On Achievable Delay/Capacity Trade-offs in Mobile Ad Hoc Networks

G. Sharma, R. R. Mazumdar

School of Electrical & Computer Engineering,
Purdue University, West-Lafayette, IN 47907-1285
gsharma@purdue.edu, mazum@purdue.edu

Abstract. Recent work of Gupta and Kumar (2000) has shown that in a multi-hop wireless network the throughput capacity per source-destination pair goes to zero as the node density increases. While it has been shown that a constant throughput scaling per source-destination pair can be achieved in mobile ad hoc networks, the delay related aspects have not been considered in detail. In this paper, we study the delay-capacity trade-off in mobile ad hoc networks. We consider two canonical random mobility models in this paper; the Brownian mobility model (BMM) and the random way-point mobility model (RWMM). We show that under the distributed 2-hop relaying protocol proposed by Grossglauser and Tse (2001), the packet delay scales as $\Theta(T_p(n)n)$ under the RWMM and $O(T_p(n)n\log^2(n))$ under the BMM, where $T_p(n)$ is the packet transmission time. We then show that the delay scales as $\Omega(T_p(n)\sqrt{n})$, under a broad class of scheduling and relaying protocols. Further, we show that the trade-off: *delay/capacity* $\geq \Theta(T_p(n)n)$, is necessary as well as sufficient under our settings. We then propose two distributed protocols which achieve the above mentioned lower bound on the packet delay, and evaluate their performance in terms of the delay-capacity trade-off.

1 Introduction

An ad hoc network is a collection of wireless nodes forming a temporary network without the aid of any existing infrastructure or centralized administration. In such a network, each node operates not only as a host but also as a router, forwarding packets for other nodes in the network which are not in the wireless transmission range of each other. This leads to a "multi-hopping" scenario, where each packet typically goes through a large number of nodes before reaching the destination. The throughput capacity of such networks has been studied in [1], and it is shown that the throughput capacity per node goes to zero as the number of nodes in the network increases. This is a pessimistic result as it implies that large ad hoc networks might be impractical.

It turns out that much higher throughput capacities can be achieved in mobile ad hoc networks (MANETs), where some or all of the nodes are mobile. In [2], it is shown that a throughput capacity of $\Theta(1)$ per node can be achieved with a 2-hop relaying algorithm. However, no bounds on the delay are provided in [2].

Clearly, the usefulness of the throughput result is limited, since both quantities, the delay and the throughput capacity, are important. A little amount of reflection also shows that there is a trade-off between the two quantities in that one can only be improved at the expense of the other. As different applications typically have different delay and throughput requirements, it will be useful to characterize this trade-off.

In [3], the authors determine the delay limited capacity of ad hoc networks using the *diversity coding* approach given in [4]. In [5], the authors consider a MANET with stationary nodes and mobile relays, and show that the delay can be improved by exploiting the velocity information and selectively relaying packets to those nodes which are moving in the direction of the destination.

Recently, there has been a lot of interest in characterizing the delay-capacity trade-off in MANETs. Various network and mobility models have been considered for this purpose. In [6], the authors consider a *cell partitioned* network with an *infinite mobility* model. Similar mobility models have been considered in [7], [8]. However, the *infinite mobility* model considered in these papers is not very realistic and fails to provide some real insights into the nature of the delay-capacity trade-off. In [9], the authors consider a more realistic random walk mobility model with a cellular TDMA scheme.

In this paper, we consider two canonical random mobility models which are used very often in the literature on MANETs. We first estimate the packet delay under the distributed version of the 2-hop relaying protocol proposed in [2]. We then consider two alternative protocols for delay improvement. The idea behind these protocols is to reduce the packet delay by trading off the throughput capacity. All protocols considered in this paper are topology-transparent and require no routing information. The only requirement is that each node should be able to determine its nearest neighbor (possibly on the basis of the received power or the signal-to-interference ratio).

The paper is organized as follows. In the next section, we describe our model and define some key concepts, as well as discuss some prior results. In section 3, we estimate the delay under the distributed 2-hop relaying protocol. We consider some alternate protocols for delay improvement in section 4. In section 5, we discuss the delay-capacity trade-off. We end this paper with some concluding remarks in section 6.

2 Model, Definitions, and Prior Results

2.1 Network and Transmission Model

We consider a MANET formed by n nodes. It is assumed that the nodes are distributed uniformly on a sphere of radius R, to start with. The motion of the nodes is governed by one of the random mobility models to be described later in this section. Throughout this paper, we study the asymptotic throughput capacity and delay as n becomes large.

We consider an interference based transmission model as in [1], [2], [3], [10]. More precisely, a node i is capable of transmitting W bits/sec to node j at time t, if

$$\frac{P_i(t)\gamma_{ij}(t)}{N_o + \frac{1}{L}\sum_{k\neq i} P_k(t)\gamma_{kj}(t)} > \beta,$$

where P_i is the transmit power of node i, $\gamma_{ij}(t)$ is the channel gain from node i to node j at time t, N_o is the background noise power, L is the processing gain of the system, and β is the SINR requirement for successful communication. The channel gain is assumed to be of the form $\gamma_{ij}(t) = 1/d_{ij}(t)^\alpha$, where $d_{ij}(t)$ is the distance between the nodes i and j, at time t, and typically, $\alpha \in (2, 4]$.

2.2 Distributed 2-Hop Relaying Protocol (D2HRP)

In this protocol, the packets are either transmitted directly from the source to the destination, or via a single relay node. We assume a slotted system. In every time-slot, each node randomly decides to be a sender with probability p, or a potential receiver otherwise. Here, p is a global parameter known to all the nodes. The optimal value of p which maximizes the number of successful sender-receiver pairs per slot is 0.34 (see [16]). Each sender node transmits a packet to its nearest neighbor with a unit transmit power. These packets may or may not be received correctly depending on the interference from other simultaneous transmissions in the network.

Let N_t be the number of successful sender-receiver pairs per slot, in [2] it is shown that

$$\lim_{n\to\infty} \frac{E[N_t]}{n} = \phi > 0 \qquad (2.1)$$

The main reason why we can have $\varTheta(n)$[1] concurrent nearest neighbor transmissions is that the received power at the nearest neighbor is of the same order as the total interference power from $\varTheta(n)$ number of interferers (see [13], for a similar result).

Clearly, whether any two arbitrary nodes are nearest neighbors[2] or not depends on the position of all the nodes in the network. This kind of coupling can significantly complicate the analysis. A simplification

[1] We use the following asymptotic notation throughout:

$$f(n) = O(g(n)) \leftrightarrow \limsup_{n\to\infty} \frac{f(n)}{g(n)} < \infty, f(n) = o(g(n)) \leftrightarrow \lim_{n\to\infty} \frac{f(n)}{g(n)} = 0, f(n) = O(g(n)) \leftrightarrow g(n) = \varOmega(f(n)),$$
$$f(n) = o(g(n)) \leftrightarrow g(n) = \omega(f(n)), f(n) = \varTheta(g(n)) \leftrightarrow f(n) = O(g(n)),\ and\ f(n) = \varOmega(g(n)).$$

[2] Here, we are ignoring the asymmetry of the nearest neighbor relationship.

could be to assume that two nodes can communicate if they are within certain distance, say $r(n)$, of each other. However, appropriate value of $r(n)$ needs to be determined in this case. It turns out that $r(n) = \Theta(1/\sqrt{n})$ is the right choice (see [16]). The basic intuition is that $r(n)$ should be of the same order as the nearest neighbor distance[3]. In the sequel, we work with this simplified model and assume that $r(n) = \Theta(1/\sqrt{n})$.

2.3 Definitions

We start with the definition of the throughput and the packet delay.

Definition 1 (Throughput). *Let $M_i(t)$ be the number of source node i packets that destination $d(i)$ receives up to time t. Then we say that a long term throughput of $\lambda(n)$ is feasible for $S-D$ pair i if,*

$$\liminf_{T\to\infty} \frac{M_i(T)}{T} \geq \lambda(n) \tag{2.2}$$

Definition 2 (Packet Delay). *It is the time taken by a packet to reach the destination node, starting from its generation at the source node.*

Remark 1. We focus mainly on the queuing delays at the source and the inter-mediate nodes, while ignoring the propagation delays and the processing delays (which are more or less constant).

In order to analyze the delay under various protocols we need to define the following quantities.

Definition 3 (Contact time). *Consider any two arbitrary nodes in the network. Let $T_m(n)$ (meeting time) be the time instant at which they come within distance $r(n)/2$ of each other[4], and let $T_d(n)$ (departing time) be the time instant at which the distance between them exceeds $r(n)$ for the first time after the current meeting time $T_m(n)$. Then, the contact time is defined to be $T_c(n) = T_d(n) - T_m(n)$.*

The purpose of using $r(n)/2$[5] instead of $r(n)$ in defining the *meeting time* is to eliminate some very short time meetings between the nodes. Note, these short time meetings are useless from the point of view of the packet exchange, and further they complicate the analysis.

Definition 4 (First Meeting Time). *Consider two arbitrary nodes in the network, say i and j, with positions $X_i(t)$ and $X_j(t)$, respectively, at time t. Then, the first meeting time of nodes i and j ($F_{ij(n)}$), is defined as*

$$F_{ij}(n) = \inf\{t \geq 0 : d_S[X_i(t), X_j(t)] \leq r(n)\},$$

where d_S is the distance along the surface of the sphere.

Definition 5 (Inter-meeting time). *Consider an arbitrary pair of nodes in the network, the inter-meeting time for this pair is defined to be the time between two successive contacts of the pair, excluding the* contact time *itself.*

We now proceed to describe the mobility models that we consider in this paper.

2.4 Mobility Models

Various mobility models have been proposed in the literature for modeling the motion of the nodes in MANETs. The two most widely used mobility models among them are the random walk mobility model, and the random way-point mobility model (see [12]). From a mathematical standpoint they belong to the same class of Markovian mobility models and yet represent two extremes. The random walk mobility model represents random memoryless roaming, whereas the random way-point mobility model represents directional movement with memory.

[3] Note, since we consider the overall network area to be fixed, the nearest neighbor distance scales as $\Theta(1/\sqrt{n})$.

[4] Note, that this happens on multiple occasions, so consider one such occasion.

[5] Using $r(n)/2$ is not necessary here, we could as well use $\epsilon r(n)$, where $\epsilon < 1$.

Brownian Mobility Model (BMM): In this model, each node executes an independent Brownian walk on a sphere. The random walk mobility model is just a discrete time approximation of this model. Let us consider the motion of a single node under the BMM. Let $\theta(t)$ and $\phi(t)$ denote the longitude and co-latitude, respectively, of the position of the node, at time t. Then, $0 \leq \theta(t) \leq \pi$ and $0 \leq \phi(t) < 2\pi$.

It can be shown (see [15]) that the (Itô) stochastic differential equations for the process are given by

$$d\theta_t = \sigma dU_t + \frac{\sigma^2}{2\tan\theta_t}dt \tag{2.3}$$

and

$$d\phi_t = \frac{\sigma}{\sin\theta_t}dV_t \tag{2.4}$$

where U_t and V_t are independent B.M's, and σ^2 their common variance (see [11]). Let $X_t = \cos\theta_t$, then from (2.3) we obtain

$$dX_t = -\sigma^2 X_t dt - \sigma\sqrt{1 - X_t^2}dU_t \tag{2.5}$$

Thus X_t defines the motion of the node w.r.t. the center of the sphere as the origin. Note, this results in a diffusion process with drift $-\sigma^2 x$ and diffusion coefficient $\sigma\sqrt{1-x^2}$, and is a strongly Feller process.

Using the diffusion process characteristics (see [14], and [15]), one can show the following results (for proof see [16]).

Lemma 1. *In the case of the BMM, the expected contact time, the inter-meeting time, and the first-meeting time scale as $\Theta(r^2(n)/\sigma^2(n))$, $\Theta(1/\sigma^2(n))$ and $\Theta(\log(1/r(n))/\sigma^2(n))$, respectively.*

Further, it is shown in [16] that the distribution of the *first-meeting* time is nearly exponential. However, the *inter-meeting* has a non exponential distribution.

Remark 2. Note, in the case of the BMM it can be easily seen that the *inter-meeting* time is *stochastically smaller* (see [17], for a discussion of stochastic ordering) than the *first meeting* time. This is because, given that two nodes just met each other, it is very likely that they are close by, and hence on an average it takes lesser time to meet again than meeting for the first time.

Using the above remark and noting that the distribution of the of the *first-meeting* time is exponential with mean $\Theta(\log(1/r(n))/\sigma^2(n))$, we can easily obtain the following result.

Lemma 2. *Let $I(n)$ denote the inter-meeting time. Then, in the case of the BMM, we have*

$$\mathbb{E}\{I^2(n)\} = O(\log^2(1/r(n))/\sigma^4(n)).$$

Random Way-point Mobility Model (RWMM): In this model, at each step the mobile node chooses a random destination on the sphere and moves toward it with a random speed. The speed is chosen uniformly in some interval $(0, v_{max}]$. The movement is along the great circle that passes through the initial position and the final destination. On reaching the destination, the node pauses for a random amount of time and the process repeats itself. In this paper, we consider the RWMM without pause times.

Lemma 3. *Consider the RWMM, and let $v(n)$ be the common average speed of the nodes, then the expected contact time is $\Theta(r(n)/v(n))$.*

Proof. Note that a pair of nodes must travel a relative distance of $r(n)/2$ to move out of contact. And since the relative velocity is $\Theta(v(n))$, the result follows. □

The exact analysis for the *first-meeting* time or the *inter-meeting* time is extremely hard in the case of the RWMM, and to the best of our knowledge remains an open problem. However, by means of extensive simulations it has been shown in [16] that the *inter-meeting* is exponentially distributed with mean $\Theta(1/r(n)v(n))$. We use this result at many places in this paper.

2.5 Scaling Laws

In order to estimate the packet delay we need to know how $\sigma(n)$ and $v(n)$ scale with n. However, it turns out that the packet size needs to be adjusted depending upon the scaling behavior of $\sigma(n)$ and $v(n)$. To understand this, let $T_p(n)$ denote the packet transmission time or the slot-length. Clearly, the communicating nodes must remain in contact with each other for a duration of time greater than or equal to the packet transmission time. This means that $T_c(n)$ should be $\omega(T_p(n))$. In this paper, we assume $T_c(n)$ to be $\Theta(T_p(n))$, because this gives us the smallest possible order of the packet delay[6] for a given order of the packet size.

In this paper, we consider the following two, perhaps extreme, possibilities:
Fixed Packet Size Strategy (FPSS): $v(n) = \sigma(n) = \Theta(1/\sqrt{n})$[7]. This is motivated by the fact that in a real network each node occupies some finite area, and hence the area of a network with n nodes should scale as $\Theta(n)$, whereas in our network model we considered the area to be fixed. Thus to compensate for this fact $\sigma(n)$ and $v(n)$ must scale down as $\Theta(1/\sqrt{n})$. Using Lemmas 1 and 3, and noting that $r(n) = \Theta(1/\sqrt{n})$, it follows that $T_c(n)$ is $\Theta(1)$, for both the models. Thus $T_p(n)$ must also be $\Theta(1)$, i.e., fixed independent of the network size; hence the name.
Variable Packet Size Strategy (VPSS): $v(n) = \sigma(n) = \Theta(1)$. This might be possible if the physical area of the nodes is negligible in comparison to the network area and the nodes are moving with extremely high speeds. Using Lemmas 1 and 3, and noting that $r(n) = \Theta(1/\sqrt{n})$, it follows that $T_c(n) = T_p(n) = \Theta(1/n)$ for the BMM, and $T_c(n) = T_p(n) = \Theta(1/\sqrt{n})$ for the RWMM. Thus $T_p(n)$ must vary with n in this case; hence the name.

3 Delay, Capacity, And Distributed 2-Hop Relaying Protocol (D2HRP)

Consider a network with n nodes, and n source-destination (S-D) pairs, such that each node generates packets for only one other node and each node is the destination of packets from only one other node[8].

Note that, in most cases the packet delay has two components; the queuing delay at the source node, and the queuing delay at the relay node. Sometimes a packet might be transmitted directly from the source to the destination, eliminating the queuing delay at the relay node. But it is easy to see that S-D transmissions contribute only $\Theta(1/n)$ to the throughput, and hence can be neglected. Henceforth, we assume that every packet travels exactly two hops, i.e, from the source to the relay, and then from the relay to the destination.

Let us consider the source queue at some arbitrary node i in the network. We assume that the input to the source queue is a Bernoulli stream of rate λ_i. The following Lemma gives the queuing delay at such a queue.

Lemma 4. *Consider a network with n nodes running under the D2HRP. If the exogenous packet stream at node i is Bernoulli with rate λ_i, where $\lambda_i < \frac{\phi}{2}$, then*
a) the expected queuing delay ($\mathbb{E}\{W_i^s\}$) at the source node i, is given by

$$\mathbb{E}\{W_i^s\} = T_p(n)\frac{1-\lambda_i}{\frac{\phi}{2}-\lambda_i} \tag{3.6}$$

b) and the output process from the queue is Bernoulli stream of rate λ_i.

Proof. We know that in the case of the D2HRP, for large enough n, the total number of successful sender-receiver pairs per slot is roughly $n\phi$. Since R-D transmissions and S-R transmissions are given equal priority, the service at any source is Bernoulli with rate $\frac{\phi}{2}$. Now (a) follows by noting that the source i is a Bernoulli(λ_i)/Bernoulli($\frac{\phi}{2}$) queue. And (b) follows from the reversibility of the Bernoulli/Bernoulli queues. □

Now let us consider any arbitrary S-D pair in the network. The packets from the source can reach the destination via any of the $n-2$ relay nodes. As in [2], we assume that all relay nodes maintain a

[6] Note, the packet delay depends on $T_c(n)$, in fact it is proportional to $T_c(n)$.

[7] Note, this does not mean that $v(n)$ and $\sigma(n)$ are equal. It only means that they both scale as $\Theta(1/\sqrt{n})$.

[8] It is easy to see that the throughput capacity reduces if multiple nodes are trying to communicate with one particular node. An extreme case could be $n-1$ nodes trying to communicate with the same node (multi-access scenario). In this case, the per node throughput capacity is $\Theta(1/n)$.

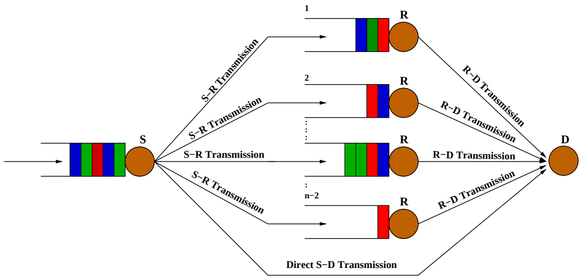

Figure 1. Network model for a particular S-D pair. The queue at the source is a Bernoulli(λ)/Bernoulli($\frac{\phi}{2}$) queue. Each packet either goes directly to the destination or through one of the $n-2$ relay nodes. The relay queues are Bernoulli/Bernoulli in the case of the RWMM, and GI/GI/1-FCFS queues in the case of the BMM.

separate queue for each of the S-D pairs (see Figure 1). By symmetry, all such relay queues are identical. Consider one particular relay queue, the arrival to the queue occurs when the source node sends a packet to the relay node, and the departure from the queue takes place when the relay node delivers a packet to the destination node. Clearly, the *inter-arrival* time and the *inter-departure* time are of the same order as the *inter-meeting* time of any two arbitrary nodes in the network. We are now ready to estimate the overall queuing delay.

Proposition 1. *Consider a network with n nodes, running under the D2HRP. If all the nodes move in accordance with the RWMM, and if the exogenous packet stream at node i is Bernoulli with rate λ_i, where $\lambda_i < \frac{\phi}{2}$, then the expected delay for source node i's packets is $\Theta(T_p(n)n)$.*

Proof. The exponential distribution of the *inter-meeting* time implies that all the relay queues are Bernoulli/Bernoulli queues. Further, as long as the arrival rates at the corresponding source nodes are less than $\frac{\phi}{2}$, the queues are stable. Now Lemma 3, $T_p(n) = \Theta(T_c(n))$, and the fact that the expected *inter-meeting* time is $\Theta(1/r(n)v(n))$ together imply that the *inter-meeting* time is $\Theta(T_p(n)n)$. Consider any arbitrary relay queue for the source node i, the inter-arrival and the inter-departure time for the queue are both $\Theta(T_p(n)n)$. Application of the standard delay formula for the Bernoulli/Bernoulli queue, shows that the delay at the relay queue is $\Theta(T_p(n)n)$. Since the delay at the source queue is $\Theta(T_p(n))$, the overall delay is $\Theta(T_p(n)n)$. □

Corollary 1. *Consider a network with n nodes running under the D2HRP. If all the nodes move in accordance with the RWMM, then the expected packet delay is $\Theta(n)$ under the FPSS, and $\Theta(\sqrt{n})$ under the VPSS.*

Let us now consider the BMM. Since the distribution of the *inter-meeting* time is not known under the BMM, the exact order of the delay is hard to find. However, since the successive *inter-meeting* times are independent, we can upper bound the delay using Kingman's bound [18] given for a GI/G/1-FCFS queue (when the average load is close to 1).

Proposition 2. *Consider a network with n nodes running under the D2HRP. If all the nodes move in accordance with the BMM, and if the exogenous packet stream at node i is Bernoulli with rate λ_i, where $\lambda_i < \frac{\phi}{2}$, then the expected delay for the source node i's packets is $O(T_p(n)nlog^2(n))$.*

Proof. Clearly, all the relay queues are GI/GI/1-FCFS. As long as $\lambda_i < \frac{\phi}{2}$, the queues are stable. Now, since the *inter-meeting* time is $\Theta(T_p(n)n)$, the *inter-arrival* times and the *inter-departure* times are also $\Theta(T_p(n)n)$. Also, from Lemmas 1, 2, and $T_p(n) = \Theta(T_c(n))$, it follows that the second moments of the *inter-arrival* time and the *inter-departure* time are bounded by $O(T_p^2(n)n^2\log^2(n))$. A simple application of Kingman's bound [18], now gives the result. □

Corollary 2. *Consider a network with n nodes running under the D2HRP. If all the nodes move in accordance with the BMM, then the expected packet delay is $O(nlog^2(n))$ under the FPSS, and $O(log^2(n))$ under the VPSS.*

Our simulation results indicate that the expected packet delay is $\Theta(T_p(n)n)$, under both the models. Figure 2 shows the variation of the packet delay with the number of nodes under the BMM with $T_p(n) = \Theta(1)$. Similar results have been obtained in the case of the RWMM as well.

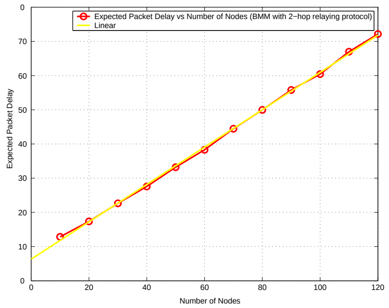

Figure 2. The expected packet delay (in s) under the BMM and the distributed 2-hop scheduling for $R = 1m$, $T_p(n) = 1ms$, $\sigma(n) = \frac{4}{\sqrt{n}}$, and $r(n) = \frac{2}{\sqrt{n}}m$.

Remark 3. As mentioned before, the BMM and the RWMM represent the two extremes of the Markovian mobility models, and one would expect a significant difference in the delay performance under the two models. However, the above results clearly show that the packet delay is almost the same under both the models. This suggests that perhaps the delay is insensitive to the nature of the mobility model, given that it is "reasonable enough". Here "reasonable enough" means that the nodes cannot simply jump from one point in the network to another in no time, as is the case with the *infinite mobility* model considered in [6]. In section 4, we show that the delay performance under the *infinite mobility* model can significantly differ from the models that we consider in this paper.

4 Alternative Protocols For Delay Improvement

In the previous section, we estimated the delay performance under the D2HRP. We now consider some alternative protocols which provide better delay performance by trading off the throughput capacity. In this paper, we restrict ourself to protocols that allow only nearest neighbor transmissions. The following Proposition gives a lower bound on the packet delay under any such protocol.

Proposition 3. *Under the BMM or the RWMM, no scheduling or relaying protocol that allows only nearest neighbor transmissions can guarantee an expected packet delay smaller than* $\Theta(T_p(n)\sqrt{n})$.

Proof. Since the S-D pairs are chosen randomly, each packet travels an expected distance of $\Theta(1)$ before reaching the destination node. Clearly, for the packet transmission to be possible $T_c(n)$ must be $\Omega(T_p(n))$. Now using Lemmas 1 and 3, we see that in every time-slot a packet travels a relative distance of $O(1/\sqrt{n})$ toward the destination. Hence, a packet requires at least $\Theta(\sqrt{n})$ time-slots to reach the destination, and the result follows. □

Clearly, since the packet delay under the D2HRP is much larger than the above mentioned lower bound, there is a possibility that the packet delay might be reduced by some alternative protocols. For this purpose, we describe the following protocols which are extensions of the protocols proposed in [6] for a *cell partitioned network*.

4.1 Distributed 2-Hop Relaying Protocol With Redundancy (D2HRP-WR)

In this protocol, each source node sends duplicate copies of the packet to new relay nodes, whenever possible. Either the source node or one of the relay nodes can deliver the packet to the destination. We assume that each source node maintains a separate send number (SN). The source node increments SN before sending a new packet, and all subsequent copies of the packet are sent with same SN. Similarly, each destination node maintains a request number (RN), which is delivered to the transmitter before R-D or S-D transmission. To start with, let us assume that the source node can send at most $k(n)$ duplicate copies of the packet to distinct nodes. Later, we derive optimal values of $k(n)$ under both the mobility models. More specific details about the protocol are given in [16]. Due to space constraints, we start with the analysis straight away.

Let us analyze the delay performance of the D2HRP-WR under the RWMM. Let $T_1^k(n)$ be the time required by the source node to send the duplicate copies of the packet to $k(n) = o(n)$ other nodes, and let $T_2^k(n)$ be the time required by one of these $k(n)$ nodes to deliver the packet to the destination. It is clear that $\mathbb{E}\{T_1^k(n)\} = \Theta(T_p(n)k(n))$. The following Lemma shows that $\mathbb{E}\{T_2^k(n)\}$ is $\Theta\left(\frac{T_p(n)n}{k(n)}\right)$.

Lemma 5. *In the case of the RWMM, we have*

$$\mathbb{E}\{T_2^k(n)\} = \Theta\left(\frac{T_p(n)n}{k(n)}\right) \tag{4.1}$$

Proof. Note that each packet goes through a Bernoulli/Bernoulli queue before reaching the destination. Hence, each packet incurs a delay which is a Geometric random variable with mean $\Theta(T_p(n)n)$. Further assuming that the delays incurred by different packets are independent[9] the overall delay is the minimum of $k(n)$ i.i.d Geometric random variables with mean $\Theta(T_p(n)n)$, and hence the result follows. □

The next proposition shows that the D2HRP-WR can achieve an expected packet delay of $\Theta(T_p(n)\sqrt{n})$ under the RWMM.

Proposition 4. *In the case of the RWMM, the distributed* 2*-hop relaying protocol can achieve an expected packet delay of $\Theta(T_p(n)\sqrt{n})$ with a throughput capacity of $\Theta(1/\sqrt{n})$.*

Proof. (a) *Lower bound on the expected packet delay*: First note that during the initial $\sqrt{n}$ time-slots, there are less than $\sqrt{n}$ nodes holding the packet. And hence the expected delay is more than the expected delay in the case of $\sqrt{n}$ nodes trying to deliver a packet to the destination. Substituting $k(n) = \sqrt{n}$ in Lemma 5 we see that the expected delay is $\Omega(T_p(n)\sqrt{n})$.
(b) *Upper bound on the expected packet delay*: The upper bound follows by noting that the expected packet delay is $O(\mathbb{E}\{T_1^{\sqrt{n}}(n)\}+\mathbb{E}\{T_2^{\sqrt{n}}(n)\})$, and appealing once again to Lemma 5.
Since each packet is transmitted $\sqrt{n}$ times, the per node throughput capacity is $\Theta(1/\sqrt{n})$. □

Simulations results in the case of the RWMM are in complete agreement with the results derived above. Due to space constraints we do not present the simulation results here.

Let us now consider the BMM. In order to bound the packet delay we make an extra assumption.

Assumption 1. As soon a relay node comes into contact with the destination, it flushes out all the packets that it has for that destination.

Remark 4. This assumption can be justified provided the queue lengths at all the relay queues are bounded by a constant. This can always be ensured by a proper adjustment of the packet arrival rate at the corresponding source node.

Now we are ready to give an upper bound on the minimal achievable delay in the case of the BMM.

Proposition 5. *In the case of the BMM, distributed 2-hop relaying with redundancy can achieve an expected delay of $O(T_p(n)\sqrt{n\log n})$.*

Proof. In the case of the BMM, in view of assumption 1, we can upper bound the expected delay ($D(n)$) as follows:

$$\begin{aligned} D(n) &\leq T_1^k(n) + T_2^k(n) \\ &\leq \Theta(T_p(n)k(n)) + \Theta\left(\frac{T_p(n)n\log n}{k(n)}\right) \end{aligned} \tag{4.2}$$

where, the last step follows by noting that the *first meeting* time is *stochastically greater* than the *inter-meeting* time, and is exponential with mean $\Theta(T_p(n)n\log n)$. Now the result follows by choosing $k(n)$ to be $\sqrt{n\log n}$. □

Since each packet is transmitted to $\sqrt{n\log n}$ relay nodes the per-node throughput capacity scales as $\Theta(1/\sqrt{n\log n})$. The simulation results in the case of the BMM (see Figure 3), show that the expected delay scales as $\Theta(T_p(n)\sqrt{n})$, under $\sqrt{n}$ as well as $\sqrt{n\log n}$ redundancy. Once again we see the delay performance to be almost the same under the BMM and the RWMM.

[9] Note that the purpose of this assumption is to simplify the proof and it is not really necessary (see [16]).

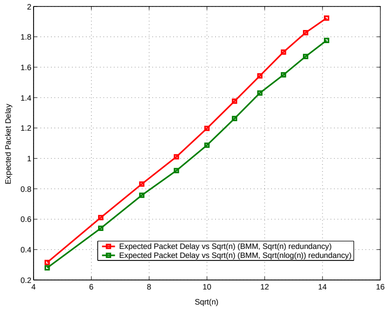

Figure 3. The expected packet delay (in s) under the BMM and the 2-hop relaying with redundancy of $\sqrt{n}$ and $\sqrt{n \log n}$, for $R = 1m$, $T_p(n) = 1ms$, $\sigma(n) = \frac{10}{\sqrt{n}}$, and $r(n) = \frac{1}{\sqrt{n}}m$.

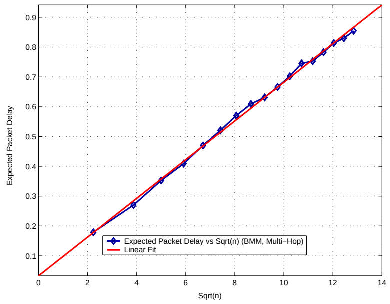

Figure 4. The expected packet delay (in s) under the BMM and the distributed multi-hop scheduling with $R = 1m$, $T_p(n) = 1ms$, $\sigma(n) = \frac{4}{\sqrt{n}}$, and $r(n) = \frac{2}{\sqrt{n}}m$.

4.2 Distributed Multi-Hop Relaying Protocol (DMHRP)

The protocols that we have considered so far, allowed a maximum of one relay node, in the path from the source to the destination. The DMHRP, as the name suggests, allows for an unlimited number of relay nodes. The idea behind this strategy is to improve the packet delay by simply flooding the network. Consider a packet generated at an arbitrary source node. First, the source sends the packet to some other relay node. Then, both the source and the relay node, transmit the packet to other nodes whenever possible. This process continues until the destination node has received the packet[10]. It is obvious that the delay incurred in this case, would be lesser than the delay incurred with D2HPR-WR. This is because, here all the nodes which receive the packet can send it further to other nodes and there is no constraint on the maximum number of duplicates that can be generated at any node.

Since the DMHRP allows only nearest neighbor transmissions, the lower bound of $\Theta(T_p(n)\sqrt{n})$ on the packet delay holds. Further since the D2HRP-WR achieves the lower bound of $\Theta(T_p(n)\sqrt{n})$ on delay under the RWMM, the DMHRP must also achieve this lower bound. In the case of the BMM, similar argument shows that the delay under the DMHRP must be $O(T_p(n)\sqrt{nlogn})$, and $\Omega(T_p(n)\sqrt{n})$.

In fact the simulation results show that the expected delay scales as $\Theta(T_p(n)\sqrt{n})$ under the BMM as well as the RWMM. The BMM results are shown in the Figure 4. The RWMM results are omitted for brevity.

It is clear that under our mobility models, the DMHRP does not offer any significant delay improvement over the distributed D2HRP-WR. Further, because of the flooding involved we would expect the throughput capacity to much smaller than in the case of the D2HRP. Thus, in our case, the DMHRP performs much worse than the D2HRP-WR.

This is unlike the case in [6], where multi-hop scheduling offers an expected delay of $\Theta(\log n)$ as opposed to an expected delay of $\Theta(\sqrt{n})$ under the 2-hop relaying with redundancy. The difference lies in the mobility models, in [6] the authors assume *infinite mobility*, and in any time-slot a node can be anywhere in the network with equal probability. It is easy to see that proposition 3 given in this paper does not hold for *infinite mobility* model considered in [6].

5 Delay and Capacity Trade-off

In the previous sections, we estimated the delay and the throughput capacity under various protocols. The results clearly show that the delay can only be improved at the cost of throughput capacity, and vice versa. In fact, considering all the protocols discussed before, it is easy to see that $delay/capacity \geq \Theta(T_p(n)n)$ is satisfied in all the cases. Proposition 6 (see [16], for a proof) shows that this is indeed a necessary trade-off. Due to space limitations we do not include the proof here.

Proposition 6. *Consider a network with n mobile nodes which are moving in accordance with the RWMM, and n S-D pairs, each generating traffic at an expected rate of $\lambda(n)$ packets per time-slot. Consider a control protocol which is used to stabilize the network, and guarantee an expected packet delay of $\mathbb{E}\{D(n)\}$. Then it is both necessary and sufficient to have $\frac{\mathbb{E}\{D(n)\}}{\lambda(n)} \geq \Theta(T_p(n)n)$.*

[10] This requires some mechanism by which all the nodes might be informed about a packet delivery. Such a mechanism might be inefficient in terms of bandwidth consumption, and we might use TTL instead.

Remark 5. The sufficiency of the above trade-off follows directly from the fact that the D2HRP and the D2HRP-WR achieve this trade-off with equality under the RWMM.

Remark 6. Ignoring the logarithmic terms, it is easy to see that $\frac{\mathbb{E}\{D(n)\}}{\lambda(n)} \geq \Theta(T_p(n)n)$ is a necessary and sufficient trade-off under the BMM as well. And in-fact our simulation result do suggest that this trade-off is achieved with equality under the BMM.

Further, the above trade-off is not limited to the settings that have been considered in this paper, in-fact it holds under a lot of other settings as well. For example, it can be shown that the trade-off holds when the nodes are static, or if the node speed is small. In [16], is is shown that this trade-off can provide a common platform for comparing the performance of various scheduling and relaying protocols that have been proposed recently in [6], [7], [8], [9].

6 Concluding Remarks

We analyzed the delay and capacity trade-off in the case of mobile ad hoc networks with users confined to a finite area. We considered the two most widely used entity mobility models for ad hoc networks: 1) Brownian mobility model (BMM), and 2) random way-point mobility model (RWMM). While there are some results available for the random walk models, the RWMM has not been considered in as much detail as here. From a practical standpoint, the RWMM is perhaps more realistic than the other mobility models in the literature.

We estimated the delay under the distributed 2-hop relating prorocol (D2HRP), and provided a fundamental lower bound on the packet delay under a class of protocols. We proposed and analyzed two distributed protocols which provide better delay performance than the D2HRP, by trading off the throughput capacity. We derived a fundamental trade-off relating the delay and the throughput capacity.

Our results inspire a rich set of questions concerning the fundamental limits of the mobile ad hoc networks. A natural question is how representative are the mobility models. We believe that the models considered in this paper are good representatives of the class of random mobility models with finite speed and variance. And the fact that the delay performance under these two models is almost the same suggests that it might be the case more generally with other random mobility models as well. Proving such a result will be the focus of our future work.

References

1. P. Gupta and P.R. Kumar. The capacity of wireless networks. *IEEE Trans. on Information Theory*, IT-46(2):388–404, March 2000.
2. M. Grossglauser and D. N. C. Tse. Mobility increases the capacity of ad-hoc wireless networks. In Proceedings of the *IEEE INFOCOM*, pages 1360–1369, 2001.
3. E. Perevalov and R. Blum. Delay limited capacity of ad hoc networks: Asymptotically optimal transmission and relaying strategy. In Proceedings of the *IEEE INFOCOM*, April 2003.
4. A.Tsirigos and Z.J. Haas. Multipath routing in presence of frequent topological changes. *IEEE Communications Magazine*, Nov. 2001.
5. N. Bansal and Z. Liu. Capacity, delay and mobility in wireless ad-hoc networks. In Proceedings of the *IEEE INFOCOM*, April 2003.
6. M.J. Neely and E. Modiano. Capacity and delay tradeoffs for ad-hoc mobile networks. Preprint, Dept. of EECS, MIT, 2003.
7. S. Toumpis and A. Goldsmith. Large wireless networks under fading, mobility, and delay constraints. In Proceedings of the *IEEE INFOCOM*, March 2004, to appear.
8. X. Lin and N. B. Shroff. Improved delay-capacity trade-off in large mobile ad hoc networks. Preprint, Purdue University, November 2003.
9. A. El Gamal, J. Mammen, B. Prabhakar, and D. Shah. Throughput-delay trade-off in wireless networks. In Proceedings of the *IEEE INFOCOM*, March 2004, to appear.
10. G. Sharma and R. Mazumdar. Scaling laws for capacity and delay in wireless ad hoc networks with random mobility. In Proceedings of the International Conference on Communications, June 2004, to appear.
11. D.R. Brillinger. A particle migrating randomly on a sphere. *J. Theoret. Probab.*, 10(2):429–443, April 1997.
12. T. Camp, J. Boleng, and V. Davies. A survey of mobility models for ad hoc network research. In *Wireless Communications and Mobile Computing (WCMC): Special issue on Mobile Ad Hoc Networking: Research,Trends and Applications*, 2002.
13. B. Hajek, A. Krishna, and R.O. LaMaire. On the capture probability of large number of stations. *IEEE Trans. on Information Theory*, 46(2):254–260, March 2000.
14. R.N. Bhattacharya and E.C. Waymire. *Stochastic Processes with Applications* Wiley, New York, 1990.
15. S. Karlin and H.M. Taylor. *A Second Course in Stochastic Processes*, Academic, New York, 1981.
16. G. Sharma and R. R. Mazumdar. *Delay and Capacity Trade-offs in Wireless Ad Hoc Networks with Random Mobility*, Technical Report, School of ECE, Purdue University, 2004.
17. S. Ross. *Stochastic Processes*, Wiley, New York, 1996.
18. J. F. C. Kingman. Inequalities in the theory of queues. *J. Roy. Statist. Soc. Ser. B*, 32:102–110, 1970.