Delay and Capacity Trade-Offs in Mobile Ad Hoc Networks: A Global Perspective

Gaurav Sharma, Ravi Mazumdar, *Fellow, IEEE*, and Ness B. Shroff, *Fellow, IEEE*

***Abstract*—Since the original work of Grossglauser and Tse, which showed that mobility can increase the capacity of an ad hoc network, there has been a lot of interest in characterizing the delay–capacity relationship in ad hoc networks. Various mobility models have been studied in the literature, and the delay–capacity relationships under those models have been characterized. The results indicate that there are trade-offs between the delay and capacity, and that the nature of these trade-offs is strongly influenced by the choice of the mobility model. Some questions that arise are: (i) How representative are these mobility models studied in the literature? (ii) Can the delay–capacity relationship be significantly different under some other "reasonable" mobility model? (iii) What sort of delay–capacity trade-off are we likely to see in a real world scenario? In this paper, we take the first step toward answering some of these questions. In particular, we analyze, among others, the mobility models studied in recent related works, under a unified framework. We relate the nature of delay–capacity trade-off to the nature of node motion, thereby providing a better understanding of the delay–capacity relationship in ad hoc networks in comparison to earlier works.**

***Index Terms*—Ad-hoc networks, capacity, delay, mobility, throughput, trade-offs, wireless.**

I. Introduction

In THIS PAPER, we study the delay and capacity trade-offs in mobile ad hoc networks. This line of investigation started with the seminal work of Gupta and Kumar [1], in which they defined and studied the notion of "capacity" for wireless ad hoc networks. Note that their notion of capacity is distinct from the information theoretic notion of capacity studied, for example, in [2], [3]. More precisely, unlike in information theoretic studies of capacity, the physical layer modulation and coding scheme was kept fixed in their work. Their main finding was that the capacity of a random wireless network with n static nodes scales as $\Theta\left(\sqrt{\frac{n}{\log n}}\right)$, provided all transmissions are carried out at the same power level. Recent works [4], [5] have shown that a throughput of $\Theta\left(\sqrt{n}\right)$ is achievable if the nodes are allowed to exercise power control.

Note that since the system capacity is shared between n nodes, the per-node throughput scales as $\Theta\left(1/\sqrt{n}\right)$. Thus, each node gets smaller and smaller throughput as the network size grows, thereby implying that static ad hoc networks are not scalable. The researchers have also investigated the possible improvement in capacity that can be achieved by adding a relatively small number base stations [6], [7] or mobile relay nodes [8] to a static ad hoc network.

An interesting observation was made in [9], where the authors considered a mobile ad hoc network. They showed that node mobility can significantly increase the capacity of an ad hoc network. In fact, they showed that even constant, non-zero, per-node throughput is achievable in case of mobile ad hoc networks. Traditionally, node mobility has been looked upon as a "liability" in wireless networks. For example, in cellular networks, mobile nodes can move from one cell to another, thereby necessitating a "hand-off". This has led to significant research on hand-off protocols (see, for example, [10]). The mobility also has an adverse effect on the performance of traditional ad hoc routing protocols (see, for example, [11]). The work in [9] has shown that mobility can be an "asset", if properly exploited. However, delay related issues were not considered in [9].

Since both capacity as well as delay are important from an application point of view, a significant effort has recently been devoted within the networking research community to understand the delay–capacity relationship in ad hoc networks. Bansal and Liu [8] were among the first to study the delay–capacity relationship in wireless networks. They considered a wireless network consisting of static sender-destination pairs and mobile relays, and proposed a geographic routing scheme that achieves a near optimal capacity and studied its delay performance. Perevalov and Blum [12] studied the delay limited capacity of mobile ad hoc networks. Their work was motivated by the *diversity coding* approach given in [13].

Recent works of Neely and Modiano [14], [15], El Gamal *et al.* [16], Toumpis and Goldsmith [17], Sharma and Mazumdar [18], [19], Lin and Shroff [20], and Lin *et al.* [21], have all studied the delay–capacity trade-offs in mobile ad hoc networks. The mobility models and the network settings studied in these works differ considerably, and so do the delay–capacity trade-offs that are reported. The mobility models that have been studied include the i.i.d. mobility model [14], [17], [20]; random way-point mobility model [18], [19]; Brownian mobility model [18], [16], [21]; and random walk mobility model [16], [22].

Since the network settings considered in these works are quite different, it is difficult to single out the impact the nature of node mobility has on the delay–capacity trade-off. For example, both [14] and [20] study the i.i.d. mobility model, but the trade-offs reported in these works are quite different. In particular, in [14], the authors consider a cell partitioned network setting and show

Manuscript received October 18, 2006; approved by IEEE/ACM Transactions on Networking Editor Z.-L. Zhang.

G. Sharma and N. B. Shroff are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: gsharma@ecn.purdue.edu; shroff@ecn.purdue.edu).

R. Mazumdar is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: mazum@ece.uwaterloo.ca).

Digital Object Identifier 10.1109/TNET.2007.905154

1063-6692/$25.00 © 2007 IEEE

that the trade-off: $\lambda \leq \Theta(\frac{D}{n})$, where D is the average packet delay and λ is the average per-node throughput, is both necessary as well as sufficient under their setting. Under a less restrictive network setting in [20], the delay–capacity trade-off is shown to be: $\lambda^3 \leq \Theta\left(\frac{D \log^3 n}{n}\right)$. Thus, we see that the delay–capacity trade-off depends not only on the nature of the node mobility, but also on the network setting.

The main insight we provide in this paper[1] is that there is a critical value of delay (henceforth denoted by *critical delay*), below which, the node mobility cannot be exploited for improving the capacity. We also show that the critical delay depends on the nature of node mobility, but not so much on the network setting. In terms of the notion of critical delay (see Section III for a rigorous definition), some recent results in the literature can be summarized as follows:

- The critical delay under the Brownian motion model and random walk model is roughly $\Theta(n)$[2] (see [21], [22]).
- The critical delay under the random way-point mobility model is roughly $\Theta(\sqrt{n})$ (see [19]).
- The critical delay under the i.i.d. mobility model is $\Theta(1)$ (see [20]).

Observing the above results, it is natural to ask: (i) How representative are the above mentioned mobility models? (ii) Can the critical delay and, in general, the delay–capacity relationship be significantly different under some other mobility model? (iii) What scaling for critical delay are we likely to see in a real world scenario? This paper makes the following contributions toward answering these fundamental questions:

- We propose and study a family of *hybrid random walk models*, and show that they exhibit a continuous range of critical delays in-between those of the i.i.d. mobility model and random walk mobility model. In particular, for every β between 0 and 1/2, there exists a mobility model in the family of *hybrid random walk models* for which the critical delay is roughly $\Theta\left(n^{2\beta}\right)$. As expected, $\beta = 1/2$ corresponds to the random walk mobility model, and $\beta = 0$ corresponds to the i.i.d. mobility model.
- We propose and study a family of *discrete random direction models* exhibiting a continuous range of critical delays in-between those of the random way-point mobility model and the Brownian mobility model. In particular, for every α between 0 and 1/2, there exists a mobility model in the class of *discrete random direction models* for which the critical delay is roughly $\Theta\left(n^{1/2+\alpha}\right)$. Note that for $\alpha = 0$ and 1/2 the corresponding discrete random direction models are similar to the random way-point mobility model and Brownian mobility model, respectively. These models approximate the motion of nodes under commonly used random direction models in the literature (see, for example, [24], [25][3]), and are simpler to analyze.

An interesting feature of the above classes of mobility models is that for all mobility models in these classes, the delay under the 2-hop relaying scheme is roughly $\Theta(n)$; which is in line with the other mobility models considered in the literature. Our results therefore show that the mobility models considered in the literature are in some sense extreme: they either exhibit the smallest critical delays (i.i.d. mobility model and random way-point mobility model) or the largest critical delays (Brownian motion model and random walk mobility model), among the mobility models in their respective classes.

The rest of the paper is organized as follows. We introduce the *hybrid random walk models* and *discrete random direction models*, and discuss our system model in Section II. We define the notions of critical delay and 2-hop delay in Section III. We study the critical delay and 2-hop delay under the hybrid random walk models in Section IV, and under the discrete random direction models in Section V. A discussion on the implications of our results for large mobile ad hoc networks is provided in Section VI. Finally, we end this paper with some concluding remarks in Section VII.

II. The Model

A. System Model

We consider an ad hoc network consisting of n mobile nodes, distributed uniformly on a unit square S. We consider a homogeneous scenario in which each node generates traffic at the same rate. Further, we assume that each node, say node i, generates traffic for exactly one other node, say node $b(i)$, and that the mapping $i \mapsto b(i)$ is bijective. We also assume that the packet arrival process at each node is independent of the node mobility process.

The communication between any *source-destination* pair can possibly be carried out via multiple other nodes, acting as relays. That is, a *source* node can, if possible, send a packet directly to its *destination* node; or, the source node can forward the packet to one or more *relay* nodes; the relay nodes can also forward the packet to other relay nodes; and finally, a relay node or the source node itself can deliver the packet to its destination node.

For simplicity, we assume that the success or failure of a transmission between a pair of nodes is governed by the protocol model of [1]. Let W be the bandwidth of the system in bits per second. Let X_t^i denote the position of node i, for $i = 1 \ldots n$, at time t. Under the protocol model, node i can communicate directly with node j at a rate of W bits per second at time t, if and only if, the following interference constraint is satisfied [1]:

$$d\left(X_t^k, X_t^j\right) \geq (1+\Delta) d\left(X_t^i, X_t^j\right) \tag{1}$$

for every other node $k \neq i, j$ that is simultaneously transmitting. Here Δ is some positive number; and $d(x,y)$ is the distance between points $x = (x_1, x_2), y = (y_1, y_2) \in S$, defined as follows:

$$d(x,y) = \min_{x^i \in [x], y^j \in [y]} \left\| x^i - y^j \right\|$$

where $[x]$ is the set of points $\{(x_1, x_2), (x_1 - 1, x_2), (x_1 + 1, x_2), (x_1, x_2 - 1), (x_1, x_2 + 1)\}$, and $[y]$ is defined similarly (the motivation for this definition will become clear in Section II-B).

[1] Most of the material in this paper appears in [23].

[2] Note that in [21], the results are stated in terms of the variance parameter $\sigma^2(n)$. We have set $\sigma^2(n) = 1/n$ for the sake of easy comparison with the other results.

[3] In [25], such models are referred to as random walk models.

Note that for a packet to be successfully received by node j, the above interference constraint must be satisfied over the entire duration of the packet transmission from node i to node j.

We use the following definition of the throughput: Let $\lambda_i(t)$ be the total number of bits delivered end-to-end for destination i up to time t, then the throughput $\lambda(n)$ of the system is given by

$$\lambda(n) = \liminf_{t\to\infty} \frac{1}{n}\sum_{i=1}^{n}\frac{\lambda_i(t)}{t}.$$

Next, we describe our mobility models, starting with the hybrid random walk models.

B. Hybrid Random Walk Models

These models are parameterized by a single parameter β, that takes values between 0 and 1/2. The unit square is divided into $n^{2\beta}$ squares of area $1/n^{2\beta}$ each (henceforth referred to as cells), resulting in a discrete torus of size $n^{\beta} \times n^{\beta}$. Each cell is then further divided into $n^{1-2\beta}$ square subcells of area $1/n$ each,[4] as shown in Fig. 1. Time is divided into slots of equal duration. At each slot a node is assumed to be in one of the subcells inside a cell. Initially, each node is equally likely to be in any of the n subcells, independent of the other nodes. At the beginning of a slot, a node jumps from its current subcell to one of the subcells in an adjacent cell, chosen in an uniformly random fashion. By adjacent cell we mean the following: Let $(i,j) : i,j = 0,1,\ldots,n^{\beta}-1$, be a numbering of the cells of the 2-D torus, as shown in Fig. 2. The cells adjacent to cell (i,j) are the cells $(i+1,j)$, $(i-1,j)$, $(i,j+1)$, and $(i,j-1)$, where the addition and subtraction operations are performed modulo n^{β}. Note that for $\beta = 0$, the above mobility model is essentially the i.i.d. mobility model considered in [14], [17], [20]; and for $\beta = 1/2$, it is the same as the random walk model of [16].

Next, we describe the random direction models (see, for example, [24], [25]).

C. Random Direction Models

These models are parameterized by a single parameter α, that takes values between 0 and 1/2. The initial position of each node is assumed to be uniformly distributed within S. The motion of each node under these models is independent and identical to the other nodes. The motion of a node is divided into multiple trips. At the beginning of a trip, the node chooses a direction θ uniformly between $[0, 2\pi]$, and moves a distance of $n^{-\alpha}$ in that direction, with a speed v_n; and the process repeats itself. Note that we are assuming a complete wrap-around of S, resulting in a unit torus (see Fig. 3).

Remark 1: In the rest of the paper, we consider $v_n = \Theta(1/\sqrt{n})$, as in [19]. This particular choice of node speed is motivated by the fact that we keep the network area fixed and let the number of nodes increase to infinity, which means that the average neighborhood size scales as $\Theta(1/\sqrt{n})$. In order to account for this "shrinking" of the neighborhood size, we scale the node velocity as $\Theta(1/\sqrt{n})$. Note that alternatively one could scale the area of the network in proportion to n, while keeping the velocity of the nodes fixed.

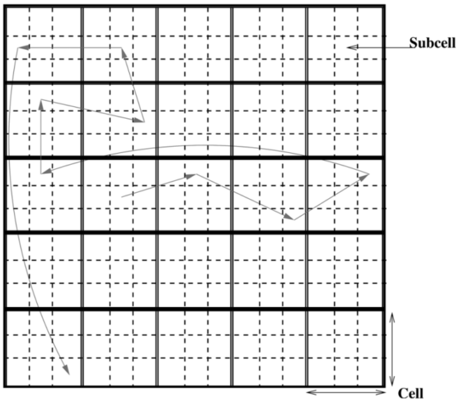

Fig. 1. The division of unit square into cells and subcells; and the motion of a node under a hybrid random walk model.

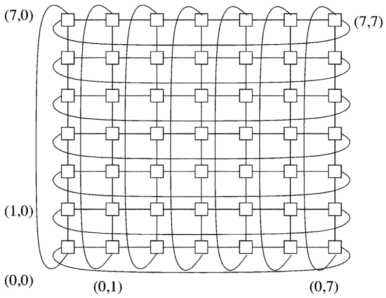

Fig. 2. A numbering of cells on a 2-D torus.

Next, we describe the discrete random direction models, that we study in this paper. As discussed before, these models approximate the motion of the nodes under the random direction models, and are simpler to analyze.

D. Discrete Random Direction Models

These models, like the random direction models, are parameterized by a single parameter α, which takes values between 0 and 1/2. The unit square S is divided into $n^{2\alpha}$ squares of area $1/n^{2\alpha}$ each (henceforth referred to as cells), resulting in a discrete torus of size $n^{\alpha} \times n^{\alpha}$.

Time is divided into slots of equal duration. At the beginning of a slot, each node jumps from its current cell to an adjacent cell, chosen uniformly from within the set of adjacent cells. The motion of a node during the slot is as follows: The node chooses start point and an end point uniformly from within the current cell. During the slot, the node moves from the start point to the end point. In order to keep the duration of all slots the same, the speed of the node is set in proportion to the distance between the start point and end point.

In view of Remark 1, the duration of a slot should be $\Theta\left(n^{1/2-\alpha}\right)$. In order to be able to compare these models with the Brownian motion model of [18], [21], we consider $\sigma_n^2 = 1/n$, where σ_n^2 is the variance parameter of the Brownian motion model, as defined in [18], [21]. Note that with the above choice of σ_n^2 and v_n, each node moves an average distance of $\Theta(1/\sqrt{n})$ in unit time, under all these models. Also observe that under these settings, the discrete random direction model

[4] Throughout this paper we ignore the issues pertaining to $n^{1-2\beta}$, $n^{2\beta}$ not being perfect squares.

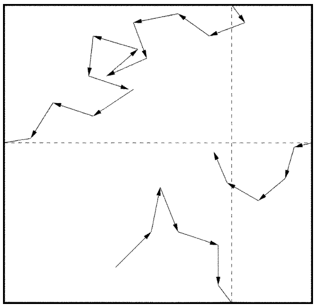

Fig. 3. An example motion path of a node under the random direction model.

with $\alpha = 1/2$ degenerates into the random walk model, which is the discrete time version of the Brownian motion model.

III. Preliminaries: Critical Delay, 2-Hop Delay, Scheduling Schemes

In this section, we build the platform required for the study of critical delay in subsequent sections. We first start by providing a rigorous definition of the critical delay and motivating its study.

It is now well known that the per-node throughput capacity of static ad hoc networks scales as $O(1/\sqrt{n})$, with the capacity achieving scheme being the famous multi-hop relaying scheme. Grossglauser and Tse [9] showed that a constant throughput per-node can be achieved under any stationary and ergodic mobility model which preserves the uniform distribution of nodes at all times. The delay performance of the capacity achieving 2-hop relaying scheme of [9] has been studied in many recent works [14], [16], [18]–[20], [22], and the 2-hop relaying scheme has been shown to incur an average delay of about $\Theta(n)$ under many different mobility models. Alternative schemes that provide a throughput in between that of multi-hop relaying and 2-hop relaying have also been studied in the literature, under various mobility models. The objective of these schemes is to provide a better delay performance than the 2-hop relaying scheme at the cost of some sacrifice in the throughput capacity. An important quantity to study in this respect is what we call the critical delay, which we next define:

Definition 1: Let $\mathcal{C}$ be the class of scheduling and relaying schemes under consideration. For $c \in \mathcal{C}$, let D_c, λ_c be the average delay and per-node throughput, respectively, under scheme c. The critical delay for the class of schemes $\mathcal{C}$, denoted by $D_{\mathcal{C}}$, is the minimum average delay that must be tolerated under a given mobility model in order to achieve a per-node capacity of $\omega(1/\sqrt{n})$, that is,

$$D_{\mathcal{C}} = \inf_{\left\{c\in\mathcal{C}:\lambda_c=\omega\left(1/\sqrt{n}\right)\right\}} D_c. \tag{2}$$

Remark 2: The requirement of $\omega(1/\sqrt{n})$ capacity in the above definition is to ensure that the asymptotic throughput is above that of static ad hoc networks.

Remark 3: Each scheme c in the class of schemes $\mathcal{C}$ is actually sequence of schemes c_n, where c_n is the scheme used with n nodes in the network. Similarly, λ_c and D_c are also sequences rather than numbers. The infimum in (2) should therefore be interpreted as a sequence of infimums, one for each n.

Remark 4: From the above definition, it is clear that the larger class $\mathcal{C}$ is, the smaller will be the critical delay for $\mathcal{C}$. Ideally, one would like the class $\mathcal{C}$ to include all possible scheduling schemes. In that case, if the delay that can be tolerated is smaller than the critical delay for a given mobility model, then one cannot exploit the node motion under that mobility model to increase the throughput capacity (in order sense) beyond its value under static conditions.

Henceforth, we will denote by *2-hop delay*, the delay under the 2-hop relaying scheme. Observe that the delay–capacity trade-off exists for delay values that are greater than the critical delay, and smaller than the 2-hop delay. Next, we introduce two key notions: *first hitting time* and the *first exit time*, which will be used for studying the 2-hop delay and critical delay, respectively.

We start with some notation. Let $B(x,r)$ denote the set of points y within S such that $d(x,y) \leq r$. Let X_i^t denote the position of node i at time t, under some arbitrary mobility model. Note that the time index t could either be continuous or discrete, depending on the mobility model. We are now ready to define the notion of first exit time:

Definition 2 (First Exit Time): Let $X_i^0 = x$. The first exit time of $B(x,r)$, denoted by τ_E^r, is given by

$$\tau_E^r = \inf\left\{t \geq 0 : X_i^t \notin B(x,r)\right\}.$$

It should be clear from the above definition that the statistical properties of the first exit time do not depend on the choice of x or node i. The notion of first exit time is well studied in the mathematics literature, under a variety of contexts. Our interest in notion of exit time stems from the fact that it has a close connection with the critical delay, which will be exploited in the subsequent sections.

We will next define the notion of first hitting time. In our case, we find it more convenient to define it in discrete time. In order to do so, we need some more notation. Let $X(t)$ be a Markov chain taking values in S_X, and with a stationary distribution of Π_X. We have the following definition:

Definition 3 (First Hitting Time): The first hitting time for the set of states $A \subseteq S_X$ is given by

$$\tau_H^A = \inf\{t \geq 0 : X(t) \in A\}$$

with $X(0)$ being distributed according to Π_X.

Again, we would like to point out that the notion of first hitting time has been widely studied in the mathematics literature, under many different contexts. As will be discussed in subsequent sections, the first hitting time has a close connection with the 2-hop delay as well as the critical delay. This connection has been exploited in several recent works for estimating the 2-hop delay under various mobility models (see, for example, [19], [21]).

Next, we recall the result concerning the first hitting time for a single state in case of a 2-D torus of size $\sqrt{n} \times \sqrt{n}$ (for a proof see, for example, [21]).

Lemma 1: Let H denote the first hitting time for a single state on a 2-D torus of size $\sqrt{n} \times \sqrt{n}$, then $\mathbb{E}\{H\} = \Theta(n \log n)$.

The above result will be used quite often in analysis in subsequent sections.

We now define the class of scheduling schemes that we study in this paper. From the above discussion, it is clear that for the purpose of estimating the critical delay, one should allow for as many scheduling schemes as possible. We limit our study to scheduling schemes that satisfy the following assumption:

Assumption A:

- Only the source node can initiate a replication; i.e., the relay nodes holding a packet do *not* initiate a replication.

We now elaborate on the notion of replication. By *replication* we mean packet duplication; i.e., creating redundant copies of a packet. This is different from multi-hop relaying, where a single copy of each packet exists in the network. It is important to note that the notions of *replication* and *relaying* are different, even though both involve forwarding packets to other relay nodes. To understand this, suppose node i decides to *replicate* a packet at node j; then node i can either transmit the packet directly to node j; or use multi-hop relaying, where the packet reaches node j through multiple intermediate nodes. Note that, if asked by node i, the intermediate nodes may also keep a copy of the packet with them, and in that case, all of them are said to receive the packet due to the *same* replication decision initiated by node i. In this example, although both node i and the other intermediate nodes forward the packet to other nodes, their roles are different. Node i is the one that *initiates* the replication, while the intermediate nodes *passively follow* the instructions of node i. Thus, we see that Assumption A only prohibits relay nodes to *initiate* replication and, in particular, multi-hop relaying is allowed under Assumption A.

We also allow *immediate capture* of the destination node using multi-hop relaying or long range wireless broadcast, at any time during the replication process. Note that by capture of the destination node we mean successful delivery of a packet to the destination node. Although we allow for other less intuitive alternatives, in a typical scheduling scheme a successful *capture* usually occurs when a relay node holding the packet comes within a small area around the destination node, so that fewer resources are needed to forward the packet to the destination node. For example, a relay node could enter a disk of a certain radius around the destination node, or a relay node could enter the same cell as the destination node. We call such an area the *capture neighborhood*. The purpose of *replication* is to reduce the time before a successful *capture* occurs (since with more nodes holding the packet, the likelihood of one of them capturing the destination node sooner is higher).

We use the word "immediate" to emphasize that capture can be carried out (using multi-hop relaying or long range wireless broadcast, if required) at a much faster time-scale than the node mobility, and the same is true of the replication process as well. The reason for this is that the packet transmission is usually carried out at a much faster time scale than the node mobility, and as a result the change in the position of a node during a packet transmission is often negligible.

Note that almost all scheduling schemes studied in the literature satisfy Assumption A [9], [14], [16], [17]–[20], [22]. The reason for this, we believe, is that in a distributed system, where nodes make replication decisions and capture decisions without any knowledge of the decisions at the other nodes, restricting the replication decisions to the source node is a natural way to *control the number of copies of a packet*. Note that higher redundancy implies smaller throughput. The source node of a packet is in the best position to control both the total number of replications of the packet and the number of relay nodes getting a copy of the packet with each replication. If the relay nodes were allowed to initiate a replication, then additional cooperation among the relay nodes would likely be required in order to limit the number of replicas of a packet. An interesting example of this would be the scheme of Bansal and Liu [8], where the relay nodes know the location of the static destination node, and also have some knowledge of the future direction of other relay nodes' movement, based on which they can *cooperate* to make selective and more efficient replication toward the destination node. Whether such a scheme can be devised for an ad hoc network with mobile source-destination pairs is still an open research challenge.

IV. Critical Delay and 2-Hop Delay Under Hybrid Random Walk Models

In this section, we study the critical delay and 2-hop delay under hybrid random walk models. We first study the critical delay.

A. Critical Delay

Recall that the critical delay is the minimum delay that must be tolerated in order to achieve a throughput of $\omega(1/\sqrt{n})$. Next, we develop lower bounds on the critical delay for hybrid random walk models. Observe that for $\beta = 0$ (which corresponds to the i.i.d. mobility model) we have a trivial lower bound of 1 on the critical delay. Further, it has been shown in [20] that a throughput of roughly $\Theta\left(1/n^{1/3}\right) = \omega(1/\sqrt{n})$ can be achieved with a constant average delay under the i.i.d. mobility model, which implies that 1 is also an asymptotically tight bound (in order terms).

In the sequel, assume therefore that $\beta > 0$. The main idea is to show that if the average delay is below a certain value, then the packets travel an average distance of $\Theta(1)$ using wireless transmissions in order to reach their respective destination nodes; and then to show that under the protocol model, this results in a throughput of $O(1/\sqrt{n})$.

Recall the mobility model described in Section II-B. Let $\tau^{r}_{E,\beta}$ be the first exit time in case of a hybrid random walk model with parameter β. We start by establishing the following lower bound on $\tau^{1/8}_{E,\beta}$, that holds with probability approaching 1 as $n \to \infty$. The proof is available in Appendix.

Lemma 2: For $0 < \beta \leq 1/2$, we have

$$\mathbf{P}\left(\tau^{1/4}_{E,\beta} \leq \frac{n^{2\beta}}{1024 \log n}\right) \leq \frac{4}{n^2}.$$

We are now ready to show that if the average delay is smaller than $\frac{f_o n^{2\beta}}{2048 \log n}$, where $f_o = 1/2400$, then, on average, the packets must be relayed over a distance of $\Theta(1)$.

Lemma 3: Suppose nodes move in accordance with the hybrid random walk model with parameter β, for some $\beta > 0$, and the average delay is smaller than $\frac{f_o n^{2\beta}}{2048 \log n}$, then there exists $N_o < \infty$ such that for all $n \geq N_o$, the packets are relayed over an average distance no smaller than $f_o/10$.

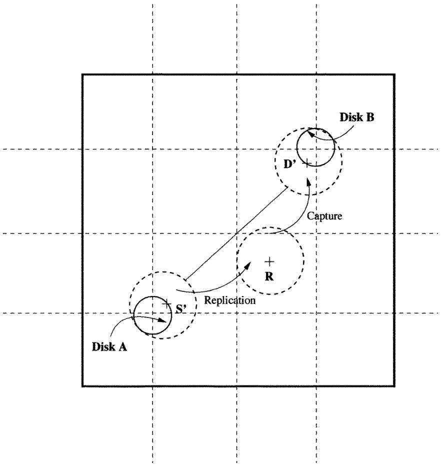

Fig. 4. Disks A and B of radius 1/16 centered at $(-1/4, -1/4)$ and $(1/4, 1/4)$, respectively, that are used in the proof of Lemma 3.

Proof: Consider disks A and B of radius 1/16, as shown in Fig. 4. Note that according to our assumption in Section II-A, the arrival process at each node is independent of the node mobility process. For a given number of nodes n, let f_n denote the fraction of packets having their source nodes inside disk A and destination nodes inside disk B, at the time of arrival. Since, under our setting, the node distribution is uniform in the subcells at all times, we have that $f_n \to \left(\pi(\frac{1}{16})^2\right)^2$ as $n \to \infty$. Thus, for all $n \geq N_o$, where N_o is chosen appropriately, we have $f_n \geq f_o$. Now, since the average delay is smaller than $\frac{f_o n^{2\beta}}{2048 \log n}$, the delay of at least of one-half of such packets must be at most

$$\frac{2}{f_n} \cdot \frac{f_o n^{2\beta}}{2048 \log n} \leq \frac{n^{2\beta}}{1024 \log n}$$

for all $n \geq N_o$. Consider one such packet; let its time of arrival be t. Using Lemma 2 and the union bound, we have that with a probability of at least

$$1 - n \cdot \frac{4}{n^2} = 1 - \frac{4}{n}$$

none of the n nodes in the network will exit the disk of radius 1/8 centered around its initial position at time t before time $t + \frac{(n^{2\beta}}{1024 \log n)}$. Let S' and D' be the positions of the source node and destination node, respectively, for which the distance between them at time t is minimized (see Fig. 4). Let the destination node be captured by the relay node r, when the latter is placed at the point R (see Fig. 4). Assuming that the total time spent in the replication (which resulted in the relay node r getting a copy of the packet) and capture is $o\left(n^{2\beta}/\log n\right)$, the contribution of node mobility during the replication and capture phase in carrying the packet toward its destination node is $o(1)$, and can be ignored. A simple geometrical argument shows that in order to reach the destination node, the packet must be relayed over a distance of at least

$$\frac{1}{\sqrt{2}} - 2 \cdot \frac{1}{16} - 2 \cdot \frac{1}{8} - \frac{1}{8} = d_0 > \frac{1}{5}$$

where the term $\frac{1}{\sqrt{2}} - 2 \cdot \frac{1}{16}$ corresponds to the minimum possible distance between the source node and the destination node at time t, that is, the distance between points S' and D'; the term $2 \cdot \frac{1}{8}$ corresponds to the maximum distance the source node and destination node can possibly travel toward each other between time $\left[t, t + \frac{n^{2\beta}}{1024 \log n}\right]$; and the term $\frac{1}{8}$ corresponds to the maximum possible distance the relay node can travel in the direction of the destination node, after receiving the packet and before time $t + \frac{n^{2\beta}}{1024 \log n}$.

By choosing N_o large enough, we can ensure that $d_0(1 - 1/n) \geq 1/5$ for $n \geq N_o$, and the average distance that the packets must be relayed over in that case would be no smaller than

$$\frac{f_o}{2} \cdot \frac{1}{5} = \frac{f_o}{10}$$

proving the Lemma. ■

Remark 5: Note that in the above proof we assumed that the total time spent in the replication and capture phase is $o\left(n^{2\beta}/\log n\right)$. This assumption is motivated by the fact that in most scheduling schemes the replication and capture are performed using a wireless broadcast or multi-hop relaying, and are therefore carried out at a much faster time-scale than the node mobility. Note also that a packet might be relayed over several hops during either the replication or capture phase, and for technical consistency, one might then need to scale down the packet size in order to keep the total time spent in replication and capture phases small enough. This kind of packet size scaling has been used quite often in the literature (see, for example, [16], [17], [20], [22]). We now show that if the packets are, on average, relayed over a distance of $\Theta(1)$, then the throughput must be $O(1/\sqrt{n})$:

Lemma 4: Suppose that there exists a constant $c > 0$, independent of n, such that on average the packets are relayed over a total distance no less than c, then $\lambda(n) = O(1/\sqrt{n})$.

Proof: Consider a large enough time interval $\mathcal{T}$. The total number of packets communicated end-to-end between all source-destination pairs during the interval is then $c_P \lambda n \mathcal{T}$, where $1/c_P$ is the number of bits per packet; and the total distance traveled by these packets is no smaller than $c_P \lambda n \mathcal{T} c$. Let h_p be the number of times packet p is relayed, and let l_p^h, for $h = 1, \ldots, h_p$, denote the transmission range for the hth relaying. Observe that we have

$$\sum_{p=1}^{c_P \lambda n \mathcal{T}} \sum_{h=1}^{h_p} l_p^h \geq c_P \lambda n \mathcal{T} c. \tag{3}$$

Let X^i denote the position of node i, for $i = 1, \ldots, n$. Consider nodes i, j transmitting directly to nodes k and l, respectively, at time t. Then, in order for the transmissions to be successful under the protocol model of interference, the following inequalities must hold at the time of transmission:

$$d(X^j, X^k) \geq (1 + \Delta) d(X^i, X^k)$$
$$d(X^i, X^l) \geq (1 + \Delta) d(X^j, X^l).$$

Hence,

$$d(X^j, X^i) \geq d(X^j, X^k) - d(X^i, X^k)$$
$$\geq \Delta d(X^i, X^k).$$

Similarly,

$$d(X^i, X^j) \geq \Delta d(X^j, X^l).$$

Therefore,

$$d(X^i, X^j) \geq \frac{\Delta}{2}\left(d(X^i, X^k) + d(X^j, X^l)\right).$$

That is, disks of radius $\frac{\Delta}{2}$ times the transmission range centered at the transmitter are disjoint from each other.[5] We can therefore measure the radio resources that each transmission consumes by the areas of these disjoint disks. Note that the area of S is 1; for each of these disks, at least 1/4 of it must lie within S; and each relaying of a packet lasts $\frac{1}{c_P W}$ amount of time. Thus,

$$\frac{1}{4}\sum_{p=1}^{c_p \lambda n \mathcal{T}} \sum_{h=1}^{h_p} \pi \left[\frac{\Delta}{2} l_p^h\right]^2 \leq c_P W \mathcal{T}. \tag{4}$$

By Cauchy–Schwarz Inequality,

$$\left[\sum_{p=1}^{c_p \lambda n \mathcal{T}} \sum_{h=1}^{h_p} l_p^h\right]^2 \leq \left[\sum_{p=1}^{c_p \lambda n \mathcal{T}} \sum_{h=1}^{h_p} \left(l_p^h\right)^2\right] \left[\sum_{p=1}^{c_p \lambda n \mathcal{T}} \sum_{h=1}^{h_p} 1\right]. \tag{5}$$

Further, since there are at most n simultaneous transmissions at any given time in the network, we have

$$\sum_{p=1}^{c_p \lambda n \mathcal{T}} h_p \leq c_P W \mathcal{T} n. \tag{6}$$

Using (3)–(6), we have

$$\frac{16 c_P W \mathcal{T}}{\pi \Delta^2} \geq \sum_{p=1}^{c_p \lambda n \mathcal{T}} \sum_{h=1}^{h_p} \left(l_p^h\right)^2 \geq \frac{\left[\sum_{p=1}^{c_p \lambda n \mathcal{T}} \sum_{h=1}^{h_p} l_p^h\right]^2}{\left[\sum_{p=1}^{c_p \lambda n \mathcal{T}} h_p\right]}$$

$$\geq \frac{(c_P \lambda n \mathcal{T} c)^2}{c_P W \mathcal{T} n}.$$

Hence,

$$\lambda \leq \sqrt{\frac{16 W^2}{\pi \Delta^2 c^2}} \frac{1}{\sqrt{n}}.$$

■

Remark 6: Following the line of analysis used for proving Theorem 4.2 in [5], it is possible to establish the above result under the generalized physical model (see [5]), which is more realistic than the protocol model considered in this paper.

The following result is an easy consequence of Lemmas 3 and 4, and the definition of critical delay.

Proposition 1: Under the class of scheduling schemes satisfying Assumption A, the critical delay for the hybrid random walk model with parameter $\beta > 0$ scales as $\Omega\left(n^{2\beta}/\log n\right)$.

We now derive an upper bound on the critical delay. Note that the delay under a scheme that can provide a throughput of $\omega(1/\sqrt{n})$ is an upper bound on the critical delay. For $\beta = 1/2$ (which corresponds to the random walk mobility model) it is claimed in [16] that a simple modification of the 2-hop relaying scheme of Grossglauser and Tse can provide a throughput capacity of $\Theta(1)$, incurring a delay of $\Theta(n \log n)$. This result immediately establishes an upper bound of $\Theta(n \log n)$ on the critical delay for $\beta = 1/2$. In what follows, we consider hybrid random walk model with parameter β, for $0 < \beta < 1/2$, and show that the critical delay for the model scales as $O\left(n^{2\beta} \log n\right)$.

The idea is to develop a scheduling and relaying scheme that can provide a throughput of $\omega(1/\sqrt{n})$, while incurring a delay of $O(n^{2\beta} \log n)$. In this paper, we only provide the main insight behind such a scheme, leaving out the detailed analysis for our future work.

Consider a scheme in which each packet is replicated at a single relay node, which delivers the packet to the destination node, once it is in the same cell as the destination node, possibly using multi-hop transmission. Note that such a scheme would require an appropriate scaling of the packet size to ensure that the packet can be delivered to its destination node within one slot, i.e., before the relay node and destination node can possibly move into different cells. We now provide an approximate analysis of the delay and throughput under such a scheme. In order to keep our discussion simple and insightful, we will ignore possible delays due to queuing at the source node or the relay nodes. More precisely, we will assume that the delay of the packet is the time it takes for the relay node to move into the same cell as the destination node, starting from the time it receives the packet from the source node.

Now, consider a packet arrival at the source node. Note that since the packet arrival process at each node is independent of the node mobility process, the source node and destination node are equally likely to be in any of the cells, at the time of the packet arrival. Thus, the expected delay of the packet is of the same order as the expected first hitting time of a single state, in case of a random walk on a 2-D torus of size $n^\beta \times n^\beta$, which, by Lemma 1, is $\Theta\left(n^{2\beta} \log n\right)$.

Now in order to the estimate the throughput, we need to account for the following factors: (i) the loss in throughput due to multiple relaying of the same packet, (ii) the loss in throughput due to the interference. Using the standard multi-hop scheme, where each hop carries the packet over $\Theta\left(\sqrt{\log n/n}\right)$ distance,[6] and noting that the packet travels a distance of no more than $\Theta\left(1/n^\beta\right)$ using multi-hop transmissions, it follows that the number of times a packet must be relayed is $O\left(n^{1/2-\beta}/\sqrt{\log n}\right)$. The loss in throughput due to multiple relaying is correspondingly $O\left(n^{1/2-\beta}/\sqrt{\log n}\right)$. Since each transmission is carried over $\Theta\left(\sqrt{\log n/n}\right)$ distance, it follows easily using the protocol model that the nodes which are within $\Theta\left(\sqrt{\log n/n}\right)$ distance of the sender must be kept quiet, resulting in a loss of throughput by a factor of $\Theta(\log n)$. Thus, the throughput of such a scheme would be $\Omega\left(n^{\beta-1/2}/\sqrt{\log n}\right) = \omega(1/\sqrt{n})$ for $\beta > 0$.

The above discussion shows that the critical delay scales as $O\left(n^{2\beta} \log n\right)$ for a hybrid random walk model of parameter β. Although, the above arguments are heuristic, they can easily be

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

[6] Note that this is the minimum possible communication range needed for ensuring almost sure connectivity (see [1]).

made precise. However, in order to do so, one would need to specify the details of the scheduling scheme, which is beyond the scope of this paper.

Remark 7: By choosing the size of capture neighborhood appropriately (e.g., $\Theta(1/\log n)$), one can show that the critical delay is bounded by $\Theta\left(n^{2\beta}\log\log n\right)$.

B. Two-Hop Delay

In this section, we analyze the 2-hop delay under hybrid random walk models. Recall that the original 2-hop relaying protocol of Grossglauser and Tse [9] allows only nearest neighbor transmissions. Subsequent works [14], [16], [18], [19], [22] have considered a slightly different version of the protocol that allow transmissions between nodes that are either in the same cell (of size $\Theta(1/\sqrt{n}) \times \Theta(1/\sqrt{n})$), or within a distance of $\Theta(1/\sqrt{n})$ from each other. Note that these different versions of the protocol are roughly the same, because when n nodes are distributed uniformly within a unit square the average nearest neighbor distance is $\Theta(1/\sqrt{n})$.

Next, we analyze the delay under 2-hop relaying protocol, assuming that the transmissions are scheduled between nodes that are in the same subcell. As in the analysis of critical delay, we will ignore the queuing delays, postponing their analysis to future work. Thus, we would mainly be interested in estimating the time it takes for the relay node and destination node to come within the same subcell, starting from two randomly and uniformly chosen subcells in the network. Let us denote this random time by T. In subsequent analysis, we will say that two nodes are in a "meeting" if they are currently inside the same cell, and will denote the time between successive meetings as the *inter-meeting time*.

Observe that

$$T = \tau_1 n^{2\beta-1} + \cdots + (\tau_1 + \cdots + \tau_i)\left(1 - n^{2\beta-1}\right)^{i-1} n^{2\beta-1} + \cdots \tag{7}$$

where τ_1 is the time required by nodes to enter the same cell, starting from their initial random and uniformly distributed positions, henceforth denoted by *first meeting time*; and τ_i for $i \geq 2$ are the successive *inter-meeting times*. Observe that $n^{2\beta-1}$ is the probability that the nodes choose the same subcell inside a given cell. It is easy to see that the mean first meeting time is of the order of mean first hitting time of a single state, in case of a random walk on a 2-D torus of size $n^\beta \times n^\beta$. Using Lemma 1, it follows that $\mathbb{E}\{\tau_1\} = \Theta\left(n^{2\beta}\log n\right)$. Further, the mean inter-meeting times are of the order of mean *first return time* (see, for example, [26, Chap 2, p. 2]) of a random walk on a 2-D torus of size $n^\beta \times n^\beta$, which is well known to be $n^{2\beta}$. We therefore have $\mathbb{E}\{\tau_i\} = \Theta\left(n^{2\beta}\right)$ for $i \geq 2$. Taking the expectations on both sides of (7), and performing some simple algebraic manipulations, we obtain

$$\begin{aligned}\mathbb{E}\{T\} &= \mathbb{E}\{\tau_1\} + \mathbb{E}\{\tau_2\} n^{1-2\beta} \\ &= \Theta\left(n^{2\beta}\log n\right) + \Theta\left(n^{2\beta}\right) n^{,-2\beta}.\end{aligned}$$

Thus, for $\beta < 1/2$, we have $\mathbb{E}\{T\} = \Theta(n)$; and for $\beta = 1/2$, we have $\mathbb{E}\{T\} = \Theta(n\log n)$.

Remark 8: Note that our results for $\beta = 0$ and $\beta = 1/2$ are in agreement with the corresponding results for the i.i.d. mobility model in [14] and random walk model in [22]. Both these works also account for the queuing delays: In [14], queuing delays at the source nodes as well as relay nodes are considered, whereas, [22] considers queuing delays at the relay nodes only. It is interesting to see that our simplified analysis yields exact results (in order sense) for two extreme choices of β, i.e., $\beta = 0$, 1/2.

Remark 9: Observe that all hybrid random walk models incur roughly $\Theta(n)$ delay under the 2-hop relaying scheme, but their critical delays vary significantly. More precisely, as β increases the critical delay increases as well (roughly as $\Theta\left(n^{2\beta}\right)$), shrinking the delay–capacity trade-off region. The two extreme cases being: (i) the i.i.d. mobility model (i.e., $\beta = 0$), for which a per-node capacity of $\omega(1/\sqrt{n})$ can be achieved even under a constant delay constraint; and (ii) the random walk model (i.e., $\beta = 1/2$), for which the delay on the order of $\Theta(n/\log n)$ or more must be tolerated in order to achieve a per-node capacity of $\omega(1/\sqrt{n})$.

V. Critical Delay and 2-Hop Delay Under Discrete Random Direction Models

In this section, we study the critical delay and 2-hop delay under discrete random direction models. As in the previous section, we first study the critical delay.

A. Critical Delay

Recall that the discrete random direction models are characterized by a single parameter α that takes values between 0 and 1/2. As in the previous section, we will first derive a lower bound on the critical delay by lower bounding the first exit time for a disk of radius 1/8. Let $\tau_{E,\alpha}^{1/8}$ denote the first exit time for such a disk in case of discrete random direction model with parameter α. Let the duration of a slot be $Cn^{1/2-\alpha}$. For $\alpha = 0$, one trivially obtains a lower bound of $\Theta(\sqrt{n})$ on $\tau_{E,\alpha}^{1/8}$. For $\alpha > 0$, we have the following result, the proof of which follows *mutatis mutandis* from the proof of Lemma 2:

Lemma 5: For $0 < \alpha \leq 1/2$, we have

$$\mathbf{P}\left(\tau_{E,\alpha}^{1/8} \leq \frac{Cn^{1/2+\alpha}}{1024\log n}\right) \leq \frac{4}{n^2}.$$

It is interesting to note that a similar result can also be proved for random direction models (see the Appendix):

Lemma 6: Let $T_{E,\alpha}^{1/8}$ denote the first exit time of a disk of radius 1/8 for the random direction model with parameter α. For $0 < \alpha \leq 1/2$, we have

$$\mathbf{P}\left(T_{E,\alpha}^{1/8} \leq \frac{Cn^{1/2+\alpha}}{768\log n}\right) \leq \frac{4}{n^2}.$$

The following Lemma and Proposition can now be proved in a similar fashion to Lemma 3 and Proposition 1, respectively.

Lemma 7: Suppose nodes move in accordance with the discrete random direction model with parameter α, for some $\alpha > 0$, and the average delay of packets under a scheduling scheme is smaller than $\frac{Cf_o n^{1/2+\alpha}}{2048\log n}$, then there exists $N_o < \infty$ such that for all $n \geq N_o$, the packets are, on average, relayed over a distance greater than $f_o/10$.

Proposition 2: Under the class of scheduling schemes satisfying Assumption A, the critical delay for a discrete random direction model with parameter $\alpha > 0$ scales as $\Omega\left(n^{\alpha+1/2}/\log n\right)$.

Remark 10: Analogs of the results in Lemma 7 and Proposition 2 can easily be proved for random direction models using Lemma 6.

Remark 11: Note that for $\alpha = 0$, using the lower bound of $\Theta(\sqrt{n})$ on $\tau_{E,\alpha}^{1/8}$, and arguing as in the proof of Lemma 7, we can easily establish a lower bound of $\Theta(\sqrt{n})$ on the critical delay. Moreover, the same reasoning shows that a lower bound of $\Theta(\sqrt{n})$ on critical delay also holds under the random way-point mobility model. This result was earlier shown in [19], but under a more restricted class of scheduling and relaying schemes than in this paper.

Next, we establish an upper bound on the critical delay. Consider the scheme discussed before in Section IV-A. Recall that each packet is replicated to at most one relay node, which delivers it to its destination node on entering the same cell as the destination node. An approximate analysis of the throughput and delay under such a scheme can be carried out following the line of analysis in Section IV-A, and it is straightforward to show that the delay under such a scheme is $\Theta\left(n^{1/2+\alpha}\log n\right)$, and the throughput is $\Omega\left(n^{\alpha-1/2}/\sqrt{\log n}\right) = \omega(1/\sqrt{n})$ for $\alpha > 0$. Thus, for $\alpha > 0$, the critical delay is bounded above by $\Theta\left(n^{1/2+\alpha}\log n\right)$.

Remark 12: One might think that by increasing the size of capture neighborhood to $1/n^{\gamma}$, where $0 < \gamma < \alpha$, one might be able reduce the delay below $\Theta\left(n^{1/2+\alpha}\log n\right)$, while maintaining a throughput of $\Omega(1/\sqrt{n})$. This is, however, not possible. In fact, it can be shown that with a capture neighborhood of size $r(n)$ the delay becomes $\Theta\left(n^{1/2+\alpha}\log\left(1/r(n)\right)\right)$, and the throughput becomes $\Theta\left(1/r(n)\sqrt{n\log n}\right)$. Thus, choosing a capture neighborhood of size $1/n^{\gamma}$ for any $\gamma > 0$ will not change the order of the delay. Also, by considering a capture neighborhood of size $\Theta(1/\log n)$, one can establish an upper bound of $\Theta\left(n^{1/2+\alpha}\log\log n\right)$ on the critical delay.

Next, we consider $\alpha = 0$. Note that for $\alpha = 0$, the discrete random direction model is similar to the random way-point mobility model, with the difference being that the successive trips (moving between a chosen pair of points) that a node makes under the discrete random direction model are independent; whereas, there is some dependency between successive trips in case of the random way-point mobility model (since the next trip starts from the point where the previous trip ends). The random way-point mobility model has been analyzed[7] in [19]. In particular, a protocol that allows one to trade-off throughput for delay has been developed in [19], and shown to achieve the following delay–capacity trade-off:

$$D(n) = O\left(n/k(n)\log n\right), \text{ and } \lambda(n) = \Omega\left(1/k(n)\log n\right)$$

where $D(n)$ is the average packet delay and the $\lambda(n)$ is the per-node throughput. Following the line of analysis in [19], one can show that the same delay–capacity trade-off can also be achieved under the discrete random direction model with parameter $\alpha = 0$. Now, by choosing $k(n) = \sqrt{n}/a(n)\log n$, where

$$a(n) \to \infty \text{ as } n \to \infty \tag{8}$$

it follows that the critical delay is $O\left(a(n)\sqrt{n}\log^2 n\right)$ for all $a(n)$ satisfying condition (8). In particular, the critical delay is $o(n^{\gamma})$ for any $\gamma > 1/2$.

B. Two-Hop Delay

In this section, we analyze the 2-hop delay under discrete random direction models. We assume that the transmissions are scheduled between nodes that are within a distance of $1/\sqrt{n}$ from each other, and ignore the queuing delays. Thus, we are mainly be interested in estimating the time it takes for the relay node and destination node to come within a distance of $1/\sqrt{n}$ of each other, starting from two randomly and uniformly chosen positions in the network.

Let us denote this random time by T. Arguing as in Section IV-B, it can be shown that $\mathbb{E}\{T\} = \Theta\left(n^{1/2-\alpha}\right)\Theta(\mathbb{E}\{\tau_1\} + \mathbb{E}\{\tau_2\}/p)$, where τ_1 is the first meeting time; τ_2 is the inter-meeting time; and p is the probability that two arbitrary nodes will come within a distance of $1/\sqrt{n}$ of each other any time during a slot, given that they are within the same cell in that slot. Note that the factor of $\Theta\left(n^{1/2-\alpha}\right)$ comes because the duration of each slot is now $\Theta\left(n^{1/2-\alpha}\right)$. As in Section IV-B, we have $\mathbb{E}\{\tau_1\} = \Theta\left(n^{2\alpha}\log n\right)$ and $\mathbb{E}\{\tau_2\} = \Theta\left(n^{2\alpha}\right)$. Furthermore, it is easy to see that $p = \Theta\left(n^{\alpha-1/2}\right)$. Thus,

$$\mathbb{E}\{T\} = \Theta\left(n^{1/2+\alpha}\log n\right) + \Theta(n).$$

Hence, we see that $\mathbb{E}\{T\} = \Theta(n\log n)$ for $\alpha = 1/2$, and $\Theta(n)$ for $0 \le \alpha < 1/2$.

Remark 13: Once again, we note that our results for $\alpha = 0$ and $\alpha = 1/2$ are in agreement with the corresponding results for the random walk model in [22] and the random way-point mobility model in [19]. Both these works also account for the queuing delays: In [19], queuing delays at the source nodes as well as relay nodes are considered, whereas, [22] considers queuing delays at the relay nodes only. Again, we see that our simplified analysis yields exact results (in order sense) for two extreme choices of α, i.e., $\alpha = 0$, 1/2.

Remark 14: Observe that all discrete random direction models incur roughly $\Theta(n)$ delay under 2-hop relaying scheme; however, their critical delays vary significantly. More precisely, as α increases the critical delay increases as well (roughly as $\Theta(n^{\alpha})$), shrinking the delay–capacity trade-off region. The two extreme cases being: (i) $\alpha = 0$ (random way-point mobility model), for which a per-node capacity of $\omega(1/\sqrt{n})$ can be achieved incurring delays of about $\Theta(\sqrt{n})$; and (ii) $\alpha = 1/2$ (random walk model), for which a delay of $\Theta(n/\log n)$ or more must be tolerated in order to achieve a per-node capacity of $\omega(1/\sqrt{n})$.

[7] Although [19] considers a slightly different version of the random way-point mobility model on a sphere, the results in [19] can easily be extended to a 2-D torus.

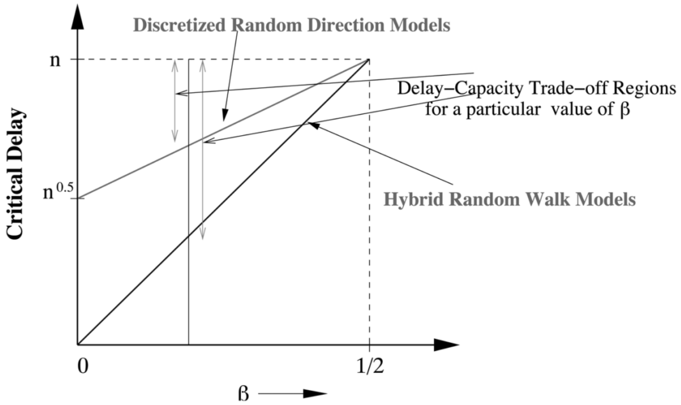

Fig. 5. The scaling of critical delay in case of hybrid random walk and discrete random direction models.

VI. Discussion

The main contribution of this paper is the definition and study of the notion of critical delay, which provides us with a platform to compare and contrast several existing mobility models. The notion of critical delay is important as it provides us with a way of determining whether a particular form of node mobility can be exploited to improve the throughput capacity under a given delay constraint. We also showed that there exists a strong connection between the notion of exit time and critical delay, and used this connection to estimate the critical delay under various mobility models.

The results obtained in the previous sections are summarized in Fig. 5. Clearly, the mobility models considered in the literature are in some sense extreme: they either exhibit the smallest critical delays or the largest critical delays among all mobility models having roughly the same 2-hop delay. Thus, on one extreme, there is almost no delay–capacity trade-off under the Brownian motion model and random walk model, and, on the other extreme, there is a smooth delay–capacity trade-off for a wide range of delays under the random way-point mobility model and i.i.d. mobility model.

An interesting insight provided by our results is that the critical delay is inversely proportional to the *characteristic path length*. By *characteristic path length*, we mean the distance that a node travels without changing direction. (Recall that in case of (discrete) random direction model with parameter α, the *characteristic path length* is of the order of $n^{-\alpha}$ and the critical delay is roughly of the order of $n^{1/2+\alpha}$.) Thus, in terms of application support, a scenario where the nodes move over long distances without changing directions (as in the random way-point mobility model) is more desirable than a scenario where nodes change directions over short distances (as in the Brownian motion model). This is because the former scenario provides more flexibility in terms of choosing the point of operation on the delay–capacity trade-off curve, and can therefore support a wider range of applications.

In a real world scenario, it is rather unlikely that the (density) number of nodes in the network will have a strong influence on the motion of nodes.[8] We therefore believe that a mobility model like the random way-point model might be more appropriate for determining the scaling laws for large mobile ad hoc networks, rather than a mobility model like the Brownian motion model (random walk model). We therefore expect that future mobile ad hoc networks would provide network designers with ample flexibility in terms of choosing the desired operational point on the delay–capacity trade-off curve, and this opportunity must be fully exploited for optimal operation of such networks, possibly using a cross-layer design approach.

VII. Conclusion

We have studied the delay–capacity trade-offs in mobile ad hoc networks. We introduced the meaningful notion of critical delay to systematically study how much delay must be tolerated for a given form of node mobility to result in an improvement of the network capacity. The notion of critical delay allowed us to look at various forms of node mobility studied in the literature from a common perspective, and to compare and contrast them.

We proposed two different classes of mobility models and showed that they both exhibit critical delays that are in-between that of the mobility models studied in the literature, thus showing that the mobility models considered in the literature are rather extreme. More importantly, we showed that the critical delay is inversely proportional to the characteristic path length, which is the distance nodes travel without changing direction. These results, among other things, provide a clear understanding of why is it that the critical delay under Brownian motion model is larger than the critical delay under random way-point mobility model.

One would expect that the density of nodes should have little, if any, influence on the motion of nodes in a practical setting. Thus the characteristic path length should have a rather weak dependence on the number of nodes or the node density. One would therefore expect the critical delay in a practical scenario to be close to $\Theta(\sqrt{n})$, as in the case of the random way-point mobility model. This result is optimistic, since it suggests that the future mobile ad hoc networks would provide network designers with ample flexibility in terms of choosing the desired operational point on the delay–capacity trade-off curve; an opportunity that must be fully exploited for optimal operation of such networks, possibly using a cross-layer design approach.

Appendix

In this appendix, we provide proofs for Lemmas 2 and 6. We start with the following simple result that is a version of Hoeffding's Inequality (see, for example, [27, ch. 3, p. 120]).

Lemma 8: Let $X_1, X_2, \dots, X_n$ be i.i.d. random variables taking values in $[-l, l]$ for $0 < l < \infty$, and suppose $\mathbb{E}\{X_i\} = 0$ for all i. Let $S_n = \sum_{i=1}^{n} X_i$ and $\sigma_{S_n}^2$ be the variance of S_n. Then

$$\mathbf{P}(S_n \geq \mu\sigma_{S_n}) \leq e^{-\mu^2/4}$$

[8] This is likely to be the case for moderate densities of nodes. At higher densities, crowding of the nodes might occur which can restrict their motion, and therefore density of the nodes can influence node motion.

for all $0 \leq \mu \leq 2\sigma_{S_n}/l$.

We are now ready to prove Lemmas 2 and 6.

Proof of Lemma 2: Recall the hybrid random walk model of Section II-B, and the definition of $\tau_{E,\beta}^{1/8}$, given in Section III. As discussed in Section III, the statistical properties of the first exit time do not depend on the choice of y. So let y be the origin, that is, the point $(0,0)$. Let (x_0, y_0) be the cell containing the origin. Also, let (x_t, y_t) be the cell in which node i lies at time t. Further, let

$$\tau_x^+ \triangleq \inf\left\{t \geq 0 : (x_t - x_0) \geq \frac{n^\beta}{16}\right\};$$

$$\tau_x^- \triangleq \inf\left\{t \geq 0 : (x_t - x_0) \leq -\frac{n^\beta}{16}\right\};$$

and τ_y^+, τ_y^- be similarly defined with y_t, y_0 in place of x_t and x_0, respectively. Observe that

$$\mathbf{P}\left(\tau_{E,\beta}^{1/8} \leq m\right) \leq \mathbf{P}\Big(\tau_x^+ \leq m \text{ or } \tau_x^- \leq m \text{ or } \tau_y^+ \leq m \text{ or } \tau_y^- \leq m\Big)$$

for $m \geq 0$. Using the union bound and appealing to the symmetry of node motion, we obtain

$$\mathbf{P}\left(\tau_{E,\beta}^{1/8} < m\right) \leq 4\mathbf{P}\left(\tau_x^+ < m\right).$$

Now, observe that before time $\tau_{E,\beta}^{1/8}$, x_t has the following form:

$$x_t = x_0 + \sum_{i=1}^{t} s_i$$

where s_i are i.i.d. random variables taking values in $\{-1, 0, 1\}$ with probabilities $\{1/4, 1/2, 1/4\}$, respectively. Although, x_t is not a simple random walk, it is clear due to its symmetry that the reflection principle for 1-D random walk holds in case of x_t as well, and we have

$$\begin{aligned}\mathbf{P}\left(\tau_x^+ \leq k\right) &= 2\mathbf{P}\left(x_{\lfloor k \rfloor} - x_0 > \frac{n^\beta}{16}\right) \\ &\quad + \mathbf{P}\left(x_{\lfloor k \rfloor} - x_0 = \frac{n^\beta}{16}\right) \\ &\leq 2\mathbf{P}\left(x_{\lfloor k \rfloor} - x_0 \geq \frac{n^\beta}{16}\right) \qquad (9)\end{aligned}$$

for $k \geq 0$, where $\lfloor\cdot\rfloor$ denotes the greatest integer function. Since each s_i has mean 0 and variance 1/2, a straightforward application of Lemma 8 gives

$$\mathbf{P}\left(x_t - x_0 \geq \frac{n^\beta}{16}\right) \leq e^{-n^{2\beta}/512t} \qquad (10)$$

for $t \geq n^\beta/16$. Substituting $k = \frac{n^{2\beta}}{1024 \log n}$ in (9), and combining with (10), we obtain

$$\mathbf{P}\left(\tau_x^+ \leq \frac{n^{2\beta}}{1024 \log n}\right) \leq e^{-2\log n} = \frac{1}{n^2}$$

and the result follows by noting that $\mathbf{P}\left(\tau_{E,\beta}^{1/8} \leq m\right) \leq 4\mathbf{P}\left(\tau_x^+ \leq m\right)$ for $m \geq 0$.

Proof of Lemma 6: Recall the random direction model of Section II-B, and the definition of $\tau_{E,\alpha}^{1/8}$, given in Section III. Arguing as in the previous proof, it suffices to consider $y = (0,0)$. Let (x_t, y_t) be the position of node i after t trips. Let

$$\tau_x \triangleq \inf\left\{t \geq 0 : |x_t| \geq 1/8\sqrt{2}\right\}$$

and

$$\tau_y \triangleq \inf\left\{t \geq 0 : |y_t| \geq 1/8\sqrt{2}\right\}.$$

It is then clear that

$$\mathbf{P}\left(\tau_{E,\alpha}^{1/8} \leq m\right) \leq \mathbf{P}(\tau_x \leq m \text{ or } \tau_y \leq m).$$

Appealing to the symmetry of the node motion and using the union bound, we obtain

$$\mathbf{P}\left(\tau_{E,\alpha}^{1/8} \leq m\right) \leq 2\mathbf{P}(\tau_x \leq m).$$

Let s_k be the x-coordinate of the nodes' position immediately after completing the k^{th} trip. Before time $\tau_{E,\alpha}^{1/8}$, s_k has the simple form:

$$s_k = \sum_{i=1}^{k} z_i$$

where z_i are i.i.d. random variables taking values in $[-n^{-\alpha}, n^{-\alpha}]$. Note also that each z_i has mean zero and variance $n^{-2\alpha}/2$. Using Lemma 8, we have

$$\mathbf{P}\left(s_k \geq 1/8\sqrt{2}\right) \leq e^{-\frac{n^{2\alpha}}{256k}}$$

for $k \geq n^\alpha/8\sqrt{2}$. Using the symmetry of the node motion once again, we have

$$\mathbf{P}\left(|s_k| \geq 1/8\sqrt{2}\right) \leq 2e^{-\frac{n^{2\alpha}}{256k}}.$$

Noting that the duration of each trip is $Cn^{1/2-\alpha}$, it follows that

$$\begin{aligned}\mathbf{P}\left(\tau_x \leq k \cdot Cn^{1/2-\alpha}\right) &= \mathbf{P}\left(\cup_{i=1}^{k} |s_i| \geq 1/8\sqrt{2}\right) \\ &\leq \sum_{i=\lceil n^\alpha/8\sqrt{2}\rceil}^{k} 2e^{-\frac{n^{2\alpha}}{256i}} \\ &\leq k \cdot 2e^{-\frac{n^{2\alpha}}{256k}}\end{aligned}$$

where $\lceil n^{2\alpha}/8\sqrt{2}\rceil$ denotes the smallest integer greater than $n^{2\alpha}/8\sqrt{2}$. Since $n^{2\alpha} \leq n$ for $\alpha \leq 1/2$, we have

$$\mathbf{P}\left(\tau_x \leq \frac{Cn^{2\alpha}}{768 \log n} \cdot n^{1/2-\alpha}\right) \leq n \cdot 2e^{-3\log n} = \frac{2}{n^2}.$$

Thus

$$\mathbf{P}\left(\tau_{E,\alpha}^{1/8} \leq \frac{Cn^{\alpha+1/2}}{768 \log n}\right) \leq \frac{4}{n^2}$$

as claimed.

Acknowledgment

The authors wish to thank Xiaojun Lin for many valuable discussions throughout the course of this work.

References

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

[2] F. Xue, L. L. Xie, and P. R. Kumar, "The transport capacity of wireless networks over fading channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 834–847, Mar. 2005.

[3] P. Gupta and P. R. Kumar, "Towards an information theory of large networks: An achievable rate region," *IEEE Trans. Inf. Theory*, vol. 49, no. 8, pp. 1877–1894, Aug. 2003.

[4] M. Franceschetti, O. Dousse, D. Tse, and P. Thiran, "On the throughput capacity of random wireless networks," Dept. Elect. Comput. Eng., Univ. California, San Diego, 2005 [Online]. Available: http://fleece.ucsd.edu/~massimo/

[5] A. Agarwal and P. R. Kumar, "Capacity bounds for ad-hoc and hybrid wireless networks," *ACM SIGCOMM Comput. Commun. Rev., Special Issue on Science of Networking Design*, vol. 34, no. 3, pp. 71–81, Jul. 2004.

[6] S. R. Kulkarni and P. Viswanath, "Throughput scaling for heterogeneous networks," in *Proc. IEEE ISIT*, 2003, p. 452.

[7] B. Liu, Z. Liu, and D. Towsley, "On the capacity of hybrid wireless networks," in *Proc. IEEE INFOCOM*, 2003, pp. 1543–1552.

[8] N. Bansal and Z. Liu, "Capacity, delay and mobility in wireless ad-hoc networks," in *Proc. IEEE INFOCOM*, 2003, pp. 1553–1563.

[9] M. Grossglauser and D. N. C. Tse, "Mobility increases the capacity of ad-hoc wireless networks," in *Proc. IEEE INFOCOM*, 2001, pp. 1360–1369.

[10] C.-K. Toh and B. Akyol, "A survey of handover techniques in wireless ATM networks," *Wireless Networks, Special Issue on Wireless ATM*, vol. 5, no. 1, 1998.

[11] F. Bai, N. Sadagopan, and A. Helmy, "IMPORTANT: A framework to systematically analyze the impact of mobility on performance of routing protocols for adhoc networks," in *Proc. IEEE INFOCOM*, 2003, pp. 825–835.

[12] E. Perevalov and R. Blum, "Delay limited capacity of ad hoc networks: Asymptotically optimal transmission and relaying strategy," in *Proc. IEEE INFOCOM*, 2003, pp. 1575–1582.

[13] A. Tsirigos and Z. J. Haas, "Multipath routing in the presence of frequent topological changes," *IEEE Commun. Mag.*, vol. 39, no. 11, pp. 132–138, Nov. 2001.

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

[15] M. J. Neely, "Dynamic power allocation and routing for satellite and wireless networks with time varying channels," Ph.D. dissertation, Massachusetts Inst. Technol., LIDS, Cambridge, MA, 2003.

[16] A. E. Gamal, J. Mammen, B. Prabhakar, and D. Shah, "Throughput-delay trade-off in wireless networks," in *Proc. IEEE INFOCOM*, 2004, pp. 464–475.

[17] S. Toumpis and A. Goldsmith, "Large wireless networks under fading, mobility, and delay constraints," in *Proc. IEEE INFOCOM*, 2004, pp. 609–619.

[18] G. Sharma and R. Mazumdar, "Scaling laws for capacity and delay in wireless ad hoc networks with random mobility," in *Proc. IEEE ICC*, 2004, pp. 3869–3873.

[19] G. Sharma and R. Mazumdar, "Delay and capacity trade-off in wireless ad hoc networks with random way-point mobility," Dept. Elect. Comput. Eng., Purdue Univ., West Lafayette, IN, 2005 [Online]. Available: http://ece.purdue.edu/~gsharma/

[20] X. Lin and N. B. Shroff, "The fundamental capacity-delay tradeoff in large mobile ad hoc networks," presented at the Third Annual Mediterranean Ad Hoc Networking Workshop Bodrum, Turkey, Jun. 2004.

[21] X. Lin, G. Sharma, R. Mazumdar, and N. Shroff, "Degenerate delay-capacity trade-offs in ad hoc networks with Brownian mobility," Dept. Elect. Comput. Eng., Purdue Univ., West Lafayette, IN, 2005 [Online]. Available: http://ece.purdue.edu/~gsharma

[22] A. E. Gamal, J. Mammen, B. Prabhakar, and D. Shah, "Throughput-delay trade-off in wireless networks—Part I: the fluid model," Dept. Elect. Eng., Stanford Univ., Stanford, CA, 2005 [Online]. Available: http://www.stanford.edu/~jmammen/

[23] G. Sharma, R. R. Mazumdar, and N. B. Shroff, "Delay and capacity trade-offs in mobile ad hoc networks: A global perspective," in *Proc. IEEE INFOCOM*, 2006, pp. 1–12.

[24] P. Nain, D. Towsley, B. Liu, and Z. Liu, "Properties of random direction models," in *Proc. IEEE INFOCOM*, 2005, pp. 1897–1907.

[25] J.-Y. Le Boudec and M. Vojnovic, "Perfect simulation and stationarity of a class of mobility models," in *Proc. IEEE INFOCOM*, 2005, pp. 2743–2754.

[26] D. Aldous and J. Fill, "Reversible Markov chains and random walks on graphs," Monograph in preparation, 2002 [Online]. Available: http://stat-www.berkeley.edu/users/aldous/RWG/book.html

[27] A. Gut, *Probability: A Graduate Course*. New York: Springer, 2005.

Gaurav Sharma received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Delhi, India, in 2002, and the Ph.D. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, in 2006.

Currently, he is working as a Quantitative Analyst with D. E. Shaw & Co., L.P. in New York, NY. He worked with IBM India Research Lab, Delhi, and Microsoft Research, Cambridge, U.K., as a summer intern in 2001 and 2004, respectively. His research interests are in mathematical modeling and performance evaluation of communication networks, game theory, and applied probability theory.

Dr. Sharma is a recipient of the Best Paper Award at IEEE INFOCOM 2006 and travel grants from NSF and ACM. He was a finalist for the 2006 Chorafas Foundation Prize at Purdue University.

Ravi R. Mazumdar (M'83–SM'94–F'05) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Bombay, India, in 1977, the M.Sc. DIC in control systems from Imperial College, London, U.K., in 1978, and the Ph.D. degree in systems science from the University of California, Los Angeles, in 1983.

He is currently a University Research Chair Professor of Electrical and Computer Engineering at the University of Waterloo, Waterloo, Canada, and an Adjunct Professor of Electrical and Computer Engineering at Purdue University, West Lafayette, IN. He has served on the faculties of Columbia University (New York, NY), INRS-Telecommunications (Montreal, Canada), University of Essex (Colchester, U.K.), and most recently at Purdue University. He has held visiting positions and sabbatical leaves at UCLA, the University of Twente (Netherlands), the Indian Institute of Science (Bangalore), and the Ecole Nationale Superieure des Telecommunications (Paris). His research interests are in wireless and wireline networks, applications of game theory to networking, applied probability, queueing theory, and stochastic analysis with applications to traffic engineering, stochastic filtering theory, and mathematical finance.

Dr. Mazumdar is a Fellow of the IEEE and the Royal Statistical Society. He is a member of the working groups WG6.3 and 7.1 of the IFIP and a member of SIAM and the IMS. He shared the Best Paper Award with G. Sharma and N. Shroff at the IEEE INFOCOM 2006 in Barcelona and was also co-author (with N. Likhanov) of a paper that was runner-up for the Best Paper Award at IEEE INFOCOM 1998 in San Francisco.

Ness B. Shroff (F'07) received the Ph.D. degree from Columbia University, New York, NY, in 1994.

He joined Purdue University, West Lafayette, IN, immediately thereafter as an Assistant Professor. At Purdue, he became Professor of the school of Electrical and Computer Engineering in 2003 and Director of CWSA in 2004, a university-wide center on wireless systems and applications. In 2007, he joined The Ohio State University, Columbus, OH, as the Ohio Eminent Scholar of Networking and Communications, and Professor of ECE and CSE. His research interests span the areas of wireless and wireline communication networks. He is especially interested in fundamental problems in the design, performance, control, and security of these networks.

Dr. Shroff is an editor for IEEE/ACM TRANSACTIONS ON NETWORKING and the *Computer Networks Journal*, and past editor of IEEE COMMUNICATIONS LETTERS. He has served on the technical and executive committees of several major conferences and workshops. He was the technical program co-chair of IEEE INFOCOM 2003, the premier conference in communication networking. He was also the conference chair of the 14th Annual IEEE Computer Communications Workshop (CCW'99), the program co-chair for the Symposium on High-Speed Networks, Globecom 2001, and the panel co-chair for ACM Mobicom'02. He was also a co-organizer of the NSF Workshop on Fundamental Research in Networking, held in Airlie House, Virginia, in 2003. In 2008, he will serve as the technical program co-chair of ACM Mobihoc 2008. He received the IEEE INFOCOM 2006 Best Paper Award, the IEEE IWQoS 2006 Best Student Paper Award, the 2005 Best Paper of the Year Award for the *Journal of Commnications and Networking*, the 2003 Best Paper of the Year Award for Computer Networks, and the NSF CAREER Award in 1996. (His INFOCOM 2005 paper was also selected as one of two runner-up papers for the Best Paper Award.)