The Fundamental Capacity-Delay Tradeoff in Large Mobile Ad Hoc Networks

Xiaojun Lin and Ness B. Shroff
School of Electrical and Computer Engineering, Purdue University
West Lafayette, IN 47907, U.S.A.
{linx, shroff}@ecn.purdue.edu

Abstract— **There has been recent interest within the networking research community to understand how mobility can improve the capacity of mobile ad hoc networks. Of particular interest is the achievable capacity under delay constraints. In this paper, we establish the following upper bound on the optimal capacity-delay tradeoff in mobile ad hoc networks for an *i.i.d.* mobility model. For a mobile ad hoc network with n nodes, if the per-bit-averaged mean delay is bounded by $\bar{D}$, then the per-node capacity λ is upper bounded by $\lambda^3 \leq O(\frac{\bar{D}}{n}\log^3 n)$. By studying the condition under which the upper bound is tight, we are able to identify the optimal values of several key scheduling parameters. We then develop a new scheme that can achieve a capacity-delay tradeoff close to the upper bound up to a logarithmic factor. Our new scheme achieves a larger per-node capacity than the schemes reported in previous works. In particular, when the delay is bounded by a constant, our scheme achieves a per-node capacity of $\Theta(n^{-1/3}/\log n)$. This indicates that, for the *i.i.d.* mobility model, *mobility results in a larger capacity than that of static networks even with constant delays.* Finally, the insight drawn from the upper bound allows us to identify limiting factors in existing schemes. These results present a relatively complete picture of the achievable capacity-delay tradeoffs under different settings.**

I. Introduction

Since the seminal paper by Gupta and Kumar [1], there has been tremendous interest in the networking research community to understand the fundamental achievable capacity in wireless ad hoc networks. For a static network (where nodes do not move), Gupta and Kumar show that the per-node capacity decreases as $O(1/\sqrt{n \log n})$[1] as the number of nodes n increases [1]. The capacity of wireless ad hoc networks can be improved when *mobility* is taken into account. When the nodes are mobile, Grossglauser and Tse show that per-node capacity of $\Theta(1)$ is achievable [2], which is much better than that of static networks. This capacity improvement is achieved at the cost of excessive packet delays. In fact, it has been pointed out in [2] that the packet delay of the proposed scheme could be unbounded.

There have been several recent studies that attempt to address the relationship between the achievable capacity and the packet delay in mobile ad hoc networks. In the work by Neely and Modiano [3], it was shown that the maximum achievable per-node capacity of a mobile ad hoc network is bounded by $O(1)$. Under an *i.i.d.* mobility model, the authors of [3] present a scheme that can achieve $\Theta(1)$ per-node capacity and incur $\Theta(n)$ delay, provided that the load is strictly less than the capacity. Further, they show that it is possible to reduce packet delay if one is willing to sacrifice capacity. In [3], the authors formulate and prove a fundamental tradeoff between the capacity and delay. Let the average end-to-end delay be bounded by D. For D between $\Theta(1)$ and $\Theta(n)$, [3] shows that the maximum per-node capacity λ is upper bounded by

$$\lambda \leq O(\frac{D}{n}). \tag{1}$$

The authors of [3] develop schemes that can achieve $\Theta(1)$, $\Theta(1/\sqrt{n})$, and $\Theta(1/(n \log n))$ per-node capacity, when the delay constraint is on the order of $\Theta(n)$, $\Theta(\sqrt{n})$, and $\Theta(\log n)$, respectively.

Inequality (1) leads to the *pessimistic* conclusion that a mobile ad hoc network can sustain at most $O(1/n)$

This work has been partially supported by the NSF grants ANI-0207728, EIA-0130599, and the Indiana 21st Century Fund for Wireless Networks.

[1] We use the following notation throughout:

$$f(n) = o(g(n)) \quad \leftrightarrow \quad \lim_{n\to\infty} \frac{f(n)}{g(n)} = 0,$$

$$f(n) = O(g(n)) \quad \leftrightarrow \quad \limsup_{n\to\infty} \frac{f(n)}{g(n)} < \infty,$$

$$f(n) = \omega(g(n)) \quad \leftrightarrow \quad g(n) = o(f(n)),$$

$$f(n) = \Theta(g(n)) \quad \leftrightarrow \quad f(n) = O(g(n)) \text{ and } g(n) = O(f(n)).$$

per-node capacity with a constant delay bound. This capacity is even worse than that of static networks. It turns out that this pessimistic conclusion is due to certain restrictive assumptions that are implicit in the work in [3] (we will elaborate on these assumptions in Section VI). In fact, Toumpis and Goldsmith [4] present a scheme that can achieve a per-node capacity of $\Theta(n^{(d-1)/2}/\log^{5/2} n)$ when the delay is bounded by $O(n^d)$. The result of [4] has incorporated the effect of fading. If we remove fading, the per-node capacity will be of the order $\Theta(n^{(d-1)/2}/\log^{3/2} n)$. Ignoring the logarithmic term, we find that in [4] the following capacity-delay tradeoff is achievable:

$$\lambda^2 = \Theta(\frac{D}{n}). \qquad (2)$$

This is better than (1). In particular, the authors of [4] present a scheme that can achieve $\Theta(1/(\sqrt{n}\log^{3/2} n))$ per-node capacity with a constant delay bound. (The capacity will be $\Theta(1/(\sqrt{n \log n}))$ with no fading.) This capacity is now *comparable* to that of the static ad hoc networks.

An open question that still remains is: *what is the optimal capacity-delay tradeoff in mobile ad hoc networks*? Inequality (1) is clearly not optimal. The methodology of [4] is constructive in nature. Hence, inequality (2) is only a lower bound. The search for the optimal capacity-delay tradeoff is important for two reasons. First, it will allow us to see where the fundamental limits (i.e., upper bounds) are, and how far existing schemes could possibly be improved. Secondly, as has happened in previous works [1], [3], a careful study of the upper bound is usually able to reveal the delicate tradeoffs inherent to the problem. A complete understanding of these tradeoffs will help us identify the possible points of inefficiency in existing schemes and provide directions for further improvement. The ultimate goal is to find a scheme that can achieve the optimal capacity-delay tradeoff.

This paper accomplishes these two goals. Under the *i.i.d.* mobility model studied in [3], we will first establish an *upper bound* on the optimal capacity-delay tradeoff in mobile ad hoc networks. We will show that, if the per-bit-averaged mean delay is bounded by $\bar{D}$, then the per-node capacity λ is upper bounded by

$$\lambda^3 \leq O(\frac{\bar{D}}{n}\log^3 n). \qquad (3)$$

In Fig. 1, we draw this upper bound alongside the capacity-delay tradeoffs achieved by the schemes in [3] and [4]. There is obviously a gap between the upper bound and what can be achieved by existing schemes.

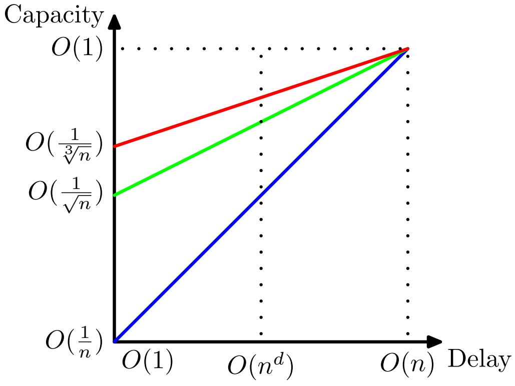

Fig. 1. The achievable capacity-delay tradeoffs of existing schemes compared with the upper bound (ignoring the logarithmic terms). The top line corresponds to our upper bound (achievable by the scheme outlined in Section V up to a logarithmic factor). The middle line is achieved by the scheme in [4], and the bottom one is achieved by the scheme in [3].

Further, in the process of proving the upper bound, we are able to identify the optimal choices for several key parameters of the scheduling policy. We then develop a new scheme that achieves the upper bound on the capacity-delay tradeoff upto a logarithmic factor, which suggests that our upper bound is fairly tight. Our new scheme achieves a larger per-node capacity than the ones in [3] and [4]. In particular, our scheme can achieve $\Theta(n^{-1/3}/\log n)$ per-node capacity with constant delay. Unlike previous works, this result shows that, even for a constant delay bound, the per-node capacity of mobile ad hoc networks can be larger than that of the static networks! Finally, the insight drawn from the upper bound allows us to identify the limiting factors of the schemes in [3] and [4].

The rest of the paper is organized as follows. In Section II, we outline the network and mobility model. In Section III, we prove several key properties that capture various tradeoffs inherent in mobile ad hoc networks. We establish the upper bound on the optimal capacity-delay tradeoff in Section IV and present a scheme in Section V that achieves a capacity-delay tradeoff close to the upper bound. In Section VI, we discuss the existing schemes described in [3] and [4]. Then we conclude.

II. Network and Mobility Model

We consider a mobile ad hoc network with n nodes moving within a unit square[2]. We assume that time is divided into slots of unit length. We assume the

[2] Note that changing the shape of the area from a square to a circle or other topologies will not affect our main results.

following *i.i.d. mobility model* proposed in [3]. At each time slot, the positions of each node are *i.i.d.* and uniformly distributed within the unit square. Between time slots, the distributions of the positions of the nodes are independent. Although the assumption on an *i.i.d.* mobility model is somewhat restrictive, its mathematical tractability allows us to gain important insights into the structure of the problem. We will comment on some extensions to the *i.i.d.* mobility model in the conclusion.

For simplicity, we assume the following traffic model similar to the models in [3], [4]. We assume that the number of nodes n is even and the nodes can be labeled in such a way that node $2i-1$ communicates with node $2i$, and node $2i$ communicates with node $2i-1$, $i = 1, ..., n/2$. The communication between any source-destination pairs can go through multiple other nodes as *relays*. That is, the *source* can either send a message directly to the *destination*; or, it can send the message to one or more *relay* nodes; the relay nodes can further forward the message to other relay nodes (possibly after moving to another position); and finally some relay node forwards the message to the destination.

We assume the following Protocol Model from [1] that governs *direct radio transmissions* between nodes. Let W be the bandwidth of the system. Let X_i denote the position of node i, $i = 1, ..., n$. Let $|X_i - X_j|$ be the Euclidean distance between nodes i and j. At each time slot, node i can communicate *directly* with another node j at W bits per second if and only if the following interference constraint is satisfied [1]:

$$|X_j - X_k| \geq (1+\Delta)|X_i - X_j|$$

for every other node $k \neq i, j$ that is simultaneously transmitting. Here, Δ is some positive number. Note that an alternative model for direct radio transmission is the Physical Model [1], [4]. In the Physical Model, a node can communicate with another node if the signal-to-interference ratio is above a given threshold. It has been shown that, under certain conditions, the Physical Model can be reduced to the Protocol Model with an appropriate choice of Δ [1]. Hence, we will not consider the Physical Model any further in this paper. We also assume that no nodes can transmit and receive over the same frequency at the same time. We further assume the following *separation of time scale*, i.e., radio transmission can be scheduled at a time scale much faster than that of node mobility. This is usually a reasonable assumption in real networks. Hence, a message may be divided into multiple bits and each bit can be forwarded multiple hops separately within a single time slot.

We assume a uniform traffic pattern, that is, all source nodes communicate with their destination nodes at the same rate λ. let $\bar{D}$ be the mean delay averaged over all messages and all source-destination pairs. Both λ and $\bar{D}$ will depend on how the transmissions between mobile nodes are scheduled. We are interested in capturing the fundamental tradeoff between the achievable capacity λ and the delay $\bar{D}$. That is, over all possible ways of scheduling the radio transmissions, what is the maximum per-node capacity λ given certain constraint on the delay $\bar{D}$.

III. Properties of the Scheduling Policies

In this section, we will prove several key results that capture the various tradeoffs inherent in mobile ad hoc networks. We will first define the class of scheduling policies that we will consider. Because we are interested in the fundamental *achievable* capacity for a given delay, we will assume that there exists a scheduler that has all the information about the current and past status of the network, and can schedule any radio transmission in the current and future time slots. At each time slot t, for each bit b that has not been delivered to its destination yet, the scheduler needs to perform the following two functions:

- *Capture:* The scheduler needs to decide whether to deliver the bit b to the destination within the current time slot. If yes, the scheduler then needs to choose one relay node (possibly the source) that has a copy of the bit b at the beginning of the time slot t, and schedule radio transmissions to forward this bit to the destination *within the same time slot*, using possibly multi-hop transmissions. When this happens successfully, we say that the chosen relay node has successfully *captured* the destination of bit b. It is important to forward the bit to the destination within a single time slot. Otherwise, since the chosen relay node may move far away from the destination in the next time slot, the nodes that received the bit b in the current time slot will only count as new relay nodes for the bit b, and they have to capture again in the next time slot.
- *Duplication:* If capture does not occur for bit b, the scheduler needs to decide whether to *duplicate* bit b to other nodes that do not have the bit at the beginning of the time slot t. The scheduler also needs to decide which nodes to relay from and relay to, and how to schedule radio transmissions to forward the bit to these new relay nodes.

In this paper, we will consider the class of *causal* scheduling policies that perform the above two functions at each time slot. The causality assumption essentially requires that, when the scheduler makes the capture decision and the duplication decision, it can only use information about the current and the past status of the network. In particular, at any time slot t, the scheduler cannot use information about the *future* positions of the nodes at any time slot $s > t$.

This class of scheduling policies is clearly very general, and encompasses nearly any practical scheduling scheme we can think of. (Note that even *predictive* scheduling schemes have to rely on current and past information only.) Some remarks on the *capture* process is in order. Although we do allow for other less intuitive alternatives, in a typical scheduling policy a successful *capture* usually occurs when some relay nodes are within an area close to the destination node, so that fewer resources will be needed to forward the information to the destination. For example, a relay node could enter a disk of a certain radius around the destination, or a relay node could enter the same cell as the destination. We call such an area a *capture neighborhood*. The relay nodes that has the bit b at the beginning of the time slot t are called *mobile relays* for bit b. The mobile relay that is chosen to forward the bit b to the destination is called the *last mobile relay* for bit b.

The following examples are illustrative of the possible scheduling policies within this broad class. The schemes in previous works [3], [4] are all special cases or variants of these examples.

Example A: The number of mobile relays R is fixed and the capture neighborhood is chosen to be a disk with a fixed radius ρ around the destination. Once a bit b enters the system, it is immediately broadcast to the nearest $R-1$ neighboring nodes. When any of the R mobile relays (including the source node) move within distance ρ from the destination, the bit b is then forwarded from the nearest mobile relay to the destination.

Example B: The unit area is divided into a number of cells. Once a bit b enters the system, it is immediately broadcast to all other nodes in the same cell. The number of mobile relays for the bit b then stay unchanged. Note that the actual number of mobile relays depends on the number of nodes that reside in the same cell as the source (at the time slot when the bit b enters the system), and is thus a random variable. When one of the mobile relays moves into the same cell as the destination, the bit b is then forwarded from the nearest mobile relay to the destination.

Example C: In the above two schemes, no duplication for bit b is carried out except at the first time slot when the bit enters the system. A more sophisticated strategy is to use an *opportunistic duplication scheme* such as the example below. The unit area is divided into a number of cells. After a bit b enters the system, at each time slot t, if one of the mobile relays moves into the same cell as the destination, bit b is then forwarded from the nearest mobile relay to the destination. Otherwise, the source node (or, alternatively, the current mobile relays) broadcasts the bit to all other nodes that reside at the same cell. Hence, duplication may occur at each time slot until bit b is delivered to its destination.

In the sequel, we will prove several key inequalities that capture the various tradeoffs inherent in the class of scheduling policies outlined above. Intuitively, the larger the number of mobile relays and the larger the capture neighborhood, the smaller the delay. On the other hand, in order to improve capacity, we need to consume fewer radio resources, which implies a smaller number of mobile relays and a shorter distance from the last mobile relay to the destination. As we will see later, these tradeoffs will determine the fundamental relationship between achievable capacity and delay in mobile ad hoc networks.

A. Notations

Fix any scheduling policy. Define the following random variables for each bit b that needs to be communicated from its source node to its destination node. Let $t_0(b)$ denote the time slot when bit b first enters the system. Let $s_b \geq t_0(b)$ be the time slot when the scheduler decides that a successful capture for bit b occurs. The scheduler also needs to choose one mobile relay that has a copy of bit b at the beginning of the time slot s_b to forward bit b towards its destination *within the same time slot* s_b. Let R_b be the number of mobile nodes holding the bit b at the time of capture, let $D_b \triangleq s_b - t_0(b)$ be the number of time slots from the time bit b enters the system to the time of capture, and let l_b denote the distance from the chosen last mobile relay to the destination of bit b. The quantities R_b, D_b, and l_b are essential for the tradeoffs that follow[3]. Note

[3] Using the notion of *filtrations* and *stopping times* [5, p231], we can rigorously define these quantities as random variables on the same probability space where the random mobility of the mobile nodes is defined. The expectation in the following tradeoffs are then taken with respect to this probability space. See [6] for details.

that D_b includes possible queueing delay at the source node or at the relay nodes.

B. Tradeoff I : D_b versus R_b and l_b

Proposition 1: Under the *i.i.d.* mobility model, the following inequality holds for any causal scheduling policy when $n \geq 3$,

$$c_1 \log n \mathbf{E}[D_b] \geq \frac{1}{(\mathbf{E}[l_b] + \frac{1}{n^2})^2 \mathbf{E}[R_b]} \text{ for all bits } b, \quad (4)$$

where c_1 is a positive constant.

The proof is available in [6]. This new result is the cornerstone for deriving the optimal capacity-delay tradeoff in mobile ad hoc networks. It captures the following tradeoff: the smaller the number R_b of mobile relays the bit b is duplicated to, and the shorter the targeted distance l_b from the last mobile relay to the destination, the longer it takes to capture the destination. This seemingly odd relationship is actually motivated by some simple examples. Consider Example A at the beginning of this section. When R_b and the area of the capture neighborhood A_b are constants, then $1-(1-A_b)^{R_b}$ is the probability that any one out of the R_b nodes can capture the destination in one time slot. It is easy to show that, the average number of time slots needed before a successful capture occurs, is,

$$\mathbf{E}[D_b] = \frac{1}{1-(1-A_b)^{R_b}} \geq \frac{1}{A_b R_b}.$$

If, as in Example B, R_b and possibly A_b are random but *fixed after the first time slot* $t_0(b)$, then

$$\mathbf{E}[D_b | R_b, A_b] \geq \frac{1}{A_b R_b}.$$

By Hőlder's Inequality [5, p15],

$$\mathbf{E}^2[\frac{1}{\sqrt{A_b}}] \leq \mathbf{E}[R_b]\mathbf{E}[\frac{1}{A_b R_b}].$$

Hence,

$$\begin{aligned} \mathbf{E}[D_b] &\geq \mathbf{E}[\frac{1}{A_b R_b}] \geq \mathbf{E}^2[\frac{1}{\sqrt{A_b}}]\frac{1}{\mathbf{E}[R_b]} \\ &\geq \frac{1}{\mathbf{E}^2[\sqrt{A_b}]\mathbf{E}[R_b]}, \end{aligned}$$

where in the last step we have applied Jensen's Inequality [5, p14]. Note that on average l_b is on the order of $\sqrt{A_b}$. Hence,

$$\mathbf{E}[D_b] \geq \frac{c_1'}{\mathbf{E}^2[l_b]\mathbf{E}[R_b]} \text{ for all bits } b, \quad (5)$$

where c_1' is a positive constant. It may appear that, when an "opportunistic duplication scheme" such as the one in Example C is employed, such a scheme might achieve a better tradeoff than (5) by starting off with fewer mobile relays and a smaller capture neighborhood, if the node positions at the early time slots after the bit's arrival turns out to be favorable. However, Proposition 1 shows that no scheduling policy can improve the tradeoff by more than a $\log n$ factor.

C. Tradeoff II : Multihop

Once a successful capture occurs, the chosen mobile relay (i.e., the *last mobile relay*) will start transmitting the bit to the destination *within a single time slot*, using possibly other nodes as relays. We will refer to these latter relay nodes as *static relays*. The static relays are only used for forwarding the bit to the destination *after a successful capture occurs*. Let h_b be the number of hops it takes from the last mobile relay to the destination. Let S_b^h denote the transmission range of each hop $h = 1, .., h_b$. The following relationship is trivial.

Proposition 2: The sum of the transmission ranges of the h_b hops must be no smaller than the straight-line distance from the last mobile relay to the destination, i.e.,

$$\sum_{h=1}^{h_b} S_b^h \geq l_b. \quad (6)$$

D. Tradeoff III : Radio Resources

It consumes radio resources to duplicate each bit to mobile relays and to forward the bit to the destination. Proposition 3 below captures the following tradeoff: the larger the number of mobile relays R_b and the further the multi-hop transmissions towards the destination have to traverse, the smaller the achievable capacity. Consider a large enough time interval T. The total number of bits communicated end-to-end between all source-destination pairs is $\lambda n T$.

Proposition 3: Assume that there exist positive numbers c_2 and N_0 such that $D_b \leq c_2 n^2$ for $n \geq N_0$. If the positions of the nodes within a time slot are *i.i.d.* and uniformly distributed within the unit square, then there exist positive numbers N_1 and c_3 that only depend on c_2, N_0 and Δ, such that the following inequality holds for any causal scheduling policy when $n \geq N_1$,

$$\sum_{b=1}^{\lambda n T} \frac{\Delta^2}{4} \frac{\mathbf{E}[R_b]-1}{n} + \mathbf{E}[\sum_{b=1}^{\lambda n T} \sum_{h=1}^{h_b} \frac{\pi \Delta^2}{4} (S_b^h)^2] \leq c_3 W T \log n. \quad (7)$$

The assumption that $D_b \leq c_2 n^2$ for large n is not as restrictive as it appears. It has been shown in [3] that the maximum achievable per-node capacity is $\Theta(1)$ and this capacity can be achieved with $\Theta(n)$ delay. Hence, we are most interested in the case when the delay is not much larger than the order $O(n)$. Further, Proposition 3 only requires that the stationary distribution of the positions of the nodes within a time slot is *i.i.d.* It does not require the distribution between time slots to be independent.

We briefly outline the motivation behind the inequality (7). The details of the proof are quite technical and available in [6]. Consider nodes i, j that directly transmit to nodes k and l, respectively, at the same time. Then, according to the interference constraint:

$$\begin{aligned} |X_j - X_k| &\geq (1+\Delta)|X_i - X_k] \\ |X_i - X_l| &\geq (1+\Delta)|X_j - X_l]. \end{aligned}$$

Hence,

$$\begin{aligned} |X_j - X_i| &\geq |X_j - X_k| - |X_i - X_k| \\ &\geq \Delta|X_i - X_k|. \end{aligned}$$

Similarly,

$$|X_i - X_j| \geq \Delta|X_j - X_l|.$$

Therefore,

$$|X_i - X_j| \geq \frac{\Delta}{2}(|X_i - X_k| + |X_j - X_l|).$$

That is, disks of radius $\frac{\Delta}{2}$ times the transmission range centered at the transmitter are disjoint from each other[4]. This property can be generalized to *broadcast* as well. We only need to define the transmission range of a broadcast as the distance from the transmitter to the furthest node that can successfully receive the bit. The above property motivates us to measure the radio resources each transmission consumes by the areas of these disjoint disks [1]. For unicast transmissions from the last mobile relay to the destination, the area consumed by each hop is $\frac{\pi\Delta^2}{4}(S_b^h)^2$. For duplication to other nodes, broadcast is more beneficial since it consumes fewer resources. Assume that each transmitter chooses the transmission range of the broadcast independently of the positions of its neighboring nodes. If the transmission range is s, then on average no greater than $n\pi s^2$ nodes can receive the broadcast, and a disk of radius $\frac{\Delta}{2}s$ (i.e., area $\frac{\pi\Delta^2}{4}s^2$) centered at the transmitter will be disjoint from other disks. Therefore, we can use $\frac{\Delta^2}{4}\frac{\mathbf{E}[R_b]-1}{n}$ as a lower bound on the expected area consumed by duplicating the bit to $R_b - 1$ mobile relays (excluding the source node). This lower bound will hold even if the duplication process is carried out over multiple time slots, because the average number of *new* mobile relays each broadcast can cover is at most proportional to the area consumed by the broadcast. Therefore, inspired by [1], the amount of radio resources consumed must satisfy

$$\sum_{b=1}^{\lambda nT} \frac{\Delta^2}{4}\frac{\mathbf{E}[R_b]-1}{n} + \mathbf{E}[\sum_{b=1}^{\lambda nT}\sum_{h=1}^{h_b}\frac{\pi\Delta^2}{4}(S_b^h)^2] \leq c_3' WT, \tag{8}$$

where c_3' is a positive constant.

However, $\frac{\Delta^2}{4}\frac{\mathbf{E}[R_b]-1}{n}$ may fail to be a lower bound on the expected area consumed by duplicating to $R_b - 1$ mobile relays if the following *opportunistic broadcast scheme* is used. The source may choose to broadcast *only when there are a larger number of nodes close by*. If the source can afford to wait for these "good opportunities", an *opportunistic broadcast scheme* may consume less radio resources than a non-opportunistic scheme to duplicate the bit to the same number of mobile relays. Nonetheless, we can show that, when n is large, the probability that the number of nodes in a neighborhood is orders of magnitude larger than its average value will be very small. We can then prove Proposition 3, which implies that no scheduling policies can improve the tradeoff by more than a $\log n$ factor. For details, please refer to [6].

E. Tradeoff IV : Half Duplex

Finally, since we assume that no node can transmit and receive over the same frequency at the same time (a practically necessary assumption for most wireless devices), the following property can be shown as in [1].

Proposition 4: The following inequality holds,

$$\sum_{b=1}^{\lambda nT}\sum_{h=1}^{h_b} 1 \leq \frac{WT}{2}n. \tag{9}$$

IV. The Upper Bound on the Capacity-Delay Tradeoff

Our first main result is to derive, from the above four tradeoffs, the upper bound on the optimal capacity-delay tradeoff of mobile ad hoc networks under the *i.i.d.* mobility model. Since the maximum achievable per-node capacity is $\Theta(1)$ and this capacity can be achieved with $\Theta(n)$ delay by the scheme of [3], we are only interested in the case when the mean delay is $o(n)$.

Proposition 5: Let $\bar{D}$ be the mean delay averaged over all bits and all source-destination pairs, and let λ be the throughput of each source-destination pair. If

[4] A similar observation is used in [1] except that they take a receiver point of view.

$\bar{D} = O(n^d), 0 \le d < 1$, the following upper bound holds for any causal scheduling policy,

$$\lambda^3 \le O(\frac{\bar{D}}{n}\log^3 n).$$

Proof: Using the Cauchy-Schwartz inequality, we have

$$\begin{aligned}\left(\sum_{b=1}^{\lambda nT}\sum_{h=1}^{h_b} S_b^h\right)^2 &\le \left(\sum_{b=1}^{\lambda nT}\sum_{h=1}^{h_b} 1\right)\left(\sum_{b=1}^{\lambda nT}\sum_{h=1}^{h_b}(S_b^h)^2\right)\\ &\le \frac{WTn}{2}\sum_{b=1}^{\lambda nT}\sum_{h=1}^{h_b}(S_b^h)^2,\end{aligned} \tag{10}$$

where in the last step we have used Tradeoff IV (9). Equality holds in (10) when inequality (9) is tight and when S_b^h is equal for all b and h. We thus have,

$$\begin{align*}\mathbf{E}[\sum_{b=1}^{\lambda nT}\sum_{h=1}^{h_b}(S_b^h)^2] &\ge \frac{2}{WTn}\mathbf{E}[\left(\sum_{b=1}^{\lambda nT}\sum_{h=1}^{h_b} S_b^h\right)^2]\\ &\ge \frac{2}{WTn}\left(\mathbf{E}[\sum_{b=1}^{\lambda nT}\sum_{h=1}^{h_b} S_b^h]\right)^2 \tag{11}\\ &\ge \frac{2}{WTn}\left(\sum_{b=1}^{\lambda nT}\mathbf{E}[l_b]\right)^2, \tag{12}\end{align*}$$

where in the last two steps we have used Jensen's Inequality and the Tradeoff II (6), respectively. Inequality (11) is tight when $\sum_{b=1}^{\lambda nT}\sum_{h=1}^{h_b} S_b^h$ is almost surely a constant, and (12) is tight when (6) is tight.

From Tradeoff I (4), we have

$$\sum_{b=1}^{\lambda nT}\mathbf{E}[R_b] \ge \sum_{b=1}^{\lambda nT}\frac{1}{c_1\log n}\frac{1}{(\mathbf{E}[l_b]+\frac{1}{n^2})^2\mathbf{E}[D_b]}. \tag{13}$$

Let

$$\bar{D} = \frac{\sum_{b=1}^{\lambda nT}\mathbf{E}[D_b]}{\sum_{b=1}^{\lambda nT} 1} = \frac{\sum_{b=1}^{\lambda nT}\mathbf{E}[D_b]}{\lambda nT}.$$

Using Jensen's Inequality and Hőlder's Inequality, we have,

$$\begin{aligned}\frac{1}{\left(\frac{\sum_{b=1}^{\lambda nT}(\mathbf{E}[l_b]+\frac{1}{n^2})}{\sum_{b=1}^{\lambda nT} 1}\right)^2} &\le \left(\frac{\sum_{b=1}^{\lambda nT}\frac{1}{(\mathbf{E}[l_b]+\frac{1}{n^2})}}{\sum_{b=1}^{\lambda nT} 1}\right)^2\\ &\le \frac{\sum_{b=1}^{\lambda nT}\frac{1}{(\mathbf{E}[l_b]+\frac{1}{n^2})^2\mathbf{E}[D_b]}}{\sum_{b=1}^{\lambda nT} 1}\frac{\sum_{b=1}^{\lambda nT}\mathbf{E}[D_b]}{\sum_{b=1}^{\lambda nT} 1}.\end{aligned} \tag{14}$$

Equality holds when $\mathbf{E}[l_b]$ is the same for all b and $\mathbf{E}[D_b] = \bar{D}$ for all b. Substituting (14) in (13), we have

$$\sum_{b=1}^{\lambda nT}\mathbf{E}[R_b] \ge \frac{1}{c_1\log n}\frac{\left(\sum_{b=1}^{\lambda nT} 1\right)^3}{\bar{D}\left(\sum_{b=1}^{\lambda nT}(\mathbf{E}[l_b]+\frac{1}{n^2})\right)^2}. \tag{15}$$

Substituting (12) and (15) into Inequality (7), we have

$$\begin{aligned}\frac{4c_3WT\log n}{\Delta^2} &\ge \sum_{b=1}^{\lambda nT}\frac{\mathbf{E}[R_b]-1}{n} + \pi\mathbf{E}[\sum_{b=1}^{\lambda nT}\sum_{h=1}^{h_b}(S_b^h)^2]\\ &\ge \frac{1}{c_1 n\log n}\frac{(\lambda nT)^3}{\bar{D}\left(\sum_{b=1}^{\lambda nT}(\mathbf{E}[l_b]+\frac{1}{n^2})\right)^2}\\ &\qquad +\frac{2\pi}{WTn}(\sum_{b=1}^{\lambda nT}\mathbf{E}[l_b])^2 - \lambda T.\end{aligned}$$

There are two cases that we need to consider.

Case 1: If $\sum_{b=1}^{\lambda nT}\mathbf{E}[l_b] \le \frac{\lambda T}{n}$, then

$$\begin{aligned}\frac{4c_3WT\log n}{\Delta^2} &\ge \frac{1}{c_1 n\log n}\frac{(\lambda nT)^3}{\bar{D}\left(\frac{2\lambda T}{n}\right)^2} - \lambda T\\ &= \frac{1}{4c_1\log n}\frac{\lambda Tn^4}{\bar{D}} - \lambda T.\end{aligned}$$

When $\bar{D} = O(n^d), d < 1$, the first term dominates when n is large. Hence, for n large enough,

$$\begin{aligned}\frac{4c_3WT\log n}{\Delta^2} &\ge \frac{1}{8c_1\log n}\frac{\lambda Tn^4}{\bar{D}}\\ \lambda &\le \frac{32c_1c_3W}{\Delta^2}\frac{\bar{D}\log^2 n}{n^4}.\end{aligned} \tag{16}$$

Case 2: If $\sum_{b=1}^{\lambda nT}\mathbf{E}[l_b] \ge \frac{\lambda T}{n}$, then

$$\begin{aligned}&\frac{4c_3WT\log n}{\Delta^2}\\ &\ge \frac{1}{c_1 n\log n}\frac{(\lambda nT)^3}{\bar{D}\left(2\sum_{b=1}^{\lambda nT}\mathbf{E}[l_b]\right)^2}\\ &\qquad +\frac{2\pi}{WTn}(\sum_{b=1}^{\lambda nT}\mathbf{E}[l_b])^2 - \lambda T\end{aligned} \tag{17}$$

$$\begin{aligned} &\geq\ 2\sqrt{\frac{1}{c_1 \log n}\frac{2\pi}{WTn^2}\frac{(\lambda nT)^3}{4\bar{D}}} - \lambda T \qquad (18) \\ &=\ 2\sqrt{\frac{\pi}{2c_1 \log n}\frac{\lambda^3 nT^2}{\bar{D}W}} - \lambda T. \qquad (19) \end{aligned}$$

Therefore, either

$$\lambda \leq O(\frac{\bar{D}\log n}{n}), \qquad (20)$$

or, if $\lambda = \omega(\frac{\bar{D}\log n}{n})$, then the first term in (19) dominates when n is large. In the latter case, for n large enough,

$$\begin{aligned} \frac{4c_3 WT\log n}{\Delta^2} &\geq \sqrt{\frac{\pi}{2c_1 \log n}\frac{\lambda^3 nT^2}{\bar{D}W}} \\ \lambda^3 &\leq \frac{32c_1c_3^2W^3}{\Delta^4}\frac{\bar{D}\log^3 n}{n}. \qquad (21) \end{aligned}$$

Finally, we compare the three inequalities we have obtained, i.e., (16), (20) and (21). Since $\bar{D} = o(n^d), d < 1$, inequality (21) will eventually be the loosest for large n. Hence, the optimal capacity-delay tradeoff is upper bounded by

$$\lambda^3 \leq O(\frac{\bar{D}}{n}\log^3 n).$$

■

V. An Achievable Lower Bound on the Capacity-Delay Tradeoff

The capacity-delay tradeoff in Proposition 5 is better than those reported in [3] and [4]. Assuming that the delay bound is $\Theta(n^d)$, $0 \leq d < 1$, the achievable per-node capacity is $O(n^{-(1-d)})$ by the scheme in [3], and $O(n^{-(1-d)/2})$ by the scheme in [4]. Our upper bound, however, implies a per-node capacity of $O(n^{-(1-d)/3})$ (we have ignored all $\log n$ factors). Since $d < 1$, there is clearly room to substantially improve existing schemes (see Fig. 1). In this section, we will show how the study of the upper bound also helps us in developing a new scheme that can achieve a capacity-delay tradeoff that is close to the upper bound. Then, in the next section, we will identify the limiting factors of existing schemes in [3], [4] that prevent them from achieving the optimal capacity-delay tradeoff.

A. *Choosing the Optimal Values of the Key Parameters*

Assume that the mean delay is bounded by $n^d, d < 1$. By Proposition 5, we have,

$$\lambda \leq \Theta(\sqrt[3]{\frac{\bar{D}}{n}\log^3 n}) = \Theta(n^{\frac{d-1}{3}}\log n).$$

TABLE I

The order of the optimal values of the parameters when the mean delay is bounded by n^d.

R_b: # of Duplicates	$\Theta(n^{(1-d)/3})$
l_b: Distance to Destination	$\Theta(\frac{n^{-(1+2d)/6}}{\sqrt{\log n}})$
h_b: # of Hops	$\Theta(\frac{n^{(1-d)/3}}{\log n})$
S_b^h: Transmission Range of Each Hop	$\Theta(\sqrt{\frac{\log n}{n}})$

In order to achieve the maximum capacity on the right hand side, all inequalities (10)-(18) should hold with equality. By studying the conditions under which these inequalities are tight, we will be able to identify the optimal choices of various key parameters of the scheduling policy. For example, by checking the conditions when (10)-(14) are tight, we can infer that the parameters (such as $S_b^h, \mathbf{E}[l_b], \mathbf{E}[D_b]$) of each bit b should be about the same and should concentrate on their respective average values. This implies that the scheduling policy should use the same parameters for all bits. Further, note that equality holds in (18) if and only if

$$\frac{1}{4c_1 n\log n}\frac{(\lambda nT)^3}{\bar{D}(\sum\limits_{b=1}^{\lambda nT}\mathbf{E}[l_b])^2} = \frac{2\pi}{WTn}(\sum_{b=1}^{\lambda nT}\mathbf{E}[l_b])^2.$$

From here we can solve for l_b, h_b, S_b^h and R_b. The results are summarized in Table I. The details of the derivation is available in [6].

Several remarks are in order. Since it is sufficient to control all parameters around these optimal values, simple cell-based schemes such as the one in Example B of Section III suffice. Secondly, the optimal values for R_b and l_b can provide guidelines on how to choose the cell partitioning. Thirdly, the optimal value for S_b^h is roughly the average distance between neighboring nodes when n nodes are uniformly distributed in a unit square. Hence, it is desirable to use multi-hop transmission over neighboring nodes to forward the information from the last mobile relay to the destination. These guidelines have sketched a blueprint of the optimal scheduling scheme for us. We next present schemes that can achieve capacity-delay tradeoffs that are close to the upper bound up to a logarithmic factor.

B. *Achievable Capacity with* $\Theta(1)$ *Delay*

We first focus on the case when the mean delay is bounded by a constant, i.e., the exponent $d = 0$. By Proposition 5, the per-node throughput is bounded by $O(n^{-1/3}\log n)$. We now present a scheme that can achieve $\Theta(n^{-1/3}/\log n)$ capacity with $\Theta(1)$ delay for large n. This is an encouraging result for mobile networks because we know that the per-node capacity of static networks is $O(1/\sqrt{n\log n})$ [1]. Hence, mobility increases the capacity even with constant delay.

We will need the following Lemma before stating the main scheduling scheme. We will repeatedly use the following type of cell-partitioning. Let m be a positive integer. Divide the unit square into $m \times m$ cells (in m rows and m columns). Each cell is a square of area $1/m^2$. As in [4], we call two cells *neighbors* if they share a common boundary, and we call two nodes *neighbors* if they lie in the same or neighboring cells. We say that a group of cells can be *active* at the same time when one node in each cell can successfully transmit to or receive from a neighboring node, subject to the interference from other cells that are active at the same time. Let $\lfloor x \rfloor$ be the largest integer smaller than or equal to x. The proof of the following Lemma is available in [6].

Lemma 6: There exists a scheduling policy such that each cell can be active for at least $1/c_4$ amount of time, where c_4 is a constant independent of m.

The capacity achieving scheme is as follows.

Capacity Achieving Scheme:

1) At each *odd* time slot, we schedule transmissions from the sources to the relays. We divide the unit square into $g_1(n) = \lfloor\left(\frac{n^{2/3}}{8\log n}\right)^{\frac{1}{2}}\rfloor^2$ cells. We refer to each cell in the odd time slot as a *sending cell*. By Lemma 6, each cell can be active for $\frac{1}{c_4}$ amount of time. When a cell is scheduled to be active, each node in the cell broadcasts a new message to all other nodes in the same cell for $\frac{1}{32c_4 n^{1/3}\log n}$ amount of time. These other nodes then serve as mobile relays for the message. The nodes within the same sending cell coordinate themselves to broadcast sequentially. If any sending cell has more than $32n^{1/3}\log n$ nodes, we refer to it as a Type-I error [4].

2) At each *even* time slot, we schedule transmissions from the mobile relays to the destination nodes. Note that the positions of the mobile relays have changed and are now independent of their positions in the previous time slot. We divide the unit square into $g_2(n) = \lfloor\left(n^{1/3}\right)^{\frac{1}{2}}\rfloor^2$ cells. We refer to each cell in the even time slot as the *receiving cell*. For any *receiving cell* $i = 1, ..., g_2(n)$ and any *sending cell* $j = 1, ..., g_1(n)$, pick a node Y_{ij} that is in the *receiving cell* i in the current time slot and that was in the *sending cell* j in the previous time slot. We refer to this node Y_{ij} as the *designated mobile relay in* receiving cell i and *for* sending cell j. If there is no such node Y_{ij} for any i or j, we refer to it as a Type-II error. There may be multiple nodes that can serve as the designated mobile relay for some i, j. In this case we only pick one. Note that if no Type-II errors occur, each destination node can now find a designated mobile relay that holds the message intended for the destination node and that resides in the same receiving cell. We then schedule multihop transmissions in the following fashion to forward each message from the designated mobile relay to its destination in the same receiving cell. We further divide each receiving cell i into $g_3(n) = \lfloor\left(\frac{n^{2/3}}{4\log n}\right)^{\frac{1}{2}}\rfloor^2$ *mini-cells* (in $\sqrt{g_3(n)}$ rows and $\sqrt{g_3(n)}$ columns). Each mini-cell is a square of area $1/(g_2(n)g_3(n))$. By Lemma 6, there exists a scheduling scheme where each mini-cell can be active for $\frac{1}{c_4}$ amount of time. When each mini-cell is active, it forwards a message (or a part of a message) to one other node in the neighboring mini-cell. If the destination of the message is in the neighboring cell, the message is forwarded directly to the destination node. The messages from each designated mobile relay are first forwarded towards neighboring cells along the X-axis, then to their destination nodes along the Y-axis. In this fashion, a successful schedule will allow each destination node to receive a message of length $\frac{W}{32c_4 n^{1/3}\log n}$ from its respective designated mobile relay residing in the same receiving cell. For details on constructing such a schedule, see [6]. If no such schedule exists, we refer to it as a Type-III error. At the end of each even time slot, if there are any packets that cannot be delivered to the destination nodes due to Type-II or Type-III errors, they are dropped.

We can show that, as $n \to \infty$, the probabilities of errors of all types will go to zero. The following proposition thus holds. For space consideration, we do not provide the proof here. It is available in [6].

Proposition 7: With probability approaching one, as $n \to \infty$, the above scheme allows each source to send a message of length $\frac{W}{32c_4 n^{1/3}\log n}$ to its respective destination node within two time slots.

Remark: Our scheme uses different cell-partitioning in the odd time slots than that in the even time slots. Note that in previous works [3], [4], the cell structure remains the same over all time slots. Our judicious choice of the cell-structures is the key to our tighter lower bound for the capacity. In particular, the size of the sending cell is chosen such that the average number of nodes in each cell, $n/g_1(n) = \Theta(n^{1/3}\log n)$, is close to the optimal value of R_b in Section V-A (with $d = 0$). The size of the receiving cell is chosen such that its area, $1/g_2(n) = \Theta(n^{1/3})$, is close to the optimal value of l_b^2. Finally, the size of the mini-cell is chosen such that each hop to the neighboring cell is of length $1/\sqrt{g_2(n)g_3(n)} = \Theta(\sqrt{\log n/n})$, which is close to the optimal value of S_b^h.

C. The Effect of Queueing

When we defined the delay D_b of each bit b in Section III, it includes the possible queueing delay at the source node and at the relay nodes. The upper bound on the capacity-delay tradeoff (Proposition 5) thus holds regardless of the queueing discipline used in the system, and $\bar{D}$ also includes the queueing delay. We now show how to analyze the queueing delay of the capacity-achieving scheme in Section V-B. This scheme attempts to deliver one message of length $\frac{W}{32c_4n^{1/3}\log n}$ for each source-destination pair every two time slots. Let p_e be the probability that a message is successfully delivered to the destination at the end of the even time slot. (Note that p_e is the same for all source-destination pairs due to symmetry, and by Proposition 7, $p_\mathrm{e} \to 1$ as $n \to \infty$.) Assume that if such delivery is unsuccessful, messages that have not been delivered to the destinations at the end of each even time slot are discarded and have to be retransmitted at the source nodes. Further, assume that packets of length $\frac{W}{32c_4n^{1/3}\log n}$ arrive at each source according to certain stochastic process. Then packets may get enqueued at the source nodes. If we observe the system at the end of each even time slot, the number of packets queued for each source-destination pair will evolve as that of a discrete-time queue with geometric service time distributions [7], and the queues for each source-destination pair can be studied independently. If we know the packet arrival process, we can then compute the queueing delay. For example, if the arrival process is Bernoulli, i.e., one new packet for each source-destination pair arrives at the source every two time slots with probability Λ, then using standard results for discrete time $M/M/1$ queues [7, p82], we can compute the queueing delay as,

$$\mathcal{D} = 2\frac{1-\Lambda}{p_\mathrm{e}-\Lambda}.$$

As $n \to \infty$, $p_\mathrm{e} \to 1$. Hence,

$$\mathcal{D} \to 2, \text{ as } n \to \infty.$$

On the other hand, if the arrival process is Poisson with rate Λ, then the number of packets arriving at a source-destination pair every two time slots is a Poisson random variable with mean 2Λ. Hence, using results for discrete time $M^{a_n}/M/1$ queues [7, p89], we can compute the queueing delay as

$$\mathcal{D} = 2\frac{1-\Lambda}{p_\mathrm{e}-2\Lambda}.$$

Assume $2\Lambda \le 1-\epsilon$, where $0 < \epsilon < 1$. As $n \to \infty$, $p_\mathrm{e} \to 1$. Hence,

$$\mathcal{D} \to 2\frac{1-\Lambda}{\epsilon}, \text{ as } n \to \infty.$$

Note that in both cases, the queueing delay is at most a constant multiple of 2 (time slots) provided that ϵ (i.e., the difference between the arrival rate and the capacity) is positive and bounded away from zero as $n \to \infty$. Hence, the capacity-achieving scheme in Section V-B can sustain $\Theta(n^{-1/3}/\log n)$ throughput (in bits per time slot) with $O(1)$ *queueing delay*.

D. The Capacity Achieving Scheme for Arbitrary Delay Bound

The above scheme can be generalized to arbitrary delay bounds. Let the mean delay be bounded by $\bar{D} = \Theta(n^d)$, $0 \le d < 1$. We can group every $\lfloor n^d \rfloor + 1$ time slots into a *super-frame*. In each *odd* super-frame, we schedule transmissions from the sources to the relays. We divide the unit square into $\Theta(n^{(2+d)/3}/\log n)$ sending cells of equal area. Within each sending cell, each source broadcasts a new message to all other nodes within the same cell for a duration of $\Theta(\frac{1}{n^{(1-d)/3}\log^2 n})$ every time slot.

In each *even* super-frame, we schedule transmissions from the relays to the destination nodes. We divide the unit square into $\Theta(n^{(1+2d)/3})$ receiving cells of equal area. In every time slot, some mobile relays will have messages intended for some other destination nodes in the same receiving cell. We then schedule multi-hop transmissions to deliver the messages from the mobile relays to the destination nodes in the same receiving cell.

Using similar techniques as the one in [4] and the one in Section V-B, we can show that, with probability approaching one as $n \to \infty$, each source can send $\lfloor n^d \rfloor + 1$ messages of length $\Theta(n^{-(1-d)/3}/\log^2 n)$ to its destination within $2(\lfloor n^d \rfloor + 1)$ time slots. The queueing delay can also be studied in a similar fashion as in Section V-C. The details are omitted because of space constraints.

VI. The Limiting Factors in Existing Schemes

In Section V, we have shown that choosing the optimal values of the scheduling parameters is the key to achieving the optimal capacity-delay tradeoff. In this section, we will show that deviating from these optimal values will lead to suboptimal capacity-delay tradeoffs. In particular, we will identify the limiting factors in the existing schemes in [3] and [4] by comparing the optimal values of scheduling parameters in Section V-A with those used by the existing schemes. Our model in Section IV can be extended to study the upper bounds on the capacity-delay tradeoff when one imposes additional restrictive assumptions that correspond to these limiting factors. We will see that these new upper bounds are inferior to the capacity-delay tradeoff reported in Sections IV and V. The existing schemes of [3] and [4] in fact achieve capacity-delay tradeoffs that are close to the respective upper bounds. These results will give us new insight on which schemes to use under different conditions.

A. The Limiting Factor in the Scheme of [3]

The scheme by Neely and Modiano [3] divides the unit square into n cells each of area $1/n$. A mobile relay will forward messages to the destination only when they both reside in the same cell. Hence, the distance from the last mobile relay to the destination, l_b, is on average on the order of $O(1/\sqrt{n})$, regardless of the delay constraints. However, we have shown in Section V-A that the optimal choice for l_b should be on the order of $\Theta(n^{-(1+2d)/6}\log^{-1/2} n)$, when the mean delay is bounded by $\Theta(n^d)$. The next Proposition shows that the restrictive choice of l_b is indeed the limiting factor of the scheme in [3]. The proof is available in [6].

Proposition 8: Let $\bar{D}$ be the mean delay averaged over all bits and all source-destination pairs, and let λ be the throughput of each source-destination pair. If $\bar{D} = O(n^d), 0 \leq d < 1$ and $\mathbf{E}[l_b] = O(1/\sqrt{n})$, then for any causal scheduling policy,

$$\lambda \leq O(\frac{\bar{D}}{n}\log^2 n).$$

Remark: The scheme of [3] achieves the above upper bound up to a logarithmic factor.

B. The Limiting Factor in the Scheme of [4]

In the scheme by Toumpis and Goldsmith [4], a mobile relay will always use single-hop transmission to forward the messages directly to the destination. That is, the number of hops from the last mobile relay to the destination node, h_b, is always 1. However, we have shown in Section V-A that the optimal value of h_b is $\Theta(n^{(1-d)/3}/\log n)$ when the mean delay is bounded by $\Theta(n^d)$. The next Proposition shows that the restriction on h_b is indeed the limiting factor of the scheme in [4]. The proof is available in [6].

Proposition 9: Let $\bar{D}$ be the mean delay averaged over all bits and all source-destination pairs, and let λ be the throughput of each source-destination pair. If $\bar{D} = O(n^d), 0 \leq d < 1$ and $h_b = O(1)$, then for any causal scheduling policy,

$$\lambda^2 \leq O(\frac{\bar{D}}{n}\log^3 n).$$

Remark: The scheme of [4] achieves the above upper bound up to a logarithmic factor.

Propositions 5, 8 and 9 present three different upper bounds on the capacity-delay tradeoff of mobile ad hoc networks under different assumptions. Assume that the mean delay is bounded by $n^d, 0 \leq d < 1$. When the capacity is the main concern, Proposition 5 shows that the per-node throughput is at most $O(n^{-(1-d)/3}\log n)$. The capacity-achieving scheme reported in Section V can achieve close to this upper bound up to a logarithmic factor. However, this capacity-achieving scheme requires sophisticated coordination among the mobile nodes. Hence, it may not be suitable *when simplicity is the main concern*. On the other hand, the scheme of [3] only requires coordination among nodes that are within a cell of area $1/n$. Note that the average number of nodes in such a cell is $\Theta(1)$. Proposition 8 then shows that, when coordination among a large number of nodes is prohibited, the scheme of [3] is close to optimal. Similarly, the scheme of [4] only requires single-hop transmissions from the mobile relays to the destinations. Proposition 9 shows that, when multi-hop transmissions are undesirable, the scheme of [4] is close to optimal. Therefore, the results reported in this paper present a relatively complete picture of the achievable capacity-delay tradeoffs under different conditions.

An interesting open problem for future work is to investigate whether these insights apply to the capacity-delay tradeoff under mobility models other than the *i.i.d.* model. For example, [8] and [9] have studied the capacity-delay tradeoff under the Brownian Motion mobility model. In these works, the authors also have implicit restrictions on the scheduling policy. In particular, the scheme in [8] uses at most one mobile relay at any time (i.e., $R_b = 1$), and the scheme in [9] schedule a transmission from the mobile relay to the destination only when they are at a distance of $O(1/\sqrt{n})$ away (i.e., $l_b = O(1/\sqrt{n})$). In this paper, we have shown under the *i.i.d.* mobility model that, in order to achieve the optimal capacity-delay tradeoff, R_b, l_b and h_b should all vary with the delay exponent d. Putting restrictions on any one of these variables will lead to suboptimal capacity for a given delay constraint. For our future work, we plan to study whether these kind of restrictions will also limit the capacity-delay tradeoff obtained in existing works under other mobility models.

VII. Conclusion and Future Work

In this paper, we have studied the fundamental capacity-delay tradeoff in mobile ad hoc networks under the *i.i.d.* mobility model. Our contributions are three-fold. We have established the upper bound on the optimal capacity-delay tradeoff over all causal scheduling policies. The upper bound not only provides the fundamental limits of capacity and delay, but also helps to identify the optimal values of the key scheduling parameters in order to achieve the optimal capacity-delay tradeoff. Our second contribution is to develop a new scheduling scheme that can achieve a capacity-delay tradeoff that differs from the upper bound only by a logarithmic factor, which also implies that our upper bound is fairly tight. The capacity achievable by our new scheme is larger than that of the existing schemes in [3] and [4]. In particular, when the delay is bounded by a constant, our scheme achieves a per-node capacity of $\Theta(n^{-1/3}/\log n)$. This indicates that, under the *i.i.d.* mobility model, mobility increases the capacity even with constant delays. Our third contribution is to use the insight drawn from the upper bound to identify the limiting factors in the existing schemes. These results present a relatively complete picture of the achievable capacity-delay tradeoffs under different considerations.

In this paper, we have assumed an *i.i.d.* mobility model. For future work, we plan to study the optimal capacity-delay tradeoff for mobile ad hoc networks under other mobility models. Among the properties that we proved in Section III, we expect that the Tradeoffs II to IV will be relatively invariant to the choice of mobility models, while Tradeoff I is likely to depend on a specific model. Hence, future work will concentrate on how to tailor Tradeoff I for other mobility models. Some immediate extensions to the *i.i.d.* mobility model are possible. For example, at each time slot, each node may independently choose to stay in its old position with probability p, and to move to a new random position with probability $1-p$. This model may approximate scenarios where nodes move at a fast speed and then stay for a relatively long period of time. Tradeoff I will hold for this extension of the *i.i.d.* mobility model, and hence our main results will hold as well. Other mobility models that we plan to investigate are, the Brownian motion mobility model [8], [9], the random waypoint model [9], [10], and the linear mobility model [11], etc.

Other aspects to consider are how the upper bound will be impacted by the use of diversity coding [12], effect of fading [4], and the use of information-theoretic approaches [13], [14].

References

[1] P. Gupta and P. R. Kumar, “The Capacity of Wireless Networks,” *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 388–404, March 2000.

[2] M. Grossglauser and D. Tse, “Mobility Increases the Capacity of Ad Hoc Wireless Networks,” *IEEE/ACM Transactions on Networking*, vol. 10, no. 4, August 2002.

[3] M. J. Neely and E. Modiano, “Capacity and Delay Trade-offs for Ad-Hoc Mobile Networks,” *submitted to IEEE Transactions on Information Theory, available at http://www-rcf.usc.edu/∼mjneely/*, 2003.

[4] S. Toumpis and A. J. Goldsmith, “Large Wireless Networks under Fading, Mobility, and Delay Constraints,” in *Proceedings of IEEE INFOCOM*, Hong Kong, China, March 2004.

[5] R. Durrett, *Probability : Theory and Examples*, 2nd ed. Belmont, CA: Duxbury Press, 1996.

[6] X. Lin and N. B. Shroff, “ The Fundamental Capacity-Delay Tradeoff in Large Mobile Ad Hoc Networks,” *Technical Report, Purdue University, http://min.ecn.purdue.edu/∼linx/papers.html*, 2004.

[7] M. E. Woodward, *Communication and Computer Networks: Modelling with Discrete-Time Queues*. Los Alamitos, CA: IEEE Computer Society Press, 1994.

[8] A. E. Gamal, J. Mammen, B. Prabhakar, and D. Shah, “Throughput-Delay Trade-off in Wireless Networks,” in *Proceedings of IEEE INFOCOM*, Hong Kong, China, March 2004.

[9] G. Sharma and R. R. Mazumdar, “Delay and Capacity Trade-offs for Wireless Ad Hoc Networks with Random Mobility,” *preprint available at http://www.ece.purdue.edu/∼mazum/*, October 2003.

[10] N. Bansal and Z. Liu, “Capacity, Delay and Mobility in Wireless Ad-Hoc Networks,” in *Proceedings of IEEE INFOCOM*, San Francisco, CA, April 2003.

[11] S. Diggavi, M. Grossglauser, and D. Tse, “Even One-Dimensional Mobility Increases Ad Hoc Wireless Capacity,” in *ISIT 02*, Lausanne, Switzerland, June 2002.

[12] E. Perevalov and R. Blum, “Delay Limited Capacity of Ad hoc Networks: Asymptotically Optimal Transmission and Relaying Strategy,” in *Proceedings of IEEE INFOCOM*, San Francisco, CA, April 2003.

[13] M. Gastpar and M. Vetterli, “On the Capacity of Wireless Networks: The Relay Case,” in *Proceedings of IEEE INFOCOM*, New York, June 2002.

[14] P. Gupta and P. R. Kumar, “Towards an Information Theory of Large Networks: An Achievable Rate Region,” *IEEE Transactions on Information Theory*, vol. 49, no. 8, pp. 1877–1894, August 2003.