Throughput-Delay Trade-off in Wireless Networks

Abbas El Gamal, James Mammen, Balaji Prabhakar, Devavrat Shah
Departments of EE and CS
Stanford University
{abbas, jmammen, balaji, devavrat}@stanford.edu

Abstract— **Gupta and Kumar (2000) introduced a random network model for studying the way throughput scales in a wireless network when the nodes are fixed, and showed that the throughput per source-destination pair is $\Theta\left(1/\sqrt{n \log n}\right)$. Grossglauser and Tse (2001) showed that when nodes are mobile it is possible to have a constant or $\Theta(1)$ throughput scaling per source-destination pair.**

The focus of this paper is on characterizing the delay and determining the throughput-delay trade-off in such fixed and mobile ad hoc networks. For the Gupta-Kumar fixed network model, we show that the optimal throughput-delay trade-off is given by $D(n) = \Theta(nT(n))$, where $T(n)$ and $D(n)$ are the throughput and delay respectively. For the Grossglauser-Tse mobile network model, we show that the delay scales as $\Theta\left(n^{1/2}/v(n)\right)$, where $v(n)$ is the velocity of the mobile nodes. We then describe a scheme that achieves the optimal order of delay for any given throughput. The scheme varies (i) the number of hops, (ii) the transmission range and (iii) the degree of node mobility to achieve the optimal throughput-delay trade-off. The scheme produces a range of models that capture the Gupta-Kumar model at one extreme and the Grossglauser-Tse model at the other. In the course of our work, we recover previous results of Gupta and Kumar, and Grossglauser and Tse using simpler techniques, which might be of a separate interest.

Keywords: Stochastic processes/Queueing theory, Combinatorics, Information theory, Statistics.

I. Introduction

An ad hoc wireless network consists of a collection of nodes, each capable of transmitting to or receiving from other nodes. When a node transmits to another node, it creates some interference to all other nodes in its vicinity. When several nodes transmit simultaneously, a receiver can successfully receive the data sent by the desired transmitter only if the interference from the other nodes is sufficiently small. An important characteristic of ad hoc wireless networks is that the topology of the nodes may not be known. For example, it may be a sensor network formed by a random configuration of nodes with wireless communication capability. The wireless nodes could also be mobile, in which case the topology could be continuously changing.

Previous research has focused on determining how the throughput of such wireless networks scales with the number of nodes, n, in the network. Gupta and Kumar [5] introduced a random network model for studying throughput scaling in a fixed wireless network; i.e. when the nodes do not move. They defined a random network to consist of n nodes distributed independently and uniformly on a unit disk. Each node has a randomly chosen destination node and can transmit at W bits-per-second provided that the interference is sufficiently small. Thus, each node is simultaneously a source, S, a potential destination, D, and a relay for other source-destination (S-D) pairs. They showed that in such a random network the throughput scales as $\Theta(1/\sqrt{n \log n})$ [1] per S-D pair.

Grossglauser and Tse [4] showed that by allowing the nodes to move, the throughput scaling changes dramatically. Indeed, if node motion is independent across nodes and has a uniform stationary distribution, a constant throughput scaling ($\Theta(1)$) per S-D pair is feasible. Later, Diggavi, Grossglauser and Tse [2] also showed that a constant throughput per S-D pair is feasible even with a more restricted mobility model.

The way in which delay scales for such throughput optimal schemes, however, has not been well-studied. Indeed, it is unclear precisely what "delay" means, especially in mobile networks. One of the main contributions of this paper is a definition of delay, which is both meaningful and makes derivations possible.

From [5] and [4], one may make the following inferences about the trade-off between throughput and delay: (i) In a fixed random network a small transmission range is necessary to limit interference and hence to obtain a high throughput. This results in multi-hopping, and consequently leads to high delays. (ii) On the other hand, mobility allows nodes to approach one another closely. This not only allows the use of small transmission ranges, but more crucially, it allows the use of a single relay node, which boosts throughput to $\Theta(1)$. However, the delay is now dictated by the node velocity (which is much lower than the

[1] We recall the following notation: (i) $f(n) = O(g(n))$ means that there exists a constant c and integer N such that $f(n) \leq cg(n)$ for $n > N$. (ii) $f(n) = o(g(n))$ means that $\lim_{n\to\infty} f(n)/g(n) = 0$. (iii) $f(n) = \Omega(g(n))$ means that $g(n) = O(f(n))$, (iv) $f(n) = \omega(g(n))$ means that $g(n) = o(f(n))$. (v) $f(n) = \Theta(f(n))$ means that $f(n) = O(g(n));\ g(n) = O(f(n))$.

speed of electromagnetic propagation).

The above observations point out three important features that influence the throughput and delay in ad hoc networks: (i) the number of hops, (ii) the transmission range, and (iii) the node mobility and velocity. We propose schemes that exploit these three features to different degrees to obtain different points on the throughput-delay curve in an *optimal* way (see Figure 1). In fixed networks, our Scheme 1 achieves the throughput-delay trade-off shown by segment PQ in Figure 1 and at the highest throughput, it reduces to the Gupta-Kumar scheme (point Q in the figure). In the presence of mobility, and using only one relay per packet (no multi-hopping), our Scheme 2 is essentially the Grossglauser-Tse scheme (point R in the figure). At this highest achievable throughput, we are able to compute the exact order of delay as network size increases. For lower throughputs, by using the number of hops and node mobility optimally, Scheme 3 obtains different points on the throughput-delay curve shown by segment PR in Figure 1. Before summarizing these statements more precisely, we shall need to define what we mean by throughput and delay.

Definition of throughput: A throughput $\lambda > 0$ is said to be feasible/achievable if every node can send at a rate of λ bits per second to its chosen destination. We denote by $T(n)$, the maximum feasible throughput with high probability [2] (*whp*). In this paper, $T(n)$ will be the maximum delay-constrained throughput. When there is no delay constraint, $T(n)$ is simply the *throughput capacity* as in [5], [4].

Definition of delay: The delay of a packet in a network is the time it takes the packet to reach the destination after it leaves the source. We do not take queueing delay at the source into account, since our interest is in the network delay. The average packet delay for a network with n nodes, $D(n)$, is obtained by averaging over all packets, all source-destination pairs, and all random network configurations.

In a fixed network, the delay equals the sum of the times spent at each relay. In a mobile network also, the delay is the sum of the times spent at each relay. However, in this case, delay depends on the velocity, $v(n)$, of each relay.

For a meaningful measure of delay per packet, it is important to scale the size of a packet depending on the throughput. If throughput is λ, the transmission delay (or service time) of a packet of fixed size would scale as $1/\lambda$. This would dominate the overall delay and hence would not let us capture the delay caused by the dynamics of the network/scheme. To counteract this, we let the packet size scale as λ so that the transmission delay (service time) is always constant.

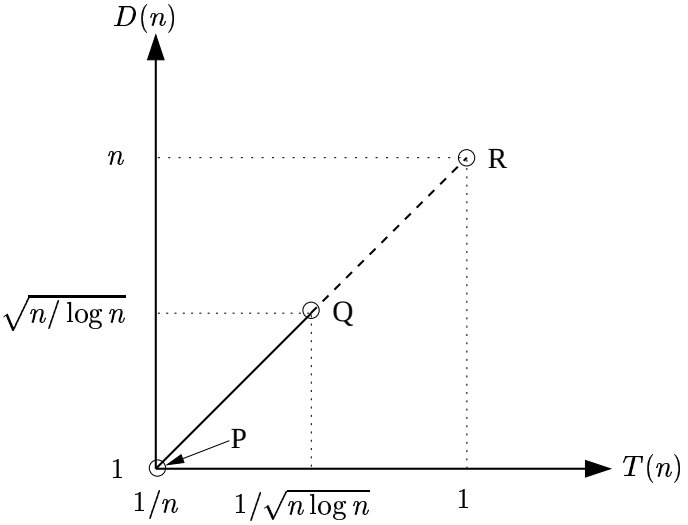

Fig. 1. Throughput-delay scaling trade-off for a wireless network assuming $v(n) = \Theta\left(1/\sqrt{n}\right)$. The marks on the axes represent the orders asymptotically in n.

A. Outline and Summary of results

Fixed random network: In Section II, we introduce Scheme 1 and show that the dependence of the optimal delay on throughput for a fixed random network is given by

$$D(n) = \Theta\left(nT(n)\right), \text{ for } T(n) = O\left(1/\sqrt{n\log n}\right). \tag{1}$$

The above result says the following: (i) The highest throughput per node achievable in a fixed network is $\Theta\left(1/\sqrt{n\log n}\right)$, as Gupta and Kumar obtained. At this throughput the average delay $D(n) = \Theta\left(\sqrt{n/\log n}\right)$ (point Q in Figure 1). (ii) By increasing the transmission radius the average number of hops can be reduced. But, because the interference is higher now, the throughput would be lower. When throughput is smaller than $\Theta\left(1/\sqrt{n\log n}\right)$, equation (1) shows how $D(n)$ is related to $T(n)$ (segment PQ in Figure 1).

Delay in a mobile network for $T(n) = \Theta(1)$: In Section III, we introduce Scheme 2 in which nodes move according to independent Brownian motions and use a single relay as in Grossglauser and Tse. This scheme achieves throughput $T(n) = \Theta(1)$. Using results from random walks [3] and queueing theory [8] we show that the delay, $D(n)$, (both due to node mobility and queueing at the relay) is given by

$$D(n) = O\left(\sqrt{n}/v(n)\right) \text{ when } T(n) = \Theta(1).$$

Here $v(n)$ denotes the way node velocity scales with n. Taking $v(n) = \Theta(1/\sqrt{n})$, the above point is shown as R in Figure 1.

[2] In this paper, *whp* means with probability $\geq 1 - 1/n$.

Throughput delay trade-off in a mobile network: In Section IV we introduce Scheme 3, where the trade-off is achieved using multiple hops. The trade-off is parametrized by $a(n)$, where $\sqrt{a(n)}$ corresponds to the average distance traveled in one hop. The range of $a(n)$ is from $\Theta(\log n/n)$ (corresponding to the Grossglauser-Tse model, point R)[3] to $\Theta(1)$ (corresponding to the Gupta-Kumar model, point Q). The optimal throughput-delay trade-off for $T(n)$, in the range between $\Theta\left(1/\sqrt{n\log n}\right)$ and $\Theta\left(1/\log n\right)$, is given by

$$T(n) = \Theta\left(1/\sqrt{na(n)\log n}\right), \text{ and}$$

$$D(n) = \Theta\left(1/v(n)\sqrt{a(n)}\right).$$

This is shown by the segment QR in Figure 1.

II. Throughput-Delay Trade-off for Fixed Networks

We consider a random network model similar to that introduced by Gupta and Kumar [5]. There are n nodes distributed uniformly at random on a unit torus and each node has a randomly chosen destination. We assume the unit torus to avoid edge effects, which otherwise complicates the analysis. We note, however, that the results in the paper hold for a unit square as well. Each node transmits at W bits per second, which is a constant, independent of n.

We assume slotted time for transmission. For successful transmission, we assume a model similar to the *Protocol* model as defined [5]. Under our *Relaxed Protocol* model, a transmission from node i to node j is successful if for any other node k that is transmitting simultaneously,

$$d(k,j) \geq (1+\Delta)d(i,j) \text{ for } \Delta > 0$$

where $d(i,j)$ is the distance between nodes i and j. This is a slightly more general version of the model presented in [5] in the sense that nodes do not require a common range of transmission.

In the other commonly used model (e.g., [5], [4]), known as the *Physical* model, a transmission is successful if the Signal to Interference and Noise Ratio (SINR) is greater than some constant. It is well known [5] that with a fading factor $\alpha > 2$, the *Protocol* model is equivalent to the *Physical* model, where each transmitter uses the same power. In the rest of the paper we shall assume the *Relaxed Protocol* model.

[3]To be precise, their scheme corresponds to $a(n) = \Theta(1/n)$, which is covered by our Scheme 2. For the technique we use to analyze Scheme 3 to work, we need $a(n) = \Omega(\log n/n)$. For the same reason, we also consider $T(n) = O(1/\log n)$ instead of $T(n) = O(1)$ in Scheme 3.

We now present a parametrized communication scheme and show that it achieves the optimal trade-off between throughput and delay. This scheme is a generalization of the Gupta-Kumar random network scheme [5].

Scheme 1:

- Divide the unit torus using a square grid into square cells, each of area $a(n)$ (see Figure 2).
- A cellular time-division multi-access (TDMA) transmission scheme is used, in which, each cell becomes *active*, i.e., its nodes can transmit successfully to nodes in the cell or in neighboring cells, at regularly scheduled *cell time-slots* (see Figure 3).
- Let the straight line connecting a source S to its destination D be denoted as an S-D line. A source S transmits data to its destination D by hops along the adjacent cells lying on its S-D line as shown in Figure 2.
- When a cell becomes active, it transmits a single packet for each of the S-D lines passing through it. This is again performed using a TDMA scheme that slots each cell time-slot into *packet time-slots* as shown in Figure 3.

The following theorem characterizes the achievable trade-off for the above scheme. The optimality of this scheme will be proved in Theorem 2.

Theorem 1. *For Scheme 1 with $a(n) \geq 2\log n/n$,*

$$T(n) = \Theta\left(\frac{1}{n\sqrt{a(n)}}\right) \text{ and } D(n) = \Theta\left(\frac{1}{\sqrt{a(n)}}\right),$$

i.e., the achievable throughput-delay trade-off is

$$T(n) = \Theta\left(\frac{D(n)}{n}\right).$$

To prove Theorem 1, we need the following three lemmas. Lemma 1 shows that each cell will have at least one node *whp*, thus guaranteeing successful transmission along each S-D line. Lemma 2 shows that each cell can be active for a constant fraction of time, independent of n. Lemma 3 bounds the maximum number of S-D lines passing through any cell. Combining these results yields a proof of Theorem 1.

Lemma 1. *(a) If $a(n) \geq 2\log n/n$, then all cells have at least one node whp.*
(b) For $a(n) = \Omega(\log n/n)$, each cell has $na(n) \pm \sqrt{2na(n)\log n}$ nodes whp. In particular, if $a(n) = \omega(\log n/n)$ then each cell has $na(n) \pm o(na(n))$ nodes.
(c) Let $a(n) = 1/n$ and let $c_k(n), k \geq 0$ be the fraction of cells with k nodes. Then whp

$$c_k(n) = e^{-1}/k!.$$

This lemma can be proved using well-known results (for example, see [7], Chapter 3). Due to space constraints, we do not repeat the proof here.

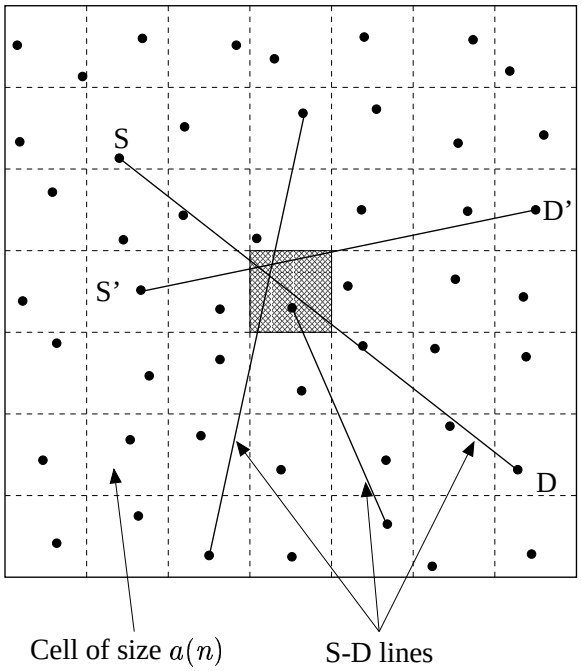

Fig. 2. The unit torus is divided into cells of size $a(n)$ for Scheme 1. The S-D lines passing through the shaded cell in the center are shown.

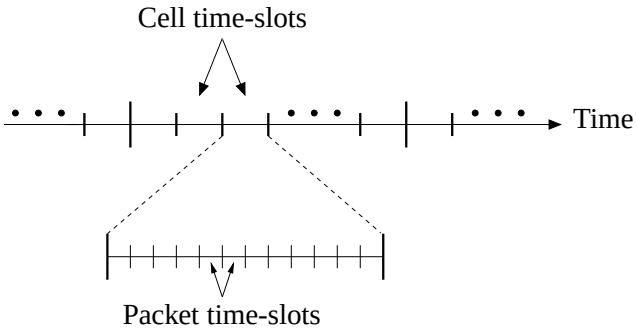

Fig. 3. The TDMA transmission schedule of Scheme 1. The number of cell time-slots is constant while the number of packet time-slots is $O(\sqrt{na(n)})$. (Note that a cell here refers to a square cell obtained by the overlay of the unit torus by a square grid and not a packet of fi xed size as commonly used in networking literature.)

Before stating Lemma 2, we make the following definition: We say that cell B interferes with another cell A if a transmission by a node in cell B can affect the success of a simultaneous transmission by a node in cell A.

Lemma 2. *Under the Relaxed Protocol model, the number of cells that interfere with any given cell is bounded above by a constant c_1, independent of n.*

Proof. Consider a node in a cell transmitting to another node within the same cell or in one of its 8 neighboring cells. Since each cell has area $a(n)$, the distance between the transmitting and receiving nodes cannot be more than $r = \sqrt{8a(n)}$. Under the Relaxed Protocol model, data is successfully received if no node within distance $\bar{r} = (1 + \Delta)r$ of the receiver transmits at the same time. Therefore, the number of interfering cells, c_1, is at most

$$c_1 \leq 2\frac{\bar{r}^2}{a(n)} = 16(1+\Delta)^2,$$

which, for a constant Δ, is a constant, independent of n (and $a(n)$). □

A consequence of Lemma 2 is that interference-free scheduling among all cells is possible, where each cell becomes active once in every $1 + c_1$ slots. In other words, each cell can have a constant throughput. Now we bound the maximum number of S-D lines passing through any cell.

Lemma 3. *The number of S-D lines passing through any cell is $O(n\sqrt{a(n)})$, whp.*

Proof. Consider n S-D pairs. Let d_i be the distance between the S-D pair i, i.e., the length of S-D line i. Let h_i be the number of hops per packet for S-D pair i. Then $h_i = d_i/\sqrt{a(n)}$. Let $H = \sum_{i=1}^{n} h_i$, i.e., the total number of hops required to send one packet from each sender S to its corresponding destination D.

Now consider a particular cell and define the Bernoulli random variables Y_k^i, for S-D pairs $1 \leq i \leq n$ and hops $1 \leq k \leq h_i$, to be equal to 1 if hop k of S-D pair i's packet originates from a node in the cell. Hence, the total number of S-D lines passing through the cell is $Y = \sum_{i=1}^{n}\sum_{k=1}^{h_i} Y_k^i$. Note that since the nodes are randomly distributed, the Y_k^is are identically distributed. For any $1 \leq i \neq j \leq n$, Y_k^i and Y_l^j (for any $1 \leq k \leq h_i$, $1 \leq l \leq h_j$) are independent. However, for any given $1 \leq i \leq n$, Y_k^i and Y_l^i (for any $1 \leq k \neq l \leq h_i$) are dependent and in fact the event $\{Y_k^i = 1, Y_l^i = 1\}$ is not possible, as S-D line i can intersect the cell at most once.

First consider the random variable $H = \sum_{i=1}^{n} d_i/\sqrt{a(n)}$. Since, for all i, $d_i \in [0, 1/\sqrt{2}]$, $H = O(n/\sqrt{a(n)})$. Now, we use this result to find a bound on $E[Y]$ as follows

$$\begin{aligned} E[Y] &= E_H[E[Y|H]] \\ &= E_H\left[\sum_{i=1}^{n}\sum_{k=1}^{h_i} E[Y_k^i|H]\right] \\ &= E_H\left[HE[Y_1^1]\right] \\ &= \Theta(n\sqrt{a(n)}), \end{aligned} \tag{2}$$

where (2) follows from the fact that, by the symmetry of the torus, any hop is equally likely to originate from any of the $1/a(n)$ cells .

Consider a random variable $\tilde{Y} = \sum_{l=1}^{H} \tilde{Y}_l$, where $\tilde{Y}_l$ are i.i.d. Bernoulli random variables with $\Pr(\tilde{Y}_1 = 1) = \Pr(Y_1^1 = 1) = a(n)$. Because of the particular dependence of Y_l^i and Y_k^i (for any given $1 \leq i \leq n$ and $1 \leq k \neq l \leq h_i$), it can be shown that, for any m,

$$E[Y^m] \leq E[\tilde{Y}^m].$$

This implies that, for any $\phi > 0$,

$$E[\exp(\phi Y)] \leq E[\exp(\phi \tilde{Y})]. \quad (3)$$

For any $\delta > 0$, define $P(\tilde{Y}, \delta) \stackrel{\triangle}{=} \Pr(\tilde{Y} \geq (1+\delta)E[\tilde{Y}])$. By the Chernoff bound for i.i.d. Bernoulli random variables,

$$P(\tilde{Y}, \delta) \quad \leq \quad \exp(-\delta^2 E[\tilde{Y}]/2). \quad (4)$$

Consider the following:

$$\begin{aligned} P(Y, \delta) &\stackrel{\triangle}{=} \Pr(Y \geq (1+\delta)E[Y]) \\ &= \Pr(\exp(\phi Y) \geq \exp(\phi(1+\delta)E[Y])) \\ &\leq \frac{E[\exp(\phi Y)]}{\exp(\phi(1+\delta)E[Y])} \quad (5) \\ &\leq \frac{E[\exp(\phi \tilde{Y})]}{\exp(\phi(1+\delta)E[\tilde{Y}])}, \quad (6) \end{aligned}$$

where (5) follows by the Markov inequality and (6) follows from (3) and the fact that $\exp(\phi(1+\delta)E[Y]) = \exp(\phi(1+\delta)E[\tilde{Y}])$. From (6) and the proof of the Chernoff bound (for example, see [7], pg. 68) it follows that $P(Y, \delta)$ can be bounded above by the bound on $P(\tilde{Y}, \delta)$ as given in (4).

By taking $\delta = 2\sqrt{\log n / E[Y]}$, we obtain

$$\Pr(Y \geq EY + 2\sqrt{\log n E[Y]}) \quad \leq \quad 1/n^2. \quad (7)$$

Thus, for any cell, the number of hops originating from it are bounded above by $n\sqrt{a(n)} + o(n\sqrt{a(n)})$ with probability $\geq 1 - 1/n^2$. Since there are at most n cells, by the union of events bound, the above bound holds for all cells with probability $\geq 1 - 1/n$. This completes the proof of the lemma. □

We are now ready to prove Theorem 1.

Proof of Theorem 1. From Lemma 2, it follows that each cell can be active for a guaranteed fraction of time, i.e., it can have a constant throughput. Lemma 3 suggests that if each cell divides its cell time-slot into $\Theta\left(n\sqrt{a(n)}\right)$ packet time-slots, each S-D pair hopping through it can use one packet time-slot. Equivalently, each S-D pair can successfully transmit for $\Theta\left(1/n\sqrt{a(n)}\right)$ fraction of time. That is, the achievable throughput per S-D pair is $T(n) = \Theta\left(1/n\sqrt{a(n)}\right)$.

Next we compute the average packet delay $D(n)$. As defined earlier, packet delay is the sum of the amount of time spent in each hop. We first bound the average number of hops then show that the time spent in each hop is constant, independent of n.

Since each hop covers a distance of $\Theta\left(\sqrt{a(n)}\right)$, the number of hops per packet for S-D pair i is $\Theta\left(d_i/\sqrt{a(n)}\right)$, where d_i is the length of S-D line i. Thus the number of hops taken by a packet averaged over all S-D pairs is $\Theta\left(\frac{1}{n}\sum_{i=1}^{n} d_i/\sqrt{a(n)}\right)$. Since for large n, the average distance between S-D pairs is $\frac{1}{n}\sum_{i=1}^{n} d_i = \Theta(1)$, the average number of hops is $\Theta\left(1/\sqrt{a(n)}\right)$.

Now note that by Lemma 2 each cell can be active once every constant number of cell time-slots and by Lemma 3 each S-D line passing through a cell can have its own packet time-slot within that cell's time-slot. Since we assumed that packet size scales in proportion to the throughput $T(n)$, each packet arriving at a node in the cell departs within a constant time.

From the above discussion, we conclude that the delay $D(n) = \Theta\left(1/\sqrt{a(n)}\right)$. This concludes the proof of Theorem 1. □

Next we show that Scheme 1 provides the optimal throughput-delay trade-off for a fixed wireless network.

Theorem 2. *Let the average delay be bounded above by $D(n)$. Then the achievable throughput $T(n)$ for any scheme scales as $O\left(\frac{D(n)}{n}\right)$.*

Proof. The proof uses similar techniques to the proof of Theorem 2.1 in [5]. Consider a given fixed placement of n nodes in the unit torus. Let $\bar{L}$ be the sample mean of the lengths of the S-D lines for the given node placement and let the throughput be λ. Consider a large enough time t over which the total number of bits transported in the network is $\lambda n t$. Let $h(b)$ be the number of hops taken by bit b, $1 \leq b \leq \lambda n t$ and let $r(b,h)$ denote the length of hop h of bit b. Therefore,

$$\sum_{b=1}^{\lambda n t} \sum_{h=1}^{h(b)} r(b,h) \geq \lambda n t \bar{L}. \quad (8)$$

Now, for two simultaneous transmissions from node i to node j and from node k to node l,

$$\begin{aligned} d(j,l) &\geq d(j,k) - d(l,k) \\ &\geq (1+\Delta)d(i,j) - d(l,k), \end{aligned}$$

and similarly,

$$\begin{aligned} d(j,l) &\geq d(l,i) - d(i,j) \\ &\geq (1+\Delta)d(l,k) - d(i,j). \end{aligned}$$

Combining the above two inequalities, we obtain

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

This result implies that if we place a disk around each receiver of radius $\Delta/2$ times the length of the hop, the disks must be disjoint for successful transmission under the Protocol model. Since a node transmits at W bits per

second, each bit transmission time is $1/W$ seconds. During each bit transmission, the total area covered by the disks surrounding the receivers must be less than the total unit area. Summing over the Wt bits transmitted in time t, we obtain

$$\sum_{b=1}^{\lambda nt}\sum_{l=1}^{h(b)}\frac{\pi}{4}\left(\frac{\Delta}{2}r(b,l)\right)^2 \leq Wt. \tag{9}$$

Let the total number of hops taken by all bits be $H = \sum_{b=1}^{\lambda nt} h(b)$. Then, by convexity, it follows that

$$\left(\sum_{b=1}^{\lambda nt}\sum_{h=1}^{h(b)}\frac{1}{H}r(b,h)\right)^2 \leq \sum_{b=1}^{\lambda nt}\sum_{h=1}^{h(b)}\frac{1}{H}r(b,h)^2. \tag{10}$$

Combining (9) and (10) gives

$$\left(\sum_{b=1}^{\lambda nt}\sum_{h=1}^{h(b)}\frac{1}{H}r(b,h)\right)^2 \leq \frac{16Wt}{\pi\Delta^2 H}. \tag{11}$$

Substituting from (8) into (11) and rearranging, we obtain

$$\left(\lambda nt\bar{L}\right)^2 \leq \left(\frac{16Wt}{\pi\Delta^2}\right)H.$$

Now defining $\overline{h(b)}$ to be the sample mean of the number of hops over λnt bits, we obtain

$$\overline{h(b)} = \frac{1}{\lambda nt}\sum_{b=1}^{\lambda nt}h(b) = \frac{1}{\lambda nt}H. \tag{12}$$

Substituting from (12) into (11) and rearranging, we obtain

$$\lambda n \leq \frac{16W}{\pi\Delta^2\bar{L}^2}\overline{h(b)}. \tag{13}$$

By definition the throughput capacity $T(n) \leq \lambda$ with high probability. As a result, $E(\lambda) \geq T(n)$. Substituting into (13), we obtain

$$nT(n) \leq E(\lambda)n \leq \frac{16W}{\pi\Delta^2\bar{L}^2}E(\overline{h(b)}).$$

Now, since the average number of hops per bit is the same as the average number of hops per packet and the packet-size scales as $T(n)$, the time spent by a packet at each relay is $\Omega(1)$. Therefore, the average delay, $D(n)$, is of the same order as the average number of hops per bit, $E(\overline{h(b)})$. This concludes the proof of Theorem 2. □

III. Delay in a Mobile network for $T(n) = \Theta(1)$

In this section we consider a random network with mobile nodes similar to the model introduced by Grossglauser and Tse in [4]. They showed that under the Physical model $T(n) = \Theta(1)$ is achievable. We assume n nodes forming n S-D pairs in a torus of unit area and assume slotted transmission time. Each node moves independently and uniformly on the unit torus. Thus, at a given time, a node is equally likely to be in any part of the torus independent of the location of any other node. We first present a scheme (which is similar to that in [4]) and show that it achieves constant throughput and then analyze its delay in Subsection III-A.

Scheme 2:

- Divide the unit torus into n square cells, each of area $1/n$.
- Each cell becomes active once in every $1 + c_1$ cell time-slots as discussed in Lemma 2.
- In an active cell, the transmission is always between two nodes within the same cell.
- In an active cell, if two or more nodes are present pick one at random. Each cell time-slot is divided into two sub-slots A and B.
  - In sub-slot A, the randomly chosen node transmits to its destination node if it is present in the same cell. Otherwise, it transmits its packet to a randomly chosen node in the same cell, which acts as a relay.
  - In sub-slot B, the randomly chosen node picks another node at random from the same cell and transmits to it a packet that is destined to it.

We now prove that this scheme achieves constant throughput scaling. The proof is simpler than the one given in [4] and, as we shall see, will help us analyze delay and characterize the throughput-delay trade-off in mobile wireless networks (see Section IV).

Theorem 3. *The throughput using Scheme 2 is* $T(n) = \Theta(1)$.

Proof. The proof is based on Part (c) of Lemma 1 and Lemma 2 as follows:

Each packet is transmitted directly to its destination or relayed at most once and hence the net traffic is at most twice the original traffic. Since: (i) a node is chosen to be a relay at random from the other nodes in the same cell and (ii) the nodes have independent and uniformly distributed motion, each source's traffic gets spread uniformly among all other nodes (similar to the argument in [4]). As a result, in steady state, each node has packets for every other node for a constant fraction of time c_2.

Since in any cell time-slot, the n nodes are uniformly distributed on the torus and the unit torus is divided into n square cells each of area $1/n$, by Lemma 1(c), $1 - 2e^{-1} \approx 0.26$ fraction of the cells contain at least 2 nodes. Thus from Lemma 2, $0.26c_2/(1 + c_1)$ fraction of cells can execute the scheme successfully. Since each cell has a throughput of $\Theta(1)$, the net throughput in any time-slot is $\Theta(n)$ *whp*. Moreover, due to reasons (i) and (ii) above, the throughput of $\Theta(n)$ is divided among all n pairs equally. Thus, the throughput per S-D pair is $T(n) = \Theta(1)$. □

A. Analysis of Delay

To analyze the delay for Scheme 2, we make the additional assumption that each node moves according to an independent 2-dimensional Brownian motion on the torus. Note that in the cellular setting with n cells, a Brownian motion on the torus yields a symmetric random walk on a 2-D torus of size $\sqrt{n} \times \sqrt{n}$.

Let the node velocity scale as $v(n)$. We assume that $v(n)$ scales down as a function of n. This is motivated by the fact that in a real network, each node would occupy a constant amount of area, and thus as the network scales, the overall area must scale accordingly. However, in our model, as in [5], [4], we keep the total area fixed and therefore to simulate a real network we must scale $v(n)$ down.

Note that a node travels to one of its neighboring cells every $t(n)$ time-slots, where

$$t(n) = \Theta\left(1/\sqrt{n}v(n)\right). \tag{14}$$

Thus, we assume that each node moves according to a random walk on the torus, where each move occurs every $t(n)$ time-slots.

We now precisely define delay for Scheme 2. Since the nodes perform independent random walks, only $\Theta(1/n)$ of the packets belonging to any S-D pair reach their destination in a single hop (which happens when both S and D are in the same cell). Thus, most of the packets reach their destination via a relay node, where the delay has two components: (i) *hop-delay*, which is constant, independent of n, and (ii) *mobile-delay*, which is the time the packet spends at the relay while it is moving. To compute mobile-delay we first model the queues formed at a relay node for each S-D pair as a GI/GI/1-FCFS. Then we characterize the inter-arrival and inter-departure times of the queue to obtain the average delay in the mobile case.

Relay queue model: For each S-D pair, each of the remaining $n-2$ nodes can act as a relay. Each node keeps a separate queue for each S-D pair as illustrated in Figure 4. Thus the mobile-delay is the average delay at such a queue. By symmetry, all such queues at all relay nodes are identical. Consider one such queue[4], i.e., fix an S-D pair and a relay node R. To compute the average delay for this queue, we need to study the characteristics of its arrival and departure processes. A packet arrives when (i) R is in the same cell as S, and (ii) the cell becomes active. Similarly, a packet departs when R is in the same active cell as D. Let p be the probability that the cell is active when both R and S are in it. Note that p does not vary with n. Define the inter-meeting time of two nodes as the time between two consecutive instants where they are both in the same cell. Since the node motion is independent of the event that the cell is active, the inter-arrival time is a sum of a Geometric number, $K \sim \text{Geom}(p)$, of inter-meeting times of S and R. Hence the inter-arrival time is of the same order as the the inter-meeting time of S and R. Similarly, the inter-departure time is also of the same order as the inter-meeting time of R and D.

[4]For delay to be finite, the arrival rate must be strictly smaller than the service rate. To ensure this, we assume that if the available throughput is $T(n)$, each source transmits at a rate $(1-\epsilon)T(n)$, for some $\epsilon > 0$.

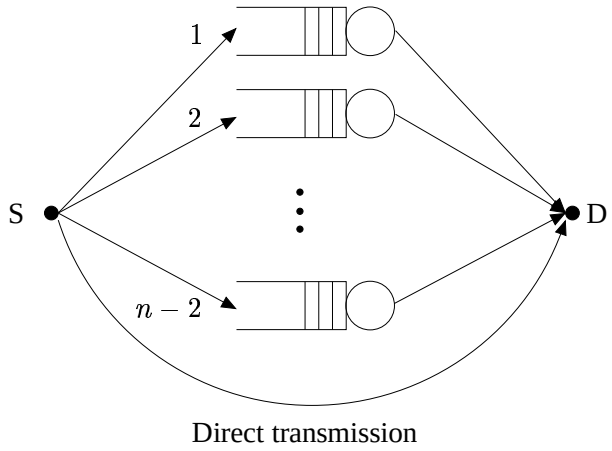

Fig. 4. For any S-D pair, the remaining $n-2$ nodes act as relays. Each node maintains a separate queue for each of the $n-2$ S-D pairs.

Average delay of GI/GI/1-FCFS queue: Since the nodes perform independent symmetric random walks, the queue at each relay node is GI/GI/1-FCFS. The average delay for a GI/GI/1-FCFS queue can be bounded using the first and second moments of the inter-arrival and inter-departure times. We recall the following upper bound on the average delay for a GI/GI/1-FCFS queue known as Kingman's upper bound (see [8], page 476).

Lemma 4. *Consider a discrete GI/GI/1-FCFS queue. Let $A(i), i \in \mathbb{Z}$ be stationary independent inter-arrival times, and $S(i), i \in \mathbb{Z}$ be stationary independent inter-departure times. Let*

$$E[A(0)] = \mu; \quad E[S(0)] = (1-\epsilon)\mu,$$
$$Var(A(0)) = \sigma_a^2; \quad Var(S(0)) = \sigma_s^2.$$

Then, the average delay is bounded above as

$$E[D] \leq \max\left\{\mu, \frac{\sigma_a^2+\sigma_s^2}{2\mu\epsilon}\right\}. \tag{15}$$

Also it is trivially true that

$$E[D] = \Omega(\mu). \tag{16}$$

Inter-meeting time analysis: In view of the above lemma, we proceed to compute the first and second moments of the inter-meeting time. The torus with n cells can be viewed as a $\sqrt{n} \times \sqrt{n}$ grid. Let the position of node i at time t be $(X_1^i(t), X_2^i(t))$, where $X_k^i(t) \in \{0, \ldots, \sqrt{n}-1\}, k = 1, 2$. Now consider the difference random walk between

nodes i and j, defined by $(X_1^{ij}(t), X_2^{ij}(t))$ where $X_k^{ij}(t) = X_k^i(t) - X_k^j(t) \mod \sqrt{n}$, for $k = 1, 2$. Since each node is performing an independent symmetric random walk on a 2-dimensional grid (or torus), each of the components $\{X_k^{ij}(t),\ k = 1, 2\}$ is independent of all others. Further since we are interested only in the first two moments, each component can be modeled as an independent symmetric random walk on a one dimensional grid of size $\sqrt{n}$, i.e., for $k = 1, 2$,

$$X_k^{ij}(t+1) = \begin{cases} X_k^{ij}(t) + 1 \bmod \sqrt{n} & \text{w.p. } 1/2 \\ X_k^{ij}(t) - 1 \bmod \sqrt{n} & \text{w.p. } 1/2. \end{cases}$$

The meeting time of two nodes i and j is identified by the event $\{(X_1^{ij}(t), X_2^{ij}(t)) = (0,0)\}$. Thus the inter-meeting time is the random stopping time $T = \inf\{t \geq 1 : (X_1^{ij}(t), X_2^{ij}(t)) = (0,0) \text{ given that } (X_1^{ij}(0), X_2^{ij}(0)) = (0,0)\}$. We need to compute $\sigma_T^2 = E[T^2] - E[T]^2$. For further analysis, we consider only the difference random walk. Also note that the unit time step of the random walk is actually of order $t(n)$ in real time.

For $k = 1, 2$, let

$$T_k = \inf\{t \geq 1 : X_k^{ij}(t) = 0 \text{ such that } X_k^{ij}(0) = 0\}.$$

Define

$$G = \sum_{t=1}^{T} 1_{\{X_1^{ij}(t)=0\}}.$$

Then $T = \sum_{l=1}^{G} T_1(l)$ where $T_1(l)$ are i.i.d. random variables with the same distribution as T_1. As a result,

$$\begin{aligned} E[T^2] &= E[(\sum_{l=1}^{G} T_1(l))^2] \\ &= E[E[(\sum_{l=1}^{G} T_1(l))^2 | G]] \\ &= E[GE[T_1^2] + G(G-1)E[T_1]^2] \\ &= E[G]E[T_1^2] + E[G^2 - G]E[T_1]^2. \quad (17) \end{aligned}$$

The sequence $(X_1^{ij}(t), X_2^{ij}(t))$ forms a Markov chain with a uniform distribution on the n states $\{(a,b) \in \{0, \ldots, \sqrt{n}-1\}^2\}$. By definition T is the inter-visit time of this Markov chain to state $(0,0)$. Since it is a finite state Markov chain with a uniform stationary distribution, $E[T] = n$. Similarly, $E[T_1] = \sqrt{n}$. By definition, we obtain, $E[T] = E[G]E[T_1]$ and hence $E[G] = \sqrt{n}$. Combining this with (17), we obtain

$$E[T^2] - E[T]^2 = \sqrt{n}E[T_1^2] - n^{3/2} + E[G^2]n - n^2. \quad (18)$$

Bound on $E[T_1^2]$: Let $\tilde{X}(t)$ be a symmetric random walk on $\mathbb{Z}$ starting at position 0 and let $\tilde{T} = \inf\{t : \tilde{X}(t) = -1 \text{ or } \tilde{X}(t) = \sqrt{n}\}$. Then $E[T_1^2] = \Theta(E[\tilde{T}^2])$. Now, consider the following lemma, which follows from standard results in probability theory for martingales.

Lemma 5. $E[\tilde{T}^2] = \Theta\left(n^{3/2}\right)$.

Using this result, it follows that

$$E[T_1^2] = \Theta\left(n^{3/2}\right). \quad (19)$$

Bound on $E[G^2]$: Consider two nodes n_1 and n_2, both starting at position 0 at time $t = 0$ and performing independent symmetric random walks on a 1-D torus of size $\sqrt{n}$. By definition, G is the number of times node n_1 visits 0 until both n_1 and n_2 are at position 0 for the first time $T > 0$. Consider the conditional probability of n_2 being at 0 at any time $t > 0$ given that it was at 0 at time $t = 0$. This probability is $\geq 1/\sqrt{n}$ since the stationary distribution of the position of n_2 has probability $1/\sqrt{n}$ for position 0. Moreover, node n_2 performs a random walk independent of n_1 and hence it is easy to see that G is stochastically upper bounded by a Geometric random variable with parameter $1/\sqrt{n}$. Therefore

$$E[G^2] \leq n. \quad (20)$$

Finally from the above discussion, by combining (18), (19) and (20), we obtain the following result.

Lemma 6.

$$E[T] = n,$$
$$\sigma_T^2 = E[T^2] - E[T]^2 = \Theta\left(n^2\right).$$

Now we are ready to compute the average delay of a packet for Scheme 2. From Lemma 6, we obtain $\mu = \Theta(n)$ and $\sigma_a^2, \sigma_s^2 = \Theta(n^2)$. Now using (15) and (16) along with the fact that the actual number of time-slots per unit time as considered for the random walk model is $t(n)$ (as given by (14)), we obtain the following theorem.

Theorem 4. *Under Scheme 2, the average delay incurred by a packet*

$$D(n) = \Theta\left(\frac{\sqrt{n}}{v(n)}\right).$$

From Theorem 4, for $v(n) = \Theta(1/\sqrt{n})$, we obtain, $D(n) = \Theta(n)$, which corresponds to the point R in Figure 1.

IV. Throughput-delay trade-off in mobile networks

In this section we find the optimal throughput-delay trade-off in random mobile networks. To achieve this trade-off, we introduce Scheme 3. This scheme is divided into two parts based on the range of throughput: Scheme 3(a) is for $T(n) = O(1/\sqrt{n \log n})$, while Scheme 3(b) is for $T(n) = \omega(1/\sqrt{n \log n})$.

For fixed networks, with throughput $T(n) = O(1/\sqrt{n \log n})$, Scheme 1 achieves the optimal trade-off of $D(n) = \Theta(nT(n))$. Since the nodes move randomly and independently, use of mobility can only result in higher delays. Hence to achieve a trade-off for throughput $T(n) = O(1/\sqrt{n \log n})$, we use Scheme 3(a) which is an adaptation of Scheme 1 for mobile networks.

To achieve constant throughput scaling, in Scheme 2, the unit torus was divided into square cells of area $1/n$. The transmissions occurred only when the source (or destination) and relay were in the same cell. The effective "neighborhood" of a node was the area of the cell containing it, and the scheme used mobility to bring the relay node into the "neighborhood" of the destination to deliver the packet. This suggests that delay can be decreased by increasing the size of the "neighborhood" of each node. But a larger neighborhood would result in lower throughput due to increased interference, thus providing a trade-off. To achieve the trade-off for $T(n) = \omega(1/\sqrt{n \log n})$, we use Scheme 3(b) which employs both mobility of nodes and relaying across cells to reduce interference.

Scheme 3(a):

- As in Scheme 1, divide the unit torus using a square grid into square cells, each of area $a(n)$ (see Figure 2).
- A cellular TDMA transmission scheme is used, in which, each cell becomes active at regularly scheduled cell time-slots (see Figure 3). From Lemma 2, each cell gets a chance to be active once every $1 + c_1$ cell time-slots.
- A source S sends its packet directly to its destination D if it is in any of the neighboring cells. Otherwise, it randomly chooses a relay node R in an adjacent cell on the S-D line at the time of transmission.
- When the cell containing the relay node R is active, R transmits the packet directly to D, if D is in a neighboring cell. Otherwise, it relays the packet again to a randomly chosen node in a neighboring cell on the straight line connecting it to D. This process continues until the packet reaches the destination.

The following theorem shows that in spite of node mobility, Scheme 3(a) achieves the same throughput-delay trade-off as Scheme 1 for fixed networks.

Theorem 5. *If $v(n)$ satisfies the condition*

$$v(n) = o(\sqrt{\log n/n}), \tag{21}$$

Scheme 3(a) achieves the following trade-off:

$$T(n) = \Theta\left(\frac{D(n)}{n}\right), \text{ for } T(n) = O\left(\frac{1}{\sqrt{n \log n}}\right).$$

Proof. First we show that condition (21) is necessary for every packet to be eventually delivered. Consider a packet relayed from a source toward its destination, and let the initial distance between the source and its destination be d. Each relaying step occurs within $1 + c_1$ time slots. Each time the packet is relayed, the distance between the center of the cell containing the packet and its destination decreases by at least $\sqrt{a(n)}$. On the other hand, since the nodes move with velocity $v(n)$, this distance can increase by at most $(1 + c_1)v(n)$. Thus after the packet is relayed l times, the distance between the center of the cell containing the packet and its destination will be less than $d - l(\sqrt{a(n)} - (1+c_1)v(n))$. Hence if $\sqrt{a(n)} = \omega(v(n))$, the packet eventually reaches its destination. Since we have $a(n) = O(\log n/n)$, this results in condition (21) being necessary for the success of the scheme.

Note that when condition (21) is satisfied, the average number of times a packet has to be relayed in order to reach its destination is of order $\Theta(1/\sqrt{a(n)})$, which is the same as in Scheme 1 for fixed networks. Hence the delay $D(n) = \Theta(1/\sqrt{a(n)})$.

Next we analyze the throughput for Scheme 3(a). Define an S-D *path* (which is not necessarily a straight line) of a packet for a particular S-D pair as the concatenation of line-segments joining the centers of the cells through which it hops. As in the analysis of Scheme 1, in order to determine the throughput, we consider the S-D paths passing through a cell in some time-slot.

From the preceding discussion about the delay, the number of hops h_i, for any packet of an S-D pair i, is $\Theta(1/\sqrt{a(n)})$. Hence $H = \sum_{i=1}^{n} h_i$, as defined in the proof of Lemma 3, has the same order. For a fixed cell and time-slot, define Y_k^i as in the proof of Lemma 3, i.e., Y_k^i is the indicator for the event that hop k of an S-D pair i's packet originates in the cell during this time-slot. The random variable Y_k^i has the same properties as that in Lemma 3, i.e., (i) independence between Y_k^i and Y_l^j for $i \neq j$, (ii) event $\{Y_l^i = 1, Y_k^i = 1\}$ cannot occur for any given $1 \leq i \leq n$, $1 \leq l \neq k \leq h_i$, and (iii) $E[Y_k^i] = 1/a(n)$. As a result, as in Lemma 3, the number of S-D paths passing through any cell at any given time-slot is $O(n\sqrt{a(n)})$. Consequently, the achievable throughput per S-D pair is at least $\Theta(1/n\sqrt{a(n)})$. By choosing a particular $a(n)$ such that $a(n) = \Omega(\log n/n)$ we obtain the trade-off region stated in the theorem. □

To obtain higher throughputs, we need to use mobility, and to obtain lower delay, we need to use multiple hops cleverly. This leads to the following scheme.

Scheme 3(b):

- As in Scheme 3(a), divide the unit torus using a square grid into square cells, each of area $a(n)$. We further lay out an additional grid formed by square sub-cells of size $b(n) = \Theta(\log n/n)$ as shown in Fig 5. Thus each square cell of area $a(n)$ contains $a(n)/b(n)$ sub-cells each of area

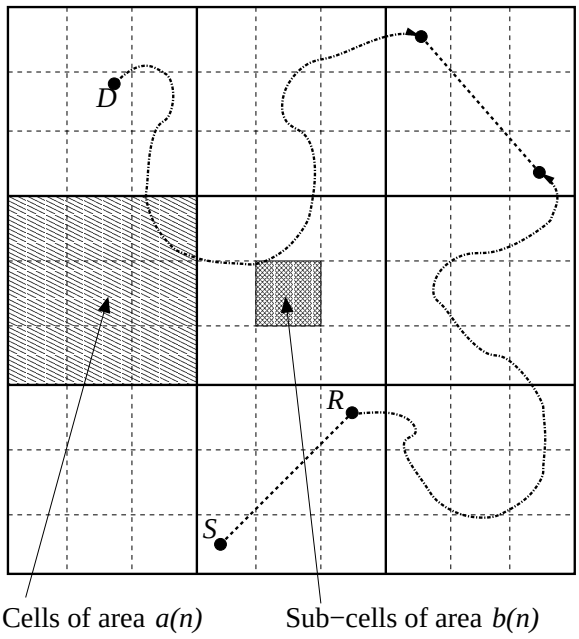

Fig. 5. Scheme 3(b) for throughput-delay trade-off in a mobile network.

$b(n)$.

• A cellular TDMA transmission scheme is used, in which, each cell becomes active at regularly scheduled cell time-slots (see Figure 3). A cell time-slot is divided into $\Theta(na(n))$ packet time-slots.

• An active packet time-slot is divided into two sub-slots A and B.

– In sub-slot A, each node sends a packet to its destination node if it is present in the same cell. Otherwise, it sends its packet to a randomly chosen node in the same cell, which acts as a relay. The packet is sent using hops along sub-cells of size $b(n)$ as in Scheme 3(a).

– In sub-slot B, each node picks another node at random from the same cell and sends a packet that is destined to it. Again, the packet is sent using hops along sub-cells as in Scheme 3(a).

We note that, the above scheme requires the packet-size to scale as $\Theta(1/na(n))$ instead of as $\Theta(1/\sqrt{na(n)})$.

The scheme is depicted in Figure 5. A source S first delivers its packet to a mobile relay node R which is chosen at random from all nodes in the same cell. The mobile relay node R delivers the packet to the destination node D when R and D are in the same cell. In this sense the scheme is similar to Scheme 2. However the packet delivery in both these cases is by hops along sub-cells as in Scheme 3(a).

The following theorem states the trade-off achievable by Scheme 3(b) for mobile networks.

Theorem 6. *If condition (21), i.e., $v(n) = o(\sqrt{\log n/n})$, is satisfied, then Scheme 3(b) achieves the throughput-delay trade-off given by*

$$T(n) = \Theta\left(\frac{1}{\sqrt{na(n)\log n}}\right) \text{ and}$$

$$D(n) = O\left(\frac{1}{a(n)^{1/2}v(n)}\right),$$

where $a(n) = O(1)$ and $a(n) = \Omega(\log n/n)$.

Proof. As discussed in the proof of Theorem 5, in order to guarantee that after leaving its source a packet is eventually delivered to its destination, we must have $b(n) = \omega(v^2(n))$. Since $b(n) = \Theta(\log n/n)$, this implies that condition (21) is necessary for the scheme to be successful.

In steady state, each node has packets for every other node for a constant fraction of the time and the traffic between each source-destination pair is spread uniformly across all other nodes. Note that this is simply a repeat of the statements from the proof of Theorem 3 for Scheme 2.

In any packet time-slot in a given cell: (i) the S-R or R-D pairs are randomly chosen according to Scheme 3(b), (ii) packets are communicated according to Scheme 3(a), and (iii) there are $d(n) = a(n)/b(n)$ sub-cells and $m = na(n)(1 + o(1))$ nodes. Hence, as in the proof of Theorem 5 for Scheme 3(a), the throughput between S-R / R-D pairs is $\Theta(1/m\sqrt{d(n)}) = \Theta(1/\sqrt{na(n)\log n})$. Thus the throughput for any S-D pair is $\Theta(1/\sqrt{na(n)\log n})$.

The delay has two components: (i) hop-delay, which is proportional to the number of hops along sub-cells from a source to the mobile relay and from the mobile relay to the destination and (ii) mobile-delay, which is the time it takes the mobile relay node to reach the cell containing the destination and to deliver the packet to it. The average number of hops taken by a packet in sub-slots A and B is then $\Theta(a(n)/b(n)) = \Theta(na(n)/\log n)$. Hence the hop-delay is $\Theta(a(n)/b(n)) = \Theta(na(n)/\log n)$. The mobile-delay can be analyzed in the same manner as for Scheme 2 with the following differences.

• The inter-meeting time of nodes for Scheme 3(b) is for a random walk on a discrete-torus of size $\sqrt{1/a(n)} \times \sqrt{1/a(n)}$, instead of $\sqrt{n} \times \sqrt{n}$.

• The time taken by a node to move out of a cell is $t(n) = \Theta(\sqrt{a(n)}/v(n))$, instead of $\Theta(1/\sqrt{n}v(n))$

Now using Lemma 6 and Lemma 4 with the two differences mentioned above, the mobile-delay is $\Theta\left(1/a(n)^{1/2}v(n)\right)$. Due to condition (21), the mobile-delay always dominates the hop-delay and hence the average delay is of the same order as the mobile-delay. □

The trade-off obtained by Scheme 3 is demonstrated graphically in Figure 1 assuming $v(n) = \Theta\left(1/\sqrt{n}\right)$.

A. Optimality of Scheme 3

Consider any communication scheme for the random mobile network introduced in Section III. The distance traveled by a packet between its source and destination is the sum of the total distance traveled by hops and the total distance traveled by the mobile relays used. Let $\bar{l}(n)$ be the sample mean distance traveled by hops averaged over all packets. In the following lemma we obtain a bound on the throughput scaling as a function of $\bar{l}(n)$ using a technique similar to the one used in Theorem 2. We then show that to achieve this optimal throughput, the minimum delay incurred is of the same order as the delay of Scheme 3, which will establish the optimality of Scheme 3.

Lemma 7. *The achievable throughput $T(n)$ for any scheme with sample mean distance traveled by hops $\bar{l}(n)$ is bounded above as*

$$T(n) = O\left(\frac{1}{\bar{l}(n)\sqrt{n\log n}}\right). \tag{22}$$

Proof. We merely outline the proof as it is similar to that of Theorem 2. Here the equivalents of (8) and (11) are:

$$\sum_{b=1}^{ntT(n)}\sum_{h=1}^{h(b)} r(b,h) \geq ntT(n)\bar{l}(n) \tag{23}$$

and

$$\left(\sum_{b=1}^{ntT(n)}\sum_{h=1}^{h(b)}\frac{1}{H}r(b,h)\right)^2 \leq \frac{16Wt}{\pi\Delta^2 H} = c_3\frac{t}{H}, \tag{24}$$

where c_3 is a constant that does not depend on n. Substituting from (23) into (24) and rearranging we obtain

$$\frac{ntT(n)\bar{l}(n)}{H}\bar{r} \leq c_3\frac{t}{H}, \tag{25}$$

where

$$\bar{r} = \sum_{b=1}^{ntT(n)}\sum_{h=1}^{h(b)}\frac{1}{H}r(b,h)$$

is the sample mean of hop-lengths over H hops. Rearranging we obtain

$$T(n) \leq \frac{c_3}{n\bar{l}(n)\bar{r}}. \tag{26}$$

Now for any cellular scheme requiring full connectivity, the hop distance is $\Omega\left(\sqrt{\frac{\log n}{n}}\right)$ and hence we obtain $T(n) = O\left(1/\bar{l}(n)\sqrt{n\log n}\right)$. □

Note that for Schemes 3(a) and 3(b) with parameter $a(n)$, the average hop distance $\bar{l}(n) = \Theta\left(\sqrt{a(n)}\right)$. Thus the above bound on throughput has the same order as the bounds in Theorems 5-6.

Optimality of Scheme 3(a): First consider the case when mobility is not used, i.e., $\bar{l}(n) = \Theta(1)$. In this case, from (26), we obtain, $T(n) \leq c/n\bar{r}$, and delay $D(n) = \Theta(1/\bar{r})$ which is only due to hops.

Now suppose mobility is used for the same throughput, i.e., $\bar{l}(n)\bar{r}$ remains of the same order in (26). If $\bar{l}(n) = \Theta(1)$, then the delay due to hopping is $\Theta(1/\bar{r})$ in addition to mobile-delay. This implies that, use of mobility will result in a worse trade-off. Thus, the use of mobility when $\bar{l}(n) = \Theta(1)$ does not help.

If $\bar{l}(n) = o(1)$, then the average distance traveled by a packet via node mobility is $\Theta(1)$. From condition 21, since $v(n) = o(\sqrt{\log n/n})$, the average mobile-delay is $\omega(\sqrt{n/\log n})$. Since $\bar{r} = \Omega(\sqrt{\log n/n})$, the hop-delay is $\Theta(\bar{l}(n)/\bar{r}) = o(\sqrt{n/\log n})$. Clearly the mobile-delay $\omega(\sqrt{n/\log n})$ dominates the hop-delay for any $\bar{l}(n)$.

From the above discussion, the optimal throughput-delay trade-off is bounded as

$$T(n) = O(D(n)/n), \text{ for } T(n) = O(1/\sqrt{n\log n}).$$

Since this throughput-delay trade-off is achieved by Scheme 3(a), it is optimal.

Optimality of Scheme 3(b): From (22), it is clear that achieving $T(n) = \omega(1/\sqrt{n\log n})$ requires that $\bar{l}(n) = o(1)$. But from the preceding discussion, for any $\bar{r}$, when $\bar{l} = o(1)$, the mobile-delay dominates the hop-delay. Thus, to maximize the throughput for a given delay, any optimal scheme must have $\bar{r} = \Theta(\sqrt{\log n/n})$. Therefore, for any optimal scheme, the throughput $T(n) = \Theta\left(1/\bar{l}(n)\sqrt{n\log n}\right)$.

Consider a throughput-delay optimal scheme with average hop distance $\bar{l}(n)$. For any such scheme, fixing a throughput $T(n)$, fixes $\bar{l}(n)$. The goal of an optimal scheme is to use hops to minimize the time for a packet to reach its destination.

Consider the transmission of a packet p starting from its source S and moving towards its destination D, initially at a distance d from S. Recall that a packet travels a distance $\bar{l} = \bar{l}(n)$ through hops and the rest through the motion of the nodes relaying it. Define t_p to be the time it takes the packet p, after leaving its source S, to reach its destination D. We ignore the time required for hops as the mobile delay dominates the total delay. Let $E[t_p]$ be the expectation of t_p for a given $\bar{l}$ and d. Note that, the expectation is over the distribution induced by random walks of the nodes.

We claim the following.

Lemma 8. *For any $\bar{l}$ and d, a scheme that minimizes $E[t_p]$ must perform all the hops the first time the packet is at a distance less than or equal to $\bar{l}$ from its destination D.*

Proof. For $\bar{l} \geq d$, the Lemma clearly holds. For $\bar{l} < d$, consider the following two schemes. Scheme A uses the entire hop distance $\bar{l}$ when the packet reaches within a distance $\bar{l}$ of D for the first time, which is consistent with the

claim of the lemma. Scheme B uses a hop of length ϵ when the packet is at a distance $\tilde{d}$ ($d > \tilde{d} > \bar{l}$) from D, and uses the remaining hop distance $\bar{l} - \epsilon$ at the end, as in Scheme A.

We want to show that, on average, a packet takes longer to reach D in Scheme B than in Scheme A. For simplicity, we assume that D is fixed. This does not affect generality as all nodes perform independent symmetric random walks.

Consider the path of a packet originating at distance d from D. Until the packet reaches within a distance $\tilde{d}$ of D, its path is the same in both schemes. As illustrated in Figure 6, under Scheme B, at point X, which is at a distance $\tilde{d}$ from D, the packet travels a distance ϵ by hops toward D to reach Y. Under Scheme A, the packet remains at point X. At this instant, the remaining time for the packet to reach D under Scheme A, t_A, is the time taken to reach a ball $B(D, \bar{l})$ starting from X, and under Scheme B, it is the time t_B taken to reach $B(D, \bar{l} - \epsilon)$, starting from Y. We now show that on average $t_{\rm A} < t_{\rm B}$. Consider a point D' on the line X–D at distance ϵ from D (as depicted in Figure 6). Since all nodes perform independent symmetric random walks, the probability that a path starting from X reaches $B(D, \bar{l})$ is the same as the probability that any path starting from Y reaches $B(D', \bar{l})$. Note that, by construction, $B(D, \bar{l} - \epsilon) \subset B(D', \bar{l})$. Hence the time for a packet at Y to reach $B(D', \bar{l})$ is stochastically dominated by the time needed to reach $B(D, \bar{l} - \epsilon)$. This proves that the time taken by Scheme A is strictly smaller than the time taken by Scheme B on average.

Using the above argument inductively for all hops establishes the lemma. □

The above lemma shows that a throughput-delay optimal scheme must utilize all the hops at the end. Since in Scheme 3(b) half the hops are performed at the end, it follows that its achievable throughput-delay trade-off is of the same order as that of an optimal scheme. This establishes the following theorem.

Theorem 7. *Scheme 3 obtains the optimal throughput-delay trade-off for mobile networks.*

V. Conclusion

The way throughput scales with the number of nodes in ad hoc fixed and mobile wireless networks has been well-studied. However, the way delay scales with the size of such networks has not been addressed previously. This paper provides a definition of delay in ad hoc networks and obtains optimal throughput-delay trade-off in fixed and mobile ad hoc networks. For the Gupta-Kumar fixed network model, we showed that the optimal throughput-delay trade-off is given by $D(n) = \Theta(nT(n))$. For the

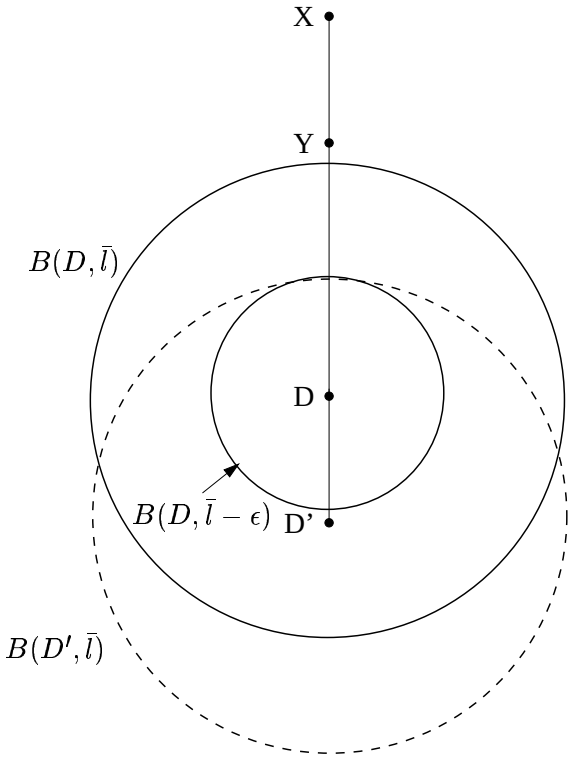

Fig. 6. Illustration for comparison of Schemes A and B.

Grossglauser-Tse mobile network model, we showed that the delay scales as $O\left(n^{1/2}/v(n)\right)$. For a mobile wireless network we described a scheme that achieves the optimal throughput-delay trade-off by varying the number of hops, the transmission range, and the degree of node mobility. The scheme captures the Gupta-Kumar model at one extreme and the Grossglauser-Tse model at the other. The proofs use a unified framework and simpler tools than used in previous work.

References

[1] N. Bansal and Z. Liu, "Capacity, Mobility and Delay in Wireless Ad hoc Networks", *In Proceedings of IEEE Infocom*, 2003.

[2] S. N. Diggavi, M. Grossglauser and D. Tse, "Even One-Dimensional Mobility Increases Ad Hoc Wireless Capacity" *In Proceedings of ISIT 2002*, Laussane, Switzerland, July 2002.

[3] R. Durrett, "Probability: Theory and Examples", 2nd edition, Duxbury Press, 1996.

[4] M. Grossglauser and D. Tse, "Mobility Increases the Capacity of Ad-hoc Wireless Networks", *IEEE INFOCOM*, Anchorage, Alaska, pp.1360-1369, 2001.

[5] P. Gupta and P. R. Kumar, "The Capacity of Wireless Networks", IEEE Trans. on Information Theory, 46(2), pp. 388-404, March 2000.

[6] S. R. Kulkarni and P. Viswanath, "A Deterministic Approach to Throughput Scaling in Wireless Networks", *Preprint:* `http://www.ifp.uiuc.edu/~pramodv/pubs/v021104.ps`.

[7] R. Motwani and P. Raghavan, "Randomized algorithms", *Cambridge Univ. Press*, 1995.

[8] R. W. Wolff, "Stochastic Modeling and the Theory of Queues", *Prentice-Hall*.