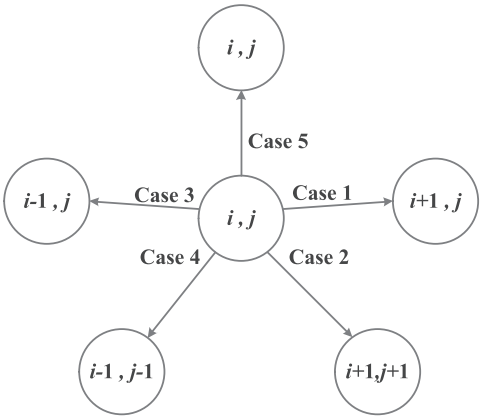


Fig. 2. Transition cases from a general state (i, j).

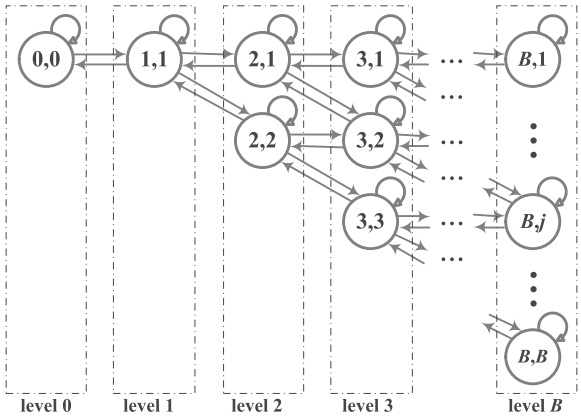


Fig. 3. State transition diagram of the QBD process.

determined under a given network scenario and the derivation of them will be elaborated in case studies.

A. QBD Process Modeling

Regarding the source queue of **S**, it can be modeled as a Bernoulli/Bernoulli queue [27] with packet arrival rate λ and service rate $\mu_s(\lambda)$, where $\mu_s(\lambda)$ is given by

$$\mu_s(\lambda) = p_{sd} + p_{sr}\,(1 - p_o(\lambda))\,. \tag{1}$$

Due to the reversibility of Bernoulli/Bernoulli queue, the packet departure process of source queue is also a Bernoulli process with rate λ.

Regarding the relay buffer of **S**, we adopt a two-tuple $\mathbf{X}(t) = (I(t), J(t))$ to define its state at time slot t, where $I(t)$ denotes the number of packets occupying the relay buffer, and $J(t)$ denotes the number of relay queues which are not empty, here $0 \le I(t) \le B$, $1 \le J(t) \le I(t)$ when $I(t) > 0$, and $J(t) = 0$ when $I(t) = 0$.

As illustrated in Fig. 2, suppose that the relay buffer of **S** is in state (i, j) at the current time slot, only one of the following transitions may happen in the next time slot:

- Case 1: $i < B$, a packet enters the relay buffer, and this packet is destined for a destination same as one of packet(s) already in relay queues.
- Case 2: $i < B$, a packet enters the relay buffer, and the destination of this packet is different from all packet(s) already in relay queues.
- Case 3: $i > 0$, a packet in one of the relay queues is delivered to its destination, and there still exist other packet(s) in this relay queue.
- Case 4: $i > 0$, a packet in one of the relay queues is delivered to its destination, and there is no remaining packet in this relay.
- Case 5: no packet enters into or departs from the relay queues.

To facilitate our discussion, we call the subset of states $L_i = \{(i, 1), (i, 2), \ldots, (i, i)\}$ level i, $L_0 = \{(0, 0)\}$ level 0, and state (i, j) that the relay buffer is in level i and phase j. Notice that when the relay buffer is in some state of level i at a time slot, the next state after one-step state transitions could only be some state in the same level or its adjacent levels. Thus, as time evolves, the state transitions of the relay buffer of **S** form a two-dimensional QBD process $\{\mathbf{X}(t), t = 0, 1, 2, \cdots\}$ [28], [29]. According to the transition cases in Fig. 2, the overall transition diagram of the QBD process is summarized in Fig. 3. There are in total $1 + 0.5B(1 + B)$ states for the QBD process, and we arrange all these states in a low-to-high level and low-to-high phase way as follows: $\{(0, 0), (1, 1), (2, 1), (2, 2), \cdots, (B, B)\}$. Then the corresponding state transition matrix **P** of the QBD process can be determined as

$$\mathbf{P} = \begin{bmatrix} \mathbf{A_{0,0}} & \mathbf{A_{0,1}} & & & & \\ \mathbf{A_{0,1}} & \mathbf{A_{1,1}} & \mathbf{A_{1,2}} & & & \\ & \ddots & \ddots & \ddots & & \\ & & \mathbf{A_{B-1,B-2}} & \mathbf{A_{B-1,B-1}} & \mathbf{A_{B-1,B}} & \\ & & & \mathbf{A_{B,B-1}} & \mathbf{A_{B,B}} & \end{bmatrix}, \tag{2}$$

where the sub-matrix $\mathbf{A_{i,l}}$ is of size $i \times l$ ($\mathbf{A_{0,0}}$, $\mathbf{A_{0,1}}$ and $\mathbf{A_{1,0}}$ are of size 1×1), denoting the transition probabilities from the states of level i to the states of level l.

It is notable that in our QBD process of relay buffer, different levels have different number of phases, and the transition probabilities of one state depend on its level, thus the QBD process is level-dependent and it is very difficult to solve its limiting distribution by determining its critical matrices and conducting recursive algorithm [28], [29]. To address this issue, we adopt a Markov chain-collapsing technique [30], [31] to convert the two-dimensional QBD process to a one-dimensional EMC in the next subsection.

B. Collapsing to an EMC

For the QBD process of Fig. 3, we integrate all states of a level into only one state, then the two-dimensional QBD process is collapsed to a one-dimensional Embedded Markov Chain (EMC). As illustrated in Fig. 4, one state L_i of the EMC corresponds to one level i of the QBD process, and $p_L^{(i,l)}$ denotes the one-step transition probability from state L_i to state L_l in the EMC. According to the EMC theory [30], [31], the

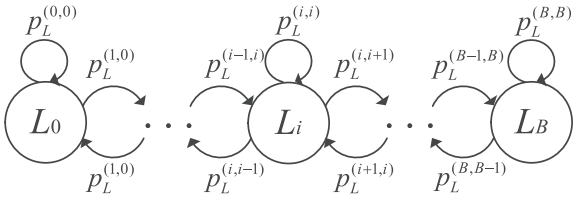


Fig. 4. State machine of the EMC.

state transition probability of the EMC is the phase-averaged state transition probability of the QBD process, then we have

$$p_L^{(i,l)} = \begin{cases} p_{(0,0),L_l}, & i = 0 \\ \sum_{j=1}^{i} p_{(i,j),L_l} \cdot P_{j|L_i}, & 1 \le i \le B \end{cases} \tag{3}$$

where $p_{(i,j),L_l}$ denotes the transition probability from state (i, j) to the states of level l, and $P_{j|L_i}$ denotes the conditional probability that the relay buffer is in phase j given that it is in level i. Based on formula (3) as well as the ergodic and uniform features of the distribution of node location, we have the following lemma regarding the transition probabilities of the EMC.

Lemma 1: The one-step transition probability $p_L^{(i,l)}$ of the EMC is determined as

$$p_L^{(i,l)} = \begin{cases} \rho_s(\lambda) \cdot p_{sr}, & l = i+1 \le B \\ \dfrac{i}{n-3+i} \cdot p_{rd}, & l = i-1 \ge 0 \\ 1 - p_L^{(i,i+1)} - p_L^{(i,i-1)}, & l = i \\ 0, & \text{others} \end{cases} \tag{4}$$

where $\rho_s(\lambda) = \frac{\lambda}{\mu_s(\lambda)} = \frac{\lambda}{p_{sd}+p_{sr}(1-p_o(\lambda))}$.

Proof: See Appendix A for the proof. ■

We arrange all the states of EMC in a low-to-high level way as follows: $\{L_0, L_1, \cdots, L_B\}$. Then the corresponding state transition matrix $\mathbf{P}_{EMC}$ can be determined as

$$\mathbf{P}_{EMC} = \begin{bmatrix} p_L^{(0,0)} & p_L^{(0,1)} & & & \\ p_L^{(1,0)} & p_L^{(1,1)} & p_L^{(1,2)} & & \\ & \ddots & \ddots & \ddots & \\ & & p_L^{(B,B-1)} & p_L^{(B,B)} \end{bmatrix}. \tag{5}$$

C. Constructing the Self-Mapping Function

With the help of transition matrix $\mathbf{P}_{EMC}$, we then construct a self-mapping function for $p_o(\lambda)$, i.e., $p_o(\lambda)$ is the fixed-point of this function [32], such that $p_o(\lambda)$ as well as the limiting distribution of the occupancy states of a relay buffer can be determined.

From the the state machine of EMC in Fig. 4 and the transition matrix $\mathbf{P}_{EMC}$, we can see that: 1) the EMC is irreducible; 2) each state L_i is recurrent; 3) the period of each state L_i is 1, so each state is aperiodic. Based on these properties, we can conclude that the EMC is ergodic, thus its limiting distribution $\mathbf{\Pi_L} = [\pi_{L_0}, \pi_{L_1}, \cdots, \pi_{L_B}]$ exists and is unique, and is same as its stationary distribution [33]. Then we have

$$\mathbf{\Pi_L} \cdot \mathbf{P}_{EMC} = \mathbf{\Pi_L}, \tag{6}$$

$$\mathbf{\Pi_L} \cdot \mathbf{1} = 1, \tag{7}$$

where $\mathbf{1}$ is a column vector of size $(B+1) \times 1$ with all elements being 1, and equation (7) follows from the normalization property of a probability vector. Combining (6) with (7) we have

$$\pi_{L_i} = \frac{\mathrm{C}_i \cdot \rho_s(\lambda)^i}{\sum_{k=0}^{B} \mathrm{C}_k \cdot \rho_s(\lambda)^k}, \tag{8}$$

where $\mathrm{C}_i = \binom{n-3+i}{i}$.

It is notable that the relay buffer overflows when it is in level B, then the critical self-mapping function for $p_o(\lambda)$ is constructed as

$$p_o(\lambda) = f\left(p_o(\lambda)\right) = \pi_{L_B} = \frac{\mathrm{C}_B \cdot \rho_s(\lambda)^B}{\sum_{k=0}^{B} \mathrm{C}_i \cdot \rho_s(\lambda)^k}. \tag{9}$$

Given a packet generating rate λ, the self-mapping function doesn't contain any unknown parameters except $p_o(\lambda)$. Thus by solving equation (9), we can determine the ROP $p_o(\lambda)$ corresponding to a given λ, and the limiting distribution of the EMC can be recursively determined as

$$\pi_{L_i} = p_o(\lambda) \cdot \rho_s(\lambda)^{i-B} \cdot \frac{\mathrm{C}_i}{\mathrm{C}_B}. \tag{10}$$

The limiting distribution $\Pi = [\pi_{0,0}, \pi_{1,1}, \cdots, \pi_{i,j}, \cdots, \pi_{B,B}]$ of the QBD process can be further determined as

$$\pi_{i,j} = \pi_{L_i} \cdot P_{j|L_i}, \tag{11}$$

where $P_{j|L_i}$ is given by formula (42) in Appendix A.

Remark 1: Notice that if we don't apply the handshake mechanism, we can also develop the corresponding theoretical framework in the same way to model the relay buffer occupancy process, where the ROP $p_o(\lambda)$ derived in (9) just corresponds to the packet dropping probability.

IV. Delay Analysis

With the help of ROP and limiting distribution of occupancy states of a relay buffer, in this section we analyze the delay performance for the concerned buffer-limited MANET. We denote by Q, D and T the queuing delay, delivery delay and E2E delay of a packet respectively. The E2E delay of a packet will be derived by computing its queuing delay and delivery delay respectively. The queuing delay will be obtained by analyzing the queuing process of the source queue, while the delivery delay will be derived by modeling the packet delivery process as an AMC and analyzing the time the chain takes to enter the absorbing state.

Before presenting our main results on the delay performance, we first provide the following lemma regarding the per node throughput capacity, which is the maximal packet generating rate the MANET can stably support, and the corresponding delay can then be determined.

Lemma 2: For the considered MANET, its per node throughput capacity μ is given by

$$\mu = p_{sd} + p_{sr}\frac{B}{n-2+B}. \tag{12}$$

Proof: See Appendix B for the proof. ■

A. Queuing Delay

Considering a given packet generating rate λ ($\lambda < \mu$), the corresponding ROP $p_o(\lambda)$ can be obtained by solving equation (9), and the service rate of source queue $\mu_s(\lambda)$ can be further determined by formula (1). Thus, in the following analysis, we use p_o and μ_s to represent $p_o(\lambda)$ and $\mu_s(\lambda)$ respectively if there is no ambiguous. Notice that the source queue is a Bernoulli/Bernoulli queue, thus its average queue length $\overline{L}_{source}$ is given by [27]

$$\overline{L}_{source} = \frac{\lambda - \lambda^2}{\mu_s - \lambda}. \tag{13}$$

According to the Little's Law [34], the average delay of a packet in its source queue $\mathbb{E}\{D_s\}$ is given by

$$\mathbb{E}\{D_s\} = \frac{1-\lambda}{\mu_s - \lambda}. \tag{14}$$

Then, the expected queuing delay $\mathbb{E}\{Q\}$ is determined as

$$\mathbb{E}\{Q\} = \mathbb{E}\{D_s\} - \frac{1}{\mu_s} = \frac{\lambda(1-\mu_s)}{\mu_s(\mu_s - \lambda)}. \tag{15}$$

B. Delivery Delay and End-to-end Delay

We present the following theorem regarding the expected E2E delay of the concerned buffer-limited MANET.

Theorem 1: (**Main result**) For the concerned MANET with number of nodes n, relay buffer size B and packet generating rate λ ($\lambda < \mu$), the expected delivery delay $\mathbb{E}\{D\}$ and the expected E2E delay $\mathbb{E}\{T\}$ of a packet are determined as

$$\mathbb{E}\{D\} = \frac{1 + (n-2+\Psi_{n,B,\lambda})(1-p_o)}{\mu_s}, \tag{16}$$

$$\mathbb{E}\{T\} = \frac{1-\lambda}{\mu_s - \lambda} + \frac{(n-2+\Psi_{n,B,\lambda})(1-p_o)}{\mu_s}, \tag{17}$$

where $\Psi_{n,B,\lambda} = \frac{\sum_{i=0}^{B-1} i\mathrm{C}_i \cdot \rho_s^i}{\sum_{i=0}^{B-1} \mathrm{C}_i \cdot \rho_s^i}$.

Proof: We focus on a packet y which is the HoL packet of the source queue at time slot t, then in the next time slot, y will be delivered to its destination with probability p_{sd}, be forwarded to a relay node with probability $p_{sr} \cdot (1-p_o)$, and still stay in the source queue with probability $1-\mu_s$. Thus, the delivery process of packet y can be modeled as an absorbing Markov chain as illustrated in Fig. 5, where S, R and D denote the states that y is in source queue, forwarded to a relay, and delivered to its destination, respectively. We denote by $\overline{X}_S$ and $\overline{X}_R$ the average transition times from the transient states S and R to the absorbing state D, respectively. Then we have

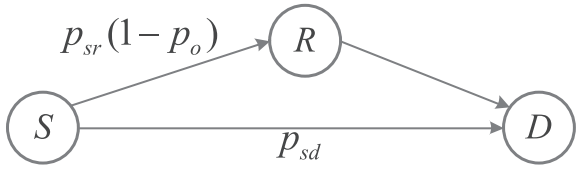


Fig. 5. The absorbing Markov chain for a focused packet delivery.

$$\overline{X}_S = 1 + \overline{X}_S \cdot (1-\mu_s) + \overline{X}_R \cdot p_{sr}(1-p_o), \tag{18}$$

$$\overline{X}_S = \frac{1 + \overline{X}_R \cdot p_{sr}(1-p_o)}{\mu_s}. \tag{19}$$

We denote by a probability vector $\mathbf{P} = (p_0, p_1, \cdots, p_{B-1})$ the steady state distribution of the corresponding relay queue of y in a relay node $\mathbf{R}$, where each element p_i denotes the probability that when y enters the relay queue, there are i packets already in this queue. Notice that the location of each node is stationary and ergodic with stationary distribution uniform on the network area, thus when $\mathbf{R}$ conducts the **R-D** transmission with probability p_{rd} in a time slot, it will deliver a packet for each of the $n-2$ traffic flows with equal probability. Thus, if there are i packets already in the relay queue of y, the expected time elapsed for y to be delivered to its destination is $(i+1) \cdot \left(\frac{p_{rd}}{n-2}\right)^{-1}$. Then we have

$$\overline{X}_R = p_0 \cdot \frac{n-2}{p_{rd}} + 2p_1 \cdot \frac{n-2}{p_{rd}} + \cdots + B \cdot p_{B-1} \cdot \frac{n-2}{p_{rd}} \tag{20}$$

$$= \frac{n-2}{p_{rd}}\{1 + p_1 + 2p_2 + \cdots + (B-1)p_{B-1}\} \tag{21}$$

$$= \frac{n-2}{p_{rd}}(1 + \overline{L}^*_{relay}), \tag{22}$$

where (21) follows from the normalization property of a probability vector, and $\overline{L}^*_{relay}$ is the average queue length of a relay queue, under the condition that the relay buffer is not full.

We denote by $\mathbf{\Pi}^*_{\mathbf{L}} = (\pi^*_{L_0}, \pi^*_{L_1}, \cdots, \pi^*_{L_{B-1}})$ the limiting distribution of the level of a relay buffer, under the condition that the relay buffer is not full, then we have

$$\pi^*_{L_i} = \frac{\pi_{L_i}}{1-\pi_{L_B}} = \frac{\mathrm{C}_i \rho_s^i}{\sum_{k=0}^{B-1} \mathrm{C}_k \cdot \rho_s^k}, \tag{23}$$

and the corresponding conditional average number of packets occupying the relay buffer $\mathbb{E}\{L^*\}$ is given by

$$\mathbb{E}\left\{L^*\right\} = \sum_{i=0}^{B-1} i \cdot \pi^*_{L_i} = \frac{\sum_{i=0}^{B-1} i\mathrm{C}_i \cdot \rho_s^i}{\sum_{i=0}^{B-1} \mathrm{C}_i \cdot \rho_s^i} = \Psi_{n,B,\lambda}. \tag{24}$$

Since these buffered packets are destined to each of the $n-2$ destinations with equal probability, then we have

$$\overline{L}^*_{relay} = \frac{\mathbb{E}\{L^*\}}{n-2}. \tag{25}$$

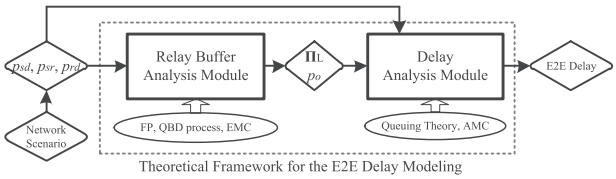


Fig. 6. Illustration of the application of our theoretical framework.

Substituting the results of (22), (24) and (25) into (19), the average transition times from the transient state S to the absorbing state D is determined as

$$\overline{X}_S = \frac{1 + (n - 2 + \Psi_{n,B,\lambda})(1 - p_o)}{\mu_s}. \tag{26}$$

Notice that $\mathbb{E}\{D\} = \overline{X}_S$, the result (16) follows, and then the result (17) follows from $\mathbb{E}\{T\} = \mathbb{E}\{Q\} + \mathbb{E}\{D\}$. ■

Remark 2: Similar to the two-hop scenario, in the multi-hop MANETs the packet delivery process and occupancy behaviors of a relay buffer can be also modeled as an AMC and a QBD process respectively, so it is expected that our proposed theoretical framework for E2E delay modeling of two-hop MANETs can be also helpful for that of the multi-hop scenarios. It is notable, however, the state transition matrix of the QBD process will be different under the two scenarios, and also that with the multi-hop network scenarios there will be multiple transient states in the AMC.

Based on Theorem 1, we can further extend our delay results to the buffer-unlimited scenario (i.e., $B \to \infty$), which is shown in the following corollary.

Corollary 1: Considering the relay buffer size tends to infinity ($B \to \infty$), then $\mathbb{E}\{D\}$ and $\mathbb{E}\{T\}$ are determined as

$$\mathop{\mathbb{E}\{D\}}_{B\to\infty} = \frac{1}{p_{sd} + p_{sr}} + \frac{n - 2}{p_{sd} + p_{sr} - \lambda}, \tag{27}$$

$$\mathop{\mathbb{E}\{T\}}_{B\to\infty} = \frac{n - 1 - \lambda}{p_{sd} + p_{sr} - \lambda}. \tag{28}$$

Proof: See Appendix C for the proof. ■

Remark 3: Notice that when $B \to \infty$, the results of Lemma 2 and Corollary 1 are coincident with the capacity and delay derived in [8], where the relay buffer size is assumed to be infinite.

V. Case Studies

In this section, we conduct case studies to illustrate the application of our theoretical framework for the E2E delay modeling in buffer-limited MANETs. As illustrated in Fig. 6, for a given network scenario, the corresponding p_{sd}, p_{sr} and p_{rd} should be determined first, then with the inputs of these probabilities, by sequentially executing the relay buffer analysis module and delay analysis module, this framework finally returns the delay results. The details of the application of our framework under specific network scenarios are shown in the following subsections.

A. Cell-partitioned MANET with LS-MAC

We first consider a cell-partitioned MANET with local scheduling based MAC protocol (LS-MAC) [8]. The whole network area is partitioned into $m \times m$ non-overlapping cells of equal size. In a time slot, each cell can support only one pair of nodes for packet transmission, concurrent transmissions in different cells will not interference with each other, and nodes within different cells cannot communicate. At the beginning of each time slot, all nodes in a cell contends for the wireless channel access using a DCF-style mechanism [35]. In addition to these network settings, the MANET also meets the set of assumptions described in Section II-A.

With the detailed information of network settings, we then determine the corresponding probabilities p_{sd}, p_{sr} and p_{rd}, provided in the following lemma.

Lemma 3: For the concerned cell-partitioned MANET with LS-MAC, the probabilities p_{sd}, p_{sr} and p_{rd} are given by

$$p_{sd} = \frac{m^2}{n} - \frac{m^2 - 1}{n - 1} + \left(\frac{m^2 - 1}{n - 1} - \frac{m^2 - 1}{n}\right)\left(1 - \frac{1}{m^2}\right)^{n-1}, \tag{29}$$

$$p_{sr} = p_{rd} = \frac{1}{2}\left\{\frac{m^2 - 1}{n - 1} - \frac{m^2}{n - 1}\left(1 - \frac{1}{m^2}\right)^{n} - \left(1 - \frac{1}{m^2}\right)^{n-1}\right\}. \tag{30}$$

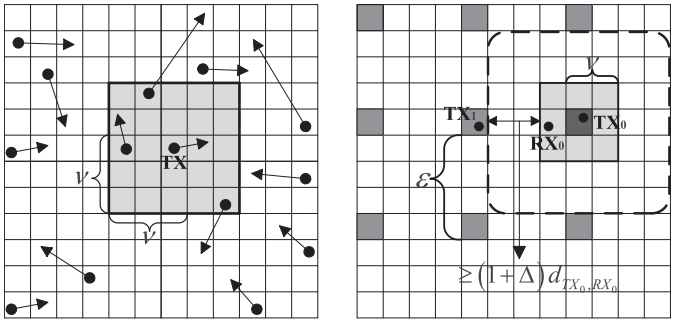
Proof: See Appendix D for the proof. ■

Given the number of nodes n and relay buffer size B, substituting formulas (29) and (30) into formula (12), we first determine the throughput capacity μ of such a MANET. Then with any packet generating rate $\lambda < \mu$, we substitute formulas (29) and (30) into equation (9) to determine the corresponding ROP p_o, and μ_s can be further determined by formula (1). Substituting λ, p_o and μ_s into formulas (15), (16) and (17), we finally obtain the results of queuing delay, delivery delay and E2E delay respectively for the concerned buffer-limited MANET.

B. Cell-partitioned MANET with Power Control and EC-MAC

We then consider a more general cell-partitioned MANET which applies the power control and Equivalent-Class based MAC protocol (EC-MAC) [15], [18], [22]. As shown in Fig. 7(a), the transmission range of a transmitter TX covers a set of cells which have a horizontal and vertical distance of no more than $\nu - 1$ cells away from its own cell. Meanwhile, as illustrated in Fig. 7(b), all cells are divided into different ECs, where any two cells in the same EC have a horizontal and vertical distance of some multiple of ε cells. Thus, the MANET contains in total ε^2 ECs and ECs are activated alternatively as time evolves. Suppose that at time slot t, a transmitter TX_0 in an active cell will transmit a packet to its receiver RX_0, in order to ensure the transmission successful, according to the Protocol Model [36] it should satisfy that

$$d_{TX_1,RX_0} \geq (1 + \Delta) d_{TX_0,RX_0}, \tag{31}$$



(a) Transmission range of a node. (b) Illustration of EC-MAC scheduling.

Fig. 7. A cell-partitioned MANET with power control and EC-MAC (a) Transmission range of a node. (b) Illustration of EC-MAC scheduling.

where TX_1 denotes a concurrent transmitter in any one of the other active cells, $d_{i,j}$ denotes the distance between nodes i and j, and Δ is a guard factor. Thus we have

$$\varepsilon - \nu \geq (1+\Delta)\sqrt{2}\nu, \tag{32}$$

and ε should be set as

$$\varepsilon = \min\{\lceil (1+\Delta)\sqrt{2}\nu + \nu \rceil, m\}. \tag{33}$$

Regarding the corresponding probabilities p_{sd}, p_{sr} and p_{rd} of this type of MANETs, we have the following lemma.

Lemma 4: For the concerned MANET with power control and EC-MAC, the probabilities p_{sd}, p_{sr} and p_{rd} are given by

$$p_{sd} = \frac{1}{\varepsilon^2}\left\{\frac{\Gamma - \frac{m^2}{n}}{n-1} + \frac{m^2 - 1 - (\Gamma-1)n}{n(n-1)}\left(1 - \frac{1}{m^2}\right)^{n-1}\right\}, \tag{34}$$

$$p_{sr} = p_{rd} = \frac{1}{2\varepsilon^2}\left\{\frac{m^2 - \Gamma}{n-1}\left(1 - \left(1 - \frac{1}{m^2}\right)^{n-1}\right) - \left(1 - \frac{\Gamma}{m^2}\right)^{n-1}\right\}, \tag{35}$$

where $\Gamma = (2\nu - 1)^2$.

Proof: See Appendix D for the proof. ■

By applying the same operations of our theoretical framework as the previous subsection, we can obtain the delay results for the concerned MANETs.

C. Other Network Scenarios

Notice that to apply our theoretical framework for the E2E delay modeling, it only needs to determine the inputs of the framework, i.e., the probabilities p_{sd}, p_{sr} and p_{rd}. Thus, this framework also has the great potential to be applied to many other network scenarios. For example, for the MANETs where the 2HR routing scheme is administered by cell [8] (not by each node in our case), and the ratio between **S-R** and **R-D** transmission can be changed [15] (we fix the ratio as 0.5), these probabilities can be determined. Recently, Chen *et al.* [19] has reported that how to compute these probabilities for a MANET under the continuous network model and the Aloha MAC protocol, then with these probabilities as the inputs of our framework, the corresponding delay analysis under the practical buffer constraint can be conducted.

VI. Simulation Results

In this section, we first conduct simulations to validate our theoretical framework for the E2E delay modeling in buffer-limited MANETs, then provide discussions about the impacts of network parameters on delay performance.

A. Simulation Settings

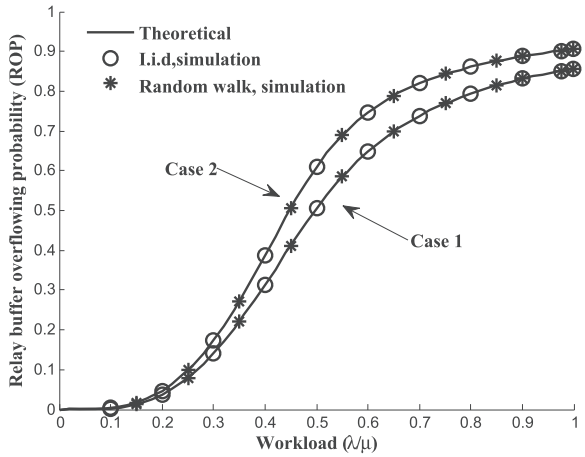
For the validation of our theoretical framework and delay results, a specific C++ simulator was developed to simulate the packet generating, queuing and delivery processes in a cell-partitioned MANET [37], where the network settings, including relay buffer size B, number of nodes n, partition parameter m, packet generating rate λ and the mobility model can be flexibly adjusted to simulate the network performance under various scenarios. For the network scenario with power control and EC-MAC, we set $\nu = 1$ and $\Delta = 1$ [38]. The duration of each task of simulation is set to be 2×10^8 time slots, and we only collect data from the last 80% of the time slots in each task (the system will be in the steady state with high probability), to ensure the accuracy of simulated results.

B. Validation

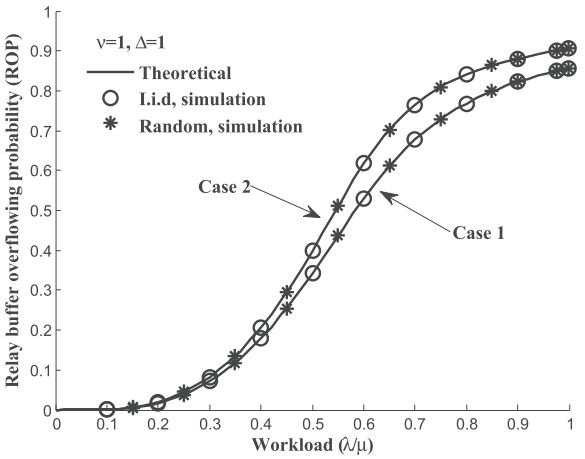
First, we provide plots of the theoretical and simulated ROP performance under two network scenarios in Fig. 8, and for each scenario we consider two cases (case 1: $n = 32, m = 4, B = 5$, and case 2: $n = 50, m = 5, B = 5$) and two mobility models (the i.i.d mobility model and the random walk model). The workload is defined as λ/μ. We can see from Fig. 8 that the simulation results match nicely with the theoretical ones for all the cases, which indicates that our framework is highly efficient in depicting the occupancy behaviors of the relay buffer in buffer-limited MANETs.

Then, with the same network settings, we provide plots of the theoretical and simulated E2E delay results in Fig. 9. It is observed from Fig. 9 that all the simulation results can match the corresponding theoretical curves very nicely, indicating that: 1) our theoretical framework is highly efficient for the E2E delay modeling in buffer-limited MANETs; 2) the framework is very general since it can be applied to various network scenarios. Another observation of Fig. 9 is that the packet E2E delay increases sharply as the packet generating rate λ approaches a specific value (e.g., under LS-MAC and case 1, the value is around 0.038), which serves as an intuitive impression of its corresponding throughput capacity μ.

To further validate our framework on throughput capacity, Fig. 10 summarizes the simulation results on the achievable per node throughput, where two network scenarios with different throughput capacity (LS-MAC: $n = 200, m = 10, B = 5, \mu_{LS} = 6.5 \times 10^{-3}$) and (power control and EC-MAC: $n = 32, m = 4, B = 10, \mu_{EC} = 3.3 \times 10^{-3}$) are presented.

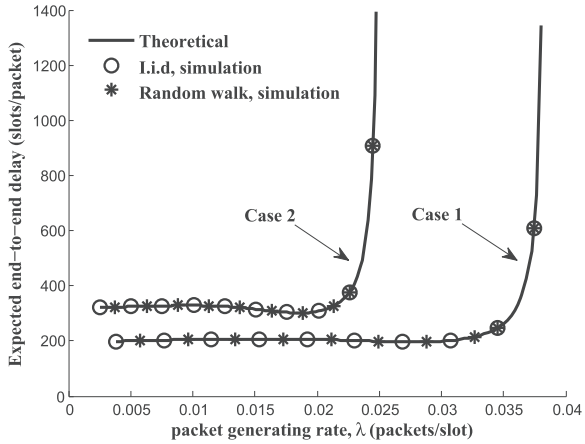


(a) ROP under LS-MAC.

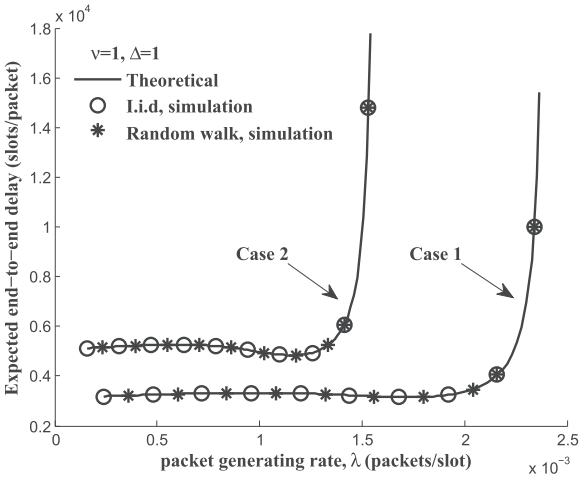


(b) ROP under power control and EC-MAC.

Fig. 8. Theoretical and simulated ROP performance. Case 1: $n = 32, m = 4, B = 5$. Case 2: $n = 50, m = 5, B = 5$. (a) ROP under LS-MAC. (b) ROP under power control and EC-MAC.



(a) End-to-end delay under LS-MAC.



(b) End-to-end delay under power control and EC-MAC.

Fig. 9. Theoretical and simulated end-to-end delay performance. Case 1: $n = 32, m = 4, B = 5$. Case 2: $n = 50, m = 5, B = 5$. (a) End-to-end delay under LS-MAC. (b) End-to-end delay under power control and EC-MAC.

We can see that for each network scenario there, the per node throughput first monotonously increases before the workload reaches 1, and then remains a constant which is just the corresponding throughput capacity when the workload exceeds 1. Thus, it indicates that the proposed framework is also efficient in depicting the per node throughput capacity behavior of the buffer-limited MANET.

C. Performance Discussion

With the help of our theoretical framework for the E2E delay modeling, we explore how the network parameters affect the the delay performance of a buffer-limited MANET. Without loss of generality, we consider here a cell-partitioned MANET with LS-MAC.

We first summarize in Fig. 11 that how the expected delivery delay $\mathbb{E}\{D\}$ varies with the workload. A very interesting observation is that under the buffer-limited scenarios ($B = 5$ and $B = 20$), as workload increases, $\mathbb{E}\{D\}$ first increases to a maximum and then decreases. This is due to the reason that the effects of workload on $\mathbb{E}\{D\}$ are two folds. On one hand, a larger workload will lead to a longer relay queue length, which further leads to a higher delay in the relay queue; on the other hand, a larger workload will lead to a higher ROP, which further

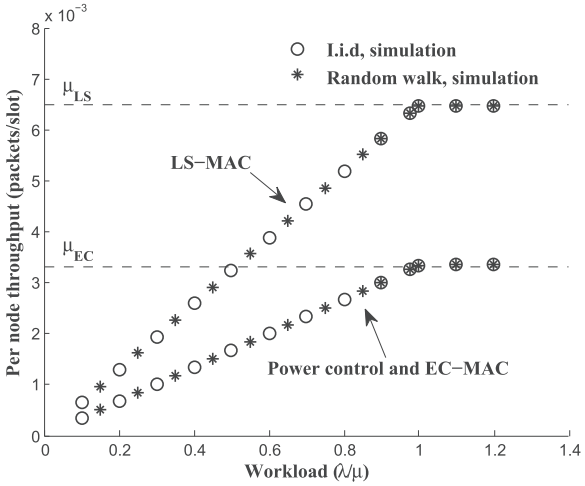


Fig. 10. Per node achievable throughput. LS-MAC: $n = 200, m = 10, B = 5$. Power control and EC-MAC: $n = 32, m = 4, B = 10$.

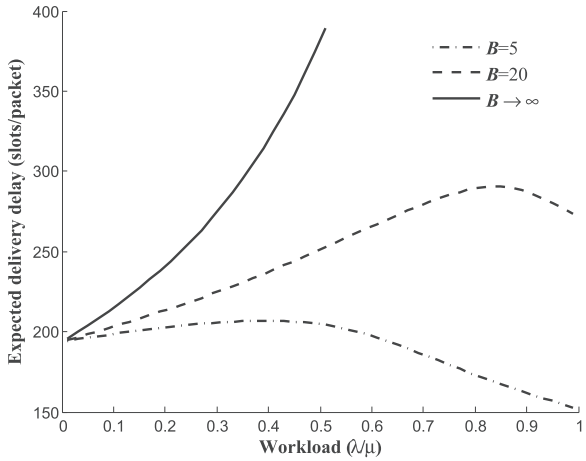


Fig. 11. Delivery delay vs. workload (λ/μ) under different settings of relay buffer size. $n = 32, m = 4$.

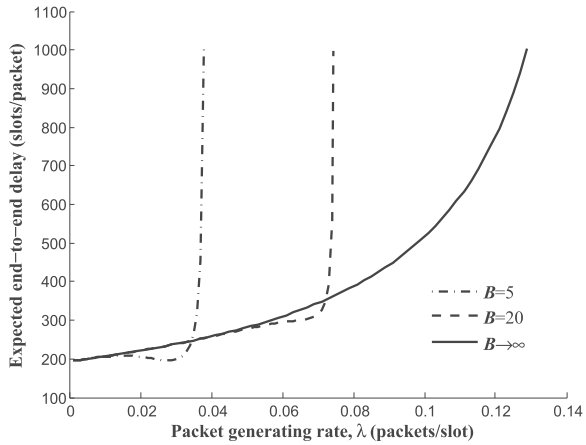


Fig. 12. End-to-end delay vs. packet generating rate λ under different settings of relay buffer size. $n = 32, m = 4$.

leads to a lower probability that a packet to be delivered by a two-hop way, such that $\mathbb{E}\{D\}$ decreases. Since the latter effect, the delivery delay under a small relay buffer is lower than that under a large one.

Fig. 12 shows the relationship between the expected E2E delay $\mathbb{E}\{T\}$ and packet generating rate λ. We can see that under buffer-limited scenarios, as λ increases, $\mathbb{E}\{T\}$ doesn't increase all the time because the delivery delay will decrease when λ exceeds a specific value, however when λ approaches the corresponding throughput capacity, $\mathbb{E}\{T\}$ increases sharply because the queuing delay tends to infinity. It also can be seen that when λ is small, $\mathbb{E}\{T\}$ under $B = 5$ is smaller than that under $B = 20$, since both of the queuing delay under two settings are small, but a small relay buffer can lead to a small delivery delay. However, with λ getting larger and larger, $\mathbb{E}\{T\}$ under $B = 5$ finally exceeds that under $B = 20$, and tends to infinity earlier. It indicates that increasing the relay buffer size can ensure the E2E delay limited for a larger region of packet generating rate.

We illustrate in Fig. 13 how $\mathbb{E}\{T\}$ varies B under the settings of $(n = 32, m = 4, \lambda = \{0.01, 0.02\})$. According to formula (12), $\mu = 0.0227$ when $B = 1$, and μ increases as B increases. Thus, for $\lambda = 0.01$ which is much smaller than 0.0227, $\mathbb{E}\{T\}$

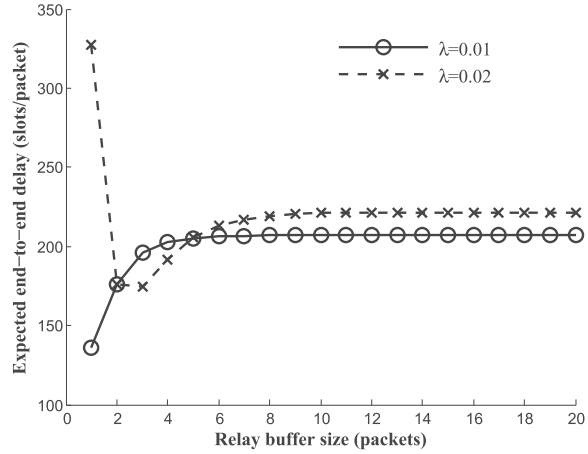


Fig. 13. End-to-end delay vs. relay buffer size. $n = 32, m = 4$.

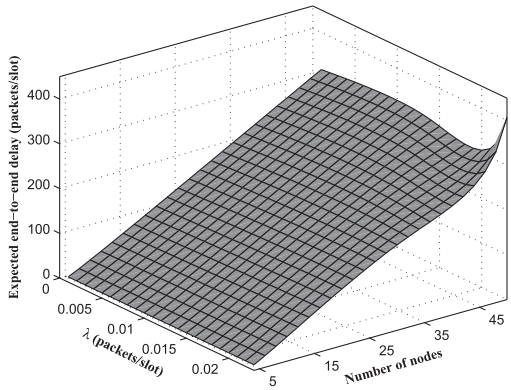


Fig. 14. End-to-end delay vs. packet generating rate λ and number of nodes n. $B = 5$.

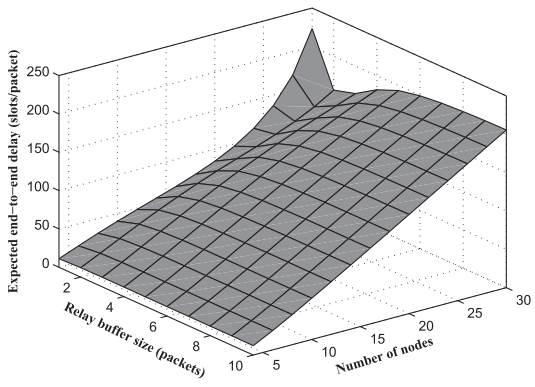


Fig. 15. End-to-end delay vs. relay buffer size B and number of nodes n. $\lambda = 0.02$.

increases as B increases and finally tends to a constant 206.92 which can be determined by formula (28). While for $\lambda = 0.02$ which is very close to the μ under $B = 1$, $\mathbb{E}\{T\}$ under $B = 1$ is very large. With B increasing, the corresponding μ increases, leading to the $\mathbb{E}\{T\}$ first decreases, then increases and finally tends to a constant 221.65.

We further illustrate in two 3D figures (Fig. 14 and Fig. 15) that how $\mathbb{E}\{T\}$ is influenced by $\{n, \lambda\}$ and $\{n, B\}$, respectively (the ratio of n to the number of cells keeps as 2). We can see

that the variations of $\mathbb{E}\{T\}$ with n are complicated, but in general $\mathbb{E}\{T\}$ increases as n increases. A more careful observation is that when n increases, $\mathbb{E}\{T\}$ first increases almost linearly when λ is much smaller than μ, then increases quickly when λ approaches μ. For example, these behaviors can be found in Fig. 14 under $\lambda = 0.23$ and in Fig. 15 under $B = 1$.

VII. Conclusion

This paper represents a significant step towards the exact end-to-end delay modeling of practical buffer-limited MANETs. With the help of the theories of Fixed-Point, QBD process, EMC and AMC, a novel theoretical framework has been developed to efficiently depict the highly dynamics in such networks. This framework is general in the sense that it can be applied to the MANETs with any mobility model that leads to the uniform distribution of the locations of nodes, and any MAC protocol as long as the probabilities p_{sd}, p_{sr} and p_{rd} there can be determined. Also, it is expected the framework can shed light on the E2E delay modeling for multi-hop MANETs. Extensive simulations have been conducted to validate the efficiency and applicability of our framework, and some interesting theoretical findings about the impacts of network parameters on delay performance have been discussed.

Notice that our E2E delay modeling for buffer-limited MANETs is based on the 2HR routing and relay buffer constraint, so one of the future research directions is to extend our study to the E2E delay modeling for MANETs with multi-hop routing schemes and more practical buffer constraint on both source buffer and relay buffer. Another appealing direction is to explore the delay modeling for the concerned MANETs with the consideration of more practical network settings (such as the wireless channel fading) and apply the well-known NS-2 network simulator for model validation.

Appendix A
Proof of Lemma 1

We denote by λ_R the packet arrival rate at the relay buffer in **S**. Due to the ergodic and uniform properties of node mobility, for node **S**, each of the remaining $n-2$ nodes (except **D**) is likely to serve as its relay with equal probability. Due to the symmetry of nodes and traffic flows, for **S** serving as a relay, all the other $n-2$ nodes are likely to forward packets to its relay buffer. Notice that the packet departure process of source queue is a Bernoulli process with rate λ, and from (1) we can see the ratio of **S-R** transmission to the whole packet departure is $\frac{p_{sr}(1-p_o(\lambda))}{\mu_s(\lambda)}$, then we have

$$\begin{aligned}\lambda_R &= (n-2)\lambda \cdot \frac{p_{sr}\,(1-p_o(\lambda))}{\mu_s(\lambda)}/(n-2)\\ &= \rho_s(\lambda)p_{sr}\,(1-p_o(\lambda))\,. \end{aligned} \tag{36}$$

Regarding the transition probability from state (i, j) to the states in its adjacent upper level L_{i+1}, we have the following equation

$$p_{(i,j),L_{i+1}} \cdot (1-p_o(\lambda)) + 0 \cdot p_o(\lambda) = \lambda_R. \tag{37}$$

Combining (37) with (36) we have

$$p_{(i,j),L_{i+1}} = \rho_s(\lambda)p_{sr}. \tag{38}$$

Substituting (38) into (3) we have

$$p_L^{(i,i+1)} = \rho_s(\lambda)p_{sr}.$$

Regarding the transition probability from state (i, j) to the states in its adjacent lower level L_{i-1}, when **S** conducts a **R-D** transmission with probability p_{rd}, due to the ergodic and uniform features of node mobility, it will choose one of the other $n-2$ nodes as its receiver with equal probability. Then we have

$$p_{(i,j),L_{i-1}} = p_{rd} \cdot \frac{j}{n-2}. \tag{39}$$

To determine the conditional probability $P_{j|L_i}$, we utilize the following Occupancy technique [39]. Considering the relay buffer on level i, where each of these i buffered packets may be destined for any one of the other $n-2$ nodes (except **S** and **D**), the number of all possible cases N_{L_i} is determined as

$$N_{L_i} = \binom{n-3+i}{i}. \tag{40}$$

Considering the condition that these i packets are destined for only j different nodes, then the number of possible cases $N_{(i,j)}$ is determined as

$$N_{(i,j)} = \binom{n-2}{j} \cdot \binom{(j-1)+(i-j)}{i-j}. \tag{41}$$

Due to the ergodic and uniform features of node mobility, each of these cases occurs with equal probability. According to the *Classical Probability*, $P_{j|L_i}$ is then determined as

$$P_{j|L_i} = \frac{N_{(i,j)}}{N_{L_i}} = \frac{\binom{n-2}{j} \cdot \binom{i-1}{j-1}}{\binom{n-3+i}{i}}. \tag{42}$$

It can be easily verified that $\sum_{j\le i} P_{j|L_i} = 1$. Combining (39), (42) and (3) we have

$$\begin{aligned} p_L^{(i,i-1)} &= \sum_{j=1}^{i} \left\{ \frac{\binom{n-2}{j} \cdot \binom{i-1}{j-1}}{\binom{n-3+i}{i}} \cdot\ p_{rd}\frac{j}{n-2} \right\}\\ &= \frac{p_{rd}}{\binom{n-3+i}{i}} \cdot \sum_{j=0}^{i-1} \left\{ \binom{n-3}{j} \cdot \binom{i-1}{j} \right\}\\ &= p_{rd} \cdot \frac{\binom{n-4+i}{i-1}}{\binom{n-3+i}{i}} = \frac{i}{n-3+i} \cdot p_{rd}. \end{aligned} \tag{43}$$

APPENDIX B
PROOF OF LEMMA 2

Since the relay buffer is limited, it is always stable. Thus, we only need to consider the stability of source queue. The source queue is a Bernoulli/Bernoulli queue with packet generating rate λ and a corresponding service rate $\mu_s(\lambda)$. Based on the queuing theory, the source queue is stable (resp. unstable) when $\lambda < \mu_s(\lambda)$ (resp. $\lambda \geq \mu_s(\lambda)$).

As λ increases, $p_o(\lambda)$ increases since the network is more congested, and $\mu_s(\lambda)$ decreases according to formula (1). Thus, from the definition of throughput capacity, μ should satisfy that

$$\mu = \lambda^* = \mu_s(\lambda^*). \tag{44}$$

It is notable that when λ approaches λ^*, $\lim_{\lambda \to \lambda^*} \rho_s(\lambda) = 1$, and from (9) we have

$$p_o(\lambda^*) = \frac{\mathrm{C}_B}{\sum_{k=0}^{B} \mathrm{C}_i} = \frac{n-2}{n-2+B}. \tag{45}$$

Then μ is determined by substituting (45) into (1).

APPENDIX C
PROOF OF COROLLARY 1

We denote by $F(\rho_s)$, $G(\rho_s)$ the sums of infinite series $\sum_{i\geq 0} \mathrm{C}_i \rho_s^i$ and $\sum_{i\geq 0} i\mathrm{C}_i \rho_s^i$, respectively. Notice that $F(\rho_s)$ is the Taylor series expansion from $(1-\rho_s)^{2-n}$, then we have

$$F(\rho_s) = \frac{1}{(1-\rho_s)^{n-2}}, \tag{46}$$

$$G(\rho_s) = \rho_s \cdot F'(\rho_s) = (n-2)\frac{\rho_s}{(1-\rho_s)^{n-1}}. \tag{47}$$

Further we have

$$\Psi_{n,\infty,\lambda} = \frac{G(\rho_s)}{F(\rho_s)} = (n-2)\frac{\rho_s}{1-\rho_s}, \tag{48}$$

and

$$\begin{aligned} \underset{B\to\infty}{p_o} &= \lim_{B\to\infty} \mathrm{C}_B \cdot \rho_s^B \cdot (1-\rho_s)^{n-2} && (49)\\ &\leq \lim_{B\to\infty} (B+n)^n \rho_s^B \leq \lim_{B\to\infty} 2^n B^n \rho_s^B && (50)\\ &= \lim_{B\to\infty} \frac{n!\rho_s^B}{(-\ln \rho_s)^n} = 0, && (51) \end{aligned}$$

where (51) is obtained by utilizing the L'Hôpital's rule recursively.

Substituting (48) and (51) into Theorem 1, we can obtain (27) and (28) directly.

APPENDIX D
PROOFS OF LEMMA 3 AND LEMMA 4

For a cell-partitioned MANET with LS-MAC, the event that node **S** conducts a **S-D** (resp. **S-R** or **R-D**) transmission in a time slot can be divided into the following sub-events: (1) **D** is (resp. is not) in the same cell with **S**; (2) other k out of $n-2$ nodes are in the same cell with **S**, while the remaining $n-2-k$ nodes are not in this cell; (3) **S** contends for the wireless channel access successfully. Thus we have

$$\begin{aligned} p_{sd} &= \sum_{k=0}^{n-2} \binom{n-2}{k} \left(\frac{1}{m^2}\right)^{k+1} \left(1-\frac{1}{m^2}\right)^{n-2-k} \cdot \frac{1}{k+2}\\ &= \sum_{k=0}^{n-2} \binom{n-1}{k+1} \left(\frac{1}{m^2}\right)^{k+1} \left(1-\frac{1}{m^2}\right)^{n-2-k} \cdot \frac{1}{k+2}\\ &\quad - \sum_{k=0}^{n-3} \binom{n-2}{k+1} \left(\frac{1}{m^2}\right)^{k+1} \left(1-\frac{1}{m^2}\right)^{n-2-k} \cdot \frac{1}{k+2}\\ &= \frac{m^2}{n}\left\{1-\left(1-\frac{1}{m^2}\right)^n\right\} - \left(1-\frac{1}{m^2}\right)^{n-1}\\ &\quad - \frac{m^2-1}{n-1}\left\{1-\left(1-\frac{1}{m^2}\right)^{n-1}\right\} + \left(1-\frac{1}{m^2}\right)^{n-1}\\ &= \frac{m^2}{n} - \frac{m^2-1}{n-1} + \left(\frac{m^2-1}{n-1} - \frac{m^2-1}{n}\right)\left(1-\frac{1}{m^2}\right)^{n-1}, \end{aligned}$$

and

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

For a cell-partitioned MANET with power control and EC-MAC, by applying the similar approach and algebraic operations we have

$$\begin{aligned} p_{sd} &= \frac{1}{\varepsilon^2}\left\{\sum_{k=0}^{n-2} \binom{n-2}{k} \left(\frac{1}{m^2}\right)^{k+1} \left(1-\frac{1}{m^2}\right)^{n-2-k} \cdot \frac{1}{k+2}\right.\\ &\quad \left. + \sum_{k=0}^{n-2} \binom{n-2}{k} \left(\frac{1}{m^2}\right)^{k+1} \left(1-\frac{1}{m^2}\right)^{n-2-k} \cdot \frac{4v^2-4v}{k+1}\right\}\\ &= \frac{1}{\varepsilon^2}\left\{\frac{\Gamma-\frac{m^2}{n}}{n-1} + \frac{m^2-1-(\Gamma-1)n}{n(n-1)}\left(1-\frac{1}{m^2}\right)^{n-1}\right\}, \end{aligned}$$

and

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

REFERENCES

[1] A. S. Tanenbaum, *Computer Networks*. Englewood Cliffs, NJ, USA: Prentice-Hall, 2003.

[2] C. E. Perkins, *Ad Hoc Networking*. Reading, MA, USA: Addison-Wesley, 2000.

[3] R. Ramanathan and J. Redi, "A brief overview of ad hoc networks: Challenges and directions," *IEEE Commun. Mag.*, vol. 40, no. 5, pp. 20–22, May 2002.

[4] J. Andrews *et al.*, "Rethinking information theory for mobile ad hoc networks," *IEEE Commun. Mag.*, vol. 46, no. 12, pp. 94–101, Dec. 2008.

[5] A. Goldsmith, M. Effros, R. Koetter, M. Medard, 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, May 2011.

[6] L. Hanzo and R. Tafazolli, "A survey of qos routing solutions for mobile ad hoc networks," *IEEE Commun. Surveys Tuts.*, vol. 9, no. 2, pp. 50–70, Second Quarter 2007.

[7] L. Chen and W. B. Heinzelman, "A survey of routing protocols that support QOS in mobile ad hoc networks," *IEEE Netw.*, vol. 21, no. 6, pp. 30–38, Nov./Dec. 2007.

[8] M. J. Neely and E. Modiano, "Capacity and delay tradeoffs for ad-hoc mobile networks," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 1917–1936, Jun. 2005.

[9] A. E. Gamal, J. Mammen, B. Prabhakar, and D. Shah, "Optimal throughput-delay scaling in wireless networks–Part I: The fluid model," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2568–2592, Jun. 2006.

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

[11] 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, Oct. 2011.

[12] W. Huang and X. Wang, "Throughput and delay scaling of general cognitive networks," in *Proc. IEEE INFOCOM*, 2011, pp. 2210–2218.

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

[14] M. Alresaini, M. Sathiamoorthy, B. Krishnamachari, and M. J. Neely, "Backpressure with adaptive redundancy (BWAR)," in *Proc. IEEE INFOCOM*, 2012, pp. 2300–2308.

[15] J. Gao, J. Liu, 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.

[16] A. Jindal and K. Psounis, "Contention-aware performance analysis of mobility-assisted routing," *IEEE Trans. Mobile Comput.*, vol. 8, no. 2, pp. 145–161, Feb. 2009.

[17] J. Liu, J. Gao, X. Jiang, H. Nishiyama, and N. Kato, "Capacity and delay of probing-based two-hop relay in manets," *IEEE Trans. Wireless Commun.*, vol. 11, no. 11, pp. 4172–4183, Nov. 2012.

[18] J. Liu, X. Jiang, H. Nishiyama, N. Kato, and X. Shen, "End-to-end delay in mobile ad hoc networks with generalized transmission range and limited packet redundancy," in *Proc. IEEE Wireless Commun. Netw. Conf.*, 2012, pp. 1731–1736.

[19] Y. Chen, Y. Shen, X. Jiang, and J. Li, "Throughput capacity of aloha manets," in *Proc. IEEE Int. Conf. Commun. China Workshops*, 2013, pp. 71–75.

[20] J. Gao and X. Jiang, "Delay modeling for broadcast-based two-hop relay manets," in *Proc. IEEE 11th Int. Symp. Model. Optim. Mobile Ad Hoc Wireless Netw.*, 2013, pp. 357–363.

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

[22] S. R. Kulkarni and P. Viswanath, "A deterministic approach to throughput scaling in wireless networks," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1041–1049, Jun. 2004.

[23] M. Grossglauser and D. Tse, "Mobility increases the capacity of ad hoc wireless networks," *IEEE/ACM Trans. Netw.*, vol. 10, no. 4, pp. 477–486, Aug. 2002.

[24] D. Ciullo, V. Martina, M. Garetto, and E. Leonardi, "Impact of correlated mobility on delay-throughput performance in mobile ad hoc networks," *IEEE/ACM Trans. Netw.*, vol. 19, no. 6, pp. 1745–1758, Dec. 2011.

[25] L. B. Le, E. Modiano, and N. B. Shroff, "Optimal control of wireless networks with finite buffers," *IEEE/ACM Trans. Netw.*, vol. 20, no. 4, pp. 1316–1329, Aug. 2012.

[26] J. D. Herdtner and E. K. Chong, "Throughput-storage tradeoff in ad hoc networks," in *Proc. IEEE INFOCOM*, 2005, pp. 2536–2542.

[27] H. Daduna, *Queueing Networks with Discrete Time Scale: Explicit Expressions for the Steady State Behavior of Discrete Time Stochastic Networks*. Berlin, Germany: Springer-Verlag, 2001.

[28] A. S. Alfa, *Queueing Theory for Telecommunications: Discrete Time Modeling of a Single Node System*. New York, NY, USA: Springer, 2010.

[29] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*. Philadelphia, PA, USA: ASA-SIAM Series on Statistics and Applied Probability, 1999.

[30] J. Hachigian, "Collapsed Markov chains and the Chapman-Kolmogorov equation," *Ann. Math. Statist.*, vol. 34, pp. 233–237, 1963.

[31] R. Subramanian and F. Fekri, "Analysis of multiple-unicast throughput in finite-buffer delay-tolerant networks," in *Proc. IEEE Int. Symp. Inf. Theory*, 2009, pp. 1634–1638.

[32] A. Granas and J. Dugundji, *Fixed Point Theory*. Berlin, Germany: Springer-Verlag, 2003.

[33] O. Häggström, *Finite Markov Chains and Algorithmic Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[34] T. Robertazzi, Computer Networks and Systems: Queueing Theory and Performance Evaluation. New York, NY, USA: Springer, 2012.

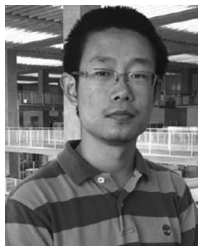
[35] G. Bianchi, "Performance analysis of the ieee 802.11 distributed coordination function," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 3, pp. 535–547, Mar. 2000.

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

[37] J. Liu and Y. Xu, (2015). *C++ Simulator for Delay Performance in Buffer-Limited Manets* [Online]. Available: http:/jliuyxu.blogspot.jp/

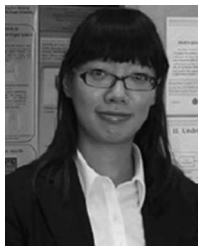
[38] S. McCanne and S. Floyd. (1997). *The Network Simulator ns-2* [Online]. Available: http://www.isi.edu/nsnam/ns/

[39] H. Stark and J. W. Woods, *Probability and Random Processes With Applications to Signal Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 2002.



Jia Liu received the B.E. degree in communications engineering from Xidian University, Xi'an, China, in 2010. From 2010 to 2012, he did research study on wireless communication as a Doctoral Student in communication and information systems, Xidian University, Xi'an, China. Since 2012, he is pursuing the Ph.D. degree at the Future University Hakodate, Hakodate, Japan. His research interests include mobile ad hoc networks, 5G communication systems, and D2D communications.

Min Sheng (M'03) received the M.S. and Ph.D. degrees in communication and information systems from Xidian University, Xi'an, China, in 1997 and 2000, respectively. She is currently a Full Professor with the School of Telecommunication Engineering, Xidian University. She has authored two books and over 50 papers in refereed journals and conference proceedings. Her research interests include mobile ad hoc networks, 5G communication systems, cross-layer algorithm design, and cooperative communications. She was the New Century Excellent Talents in University by the Ministry of Education of China, and obtained the Young Teachers Award by the Fok Ying-Tong Education Foundation, China, in 2008.



Yang Xu received the B.E. degree in communications engineering and the Ph.D. degree in communication and information systems from Xidian University, Xi'an, China, in 2006 and 2014, respectively. Currently, she is a Postdoctor with the School of Computer Science and Technology, Xidian University, and a Visiting Scholar at the Future University Hakodate, Hakodate, Japan. She has authored about ten papers at premium international journals and conferences. Her research interests include routing protocol design, network performance analysis, and physical-layer security.

Jiandong Li (SM'05) received the B.E., M.S., and Ph.D. degrees in communications engineering from Xidian University, Xi'an, China, in 1982, 1985, and 1991, respectively. Since 1985, he has been a Faculty Member of the School of Telecommunications Engineering, Xidian University, Xi'an, China, where he is currently a Professor and Vice Director of the Academic Committee of the State Key Laboratory of Integrated Service Networks. From 2002 to 2003, he was a Visiting Professor at the Department of Electrical and Computer Engineering, Cornell University, Ithaca, NY, USA. His research interests include wireless communication theory, cognitive radio, and signal processing. He served as the General Vice Chair for ChinaCom in 2009 and TPC Chair of IEEE ICCC in 2013. He was recognized as a Distinguished Young Researcher by NSFC and a Changjiang Scholar by the Ministry of Education, China.

Xiaohong Jiang (M'03–SM'09) received the B.S., M.S., and Ph.D. degrees from Xidian University, Xi'an, China, in 1989, 1992, and 1999, respectively. He is currently a Full Professor with the Future University Hakodate, Hakodate, Japan. Before joining Future University, he was an Associate Professor, Tohoku University, Sendai, Japan, from February 2005 to March 2010. He has authored over 260 technical papers at premium international journals and conferences, which include over 50 papers published in top IEEE journals and top IEEE conferences, such as the IEEE/ACM TRANSACTIONS ON NETWORKING, the IEEE JOURNAL OF SELECTED AREAS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, and IEEE INFOCOM. His research interests include computer communications networks, mainly wireless networks and optical networks, network security, routers/switches design, etc. He was the recipient of the Best Paper Award of the IEEE HPCC 2014, the IEEE WCNC 2012, the IEEE WCNC 2008, the IEEE ICC 2005-Optical Networking Symposium, and the IEEE/IEICE HPSR 2002. He is a Member of ACM and IEICE.