Throughput-Delay Tradeoffs in Content-Centric Wireless Networks

Milad Mahdian, Member, IEEE, Edmund Yeh, Senior Member, IEEE,

Abstract—We study the throughput and delay characteristics of wireless networks based on a content-centric network architecture, where users are mainly interested in retrieving content stored in the network, rather than in maintaining source-destination communication. Nodes are assumed to be uniformly distributed in the network area. Each node has a limited-capacity content store, which it uses to cache contents according to the proposed caching scheme. Requested content follows a general popularity distribution, and users employ multi-hop communication to retrieve the requested content from the closest cache. We derive the throughput-delay tradeoff of the content-centric wireless network model and solve the caching optimization problem. We then evaluate the network performance for a Zipf content popularity distribution, letting the number of content types and the network size both go to infinity. Finally, we extend our analysis to heterogeneous wireless networks where, in addition to wireless nodes, there are a number of base stations uniformly distributed at random in the network area. We show that in order to achieve a better performance in a heterogeneous network in the order sense, the number of base stations needs to be greater than the ratio of the number of nodes to the number of content types. Furthermore, we show that the heterogeneous wireless network does not yield performance advantages in the order sense if the Zipf content popularity distribution exponent exceeds 3/2.

I. INTRODUCTION AND RELATED WORK

Two fundamental trends in networking are: first, the bulk of network traffic today, and of its projected enormous growth, consists mainly of content disseminated to multiple users. Second, network content is accessed increasingly in wireless environments. A basic problem, of both theoretical and practical interest, is the characterization of performance and scaling in large-scale wireless networks for content distribution. This paper addresses this key question. We focus on the well-known random wireless network model, where nodes are uniformly distributed in a network area. Rather than assuming a wireless communication network consisting of source-destination pairs, however, we investigate a content-centric wireless infrastructure where users are mainly interested in retrieving content stored in the network. Combining caching schemes with shortest-path request forwarding, we derive the throughput-delay tradeoff of the content-centric wireless network and solve the caching optimization problem. We then extend our analysis to heterogeneous wireless networks with base stations as well as wireless nodes.

As the number of users of wireless technology continues to grow exponentially, the scaling behavior of wireless networks has been of wide interest. Gupta and Kumar [1] pioneered this study within the context of wireless communication networks consisting of source-destination pairs. They focus on a random network model where nodes are 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. Each node can simultaneously serve as a source, a destination, and as a relay for other source-destination pairs. It was shown [1] that the per-source-destination-pair throughput scales as $\Theta(1/\sqrt{n \log n})$ if the Zipf content popularity distribution exponent exceeds 3/2.

Unfortunately, host-based network designs such as IP cannot directly use caching as a network functionality, since a specific piece of data cannot have an independent existence outside a conversation between sources and destinations. In order for IP to exploit caching, content objects need to be mapped to host locations using complicated network configurations and services [8]. Recently, new content-centric networking architectures such as Named Data Networking (NDN) [8] and Content-Centric Networking (CCN) [9] have been developed to more directly enable efficient content distribution using caching.

Given the above, a natural and important problem is the characterization of performance and scaling in large-scale content-centric wireless networks with caching. The problem has received attention recently in [10], [11]. In [10], asymptotic properties of the joint delivery and replication problem in a static grid-based wireless network with multi-

---

1 We use the following notation. We say $f(n) = O(g(n))$ if there exists $n_0 > 0$ and a constant $M$ such that $f(n) \leq Mg(n)$ for all $n \geq n_0$. We say $f(n) = o(g(n))$ if for any constant $c > 0$ there exists $n(c) > 0$ such that $|f(n)| \leq cg(n)$ for all $n > n(c)$. We say $f(n) = \Omega(g(n))$ if $g(n) = \Theta(f(n))$, and $f(n) = \omega(g(n))$ if $g(n) = o(f(n))$. Finally, we say $f(n) = \Theta(g(n))$ if $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$.
hop communication and caching are presented. The objective here is the minimization of average link capacity subject to content replication constraints. Scaling laws for link capacities are derived, with the content popularity following a Zipf distribution.

The paper [11] derives the throughput and delay performance of content-centric mobile ad-hoc networks under various mobility models on a random geometric graph, for Zipf content popularity distributions. The paper makes the assumption that at any given time, each node has at most one pending content request in the network. It further considers a request model in which the relation between the throughput and delay is pre-determined as $\lambda = \frac{1}{I + D}$, where $\lambda$ is the average request throughput, $D$ is the average request delay, and $I$ is the average time between consecutive content requests [11].

In [12], the asymptotic throughput capacity of content-centric wireless networks is studied under the assumption that a constant number of content objects with similar popularity are requested and cached with limited lifetime by network users. By computing the average lifetime of the cached content objects of each user, the network throughput is derived for both the grid and random network models.

In [13], a content placement problem in a wireless femtocellular network using helper nodes is studied. The paper considers a one-hop communication scheme where nodes are connected to a set of helper nodes according to a bipartite graph. Each node is also connected to the base station. The paper focuses on the minimization of the average total downloading delay for a given content popularity distribution and network topology. The authors show that the uncoded optimal file assignment is NP-hard, and demonstrate a greedy strategy with performance which is provably within a factor 2 of the optimum.

The authors of [14] analyze base-station-assisted device-to-device wireless networks with caching capability. They examine a cellular grid network model in which communication among wireless nodes or between wireless nodes and the base station is limited to one hop, and derive the asymptotic throughput-outage tradeoff for the network model.

Finally, the paper [15] develops a systematic framework to solve the fundamental problem of jointly optimizing interest request forwarding and dynamic cache placement and eviction, for arbitrary network topologies and content popularity distributions.

In this paper, we characterize the throughput and delay scaling behavior of content-centric wireless networks, using the random geometric model as studied in [11, 5] and in many related papers (previously within the context of traditional source-destination communication networks). We assume that contents follow a general popularity distribution, and that each node has a limited-capacity content store, which it uses to cache contents according to a proposed caching scheme. Users employ multi-hop communication to retrieve the requested content from the closest cache. We first establish the throughput-delay trade-off of the network model as follows:

$$D(n) \lambda(n) = \Theta( (na(n))^{-1}) \quad \text{w.h.p.}$$

where $D$, and $\lambda$ are the network delay and per-node throughput, respectively, $n$ is the number of wireless nodes in the network, and $a(n)$ is the appropriately chosen cell size such that $a(n) = \Omega \left( \frac{\log n}{n} \right)$.

We optimize the caching strategy to simultaneously minimize the average network delay and maximize the network throughput. Using the optimal caching strategy, we evaluate the network performance under a Zipf content popularity distribution. Furthermore, we investigate heterogeneous wireless networks where, in addition to wireless nodes, there are a number of base stations uniformly distributed at random in the network area. We show that in order to achieve a better performance in a heterogeneous network in the order sense, the number of base stations needs to be greater than the ratio of the number of nodes to the number of content types. For the case where the number of content objects is greater than the number of wireless nodes, this condition reduces to having at least one base station in the network. In addition, we show that for the Zipf content popularity distribution with exponent $\alpha \geq 3/2$, the performance of the wireless ad hoc network is of the same order as for the heterogeneous wireless network, independent of number of base stations.

To the best of our knowledge, this paper is the first to characterize the performance of content-centric heterogeneous wireless networks with caching capability. Our focus on the random geometric network model allows for ready comparison with existing literature on the scaling behavior of traditional host-based wireless communication networks.

### II. NETWORK MODEL

We analyze a content-centric wireless network model where $n$ nodes are independently and uniformly distributed over a unit-sized torus. From these nodes originate requests for content objects. There are $M$ distinct content objects, where $M$ scales as $n^\beta$, $0 < \beta < 1$. Note that we assume $\beta < 1$ in order for the network to have sufficient memory to store at least one copy of each content object. All content objects are assumed to have the same size (one unit). Each node is assumed to have a local cache, named the Content Store, which can store copies of content objects. All Content Stores are assumed to have the same size: $K$ content units.

Assume that time is slotted: $t = 0, 1, 2, \ldots$. Independently from one another, all nodes generate requests for content objects at each time $t$. Assuming an infinite backlog of requests at each node, each content request is for content object $m, 1 \leq m \leq M$, with probability $p_m$, independent of all other requests. Due to the time invariant nature of the content request pattern, we consider static caching allocations. Let $\chi_m$ be the set of nodes which cache content object $m$ in their Content Store. We call the nodes in $\chi_m$ the holders of content $m$. Let $X_m = |\chi_m|$ be the number of holders of content object $m$. Since content popularity distribution is homogeneous in

---

2 We say an event holds with high probability (w.h.p.) if the event occurs with probability 1 as $n$ goes to infinity.
space, we may assume (without loss of optimality) that for any given content object $m$, the $X_m$ holders of the content object are uniformly and independently distributed over the network area.\footnote{One way to choose the $X_m$ holders is to choose one of the $\binom{M}{X_m}$ sets of $X_m$ nodes, uniformly at random.} In order for a caching allocation $\{X_m\}_{m=1}^M$ to be feasible, the constraint on total caching space must be satisfied: \begin{equation} \sum_{m=1}^{M} X_m \leq nK. \end{equation}

The total caching constraint in (2) is a relaxed version of the individual caching constraints. For ease of presentation and analysis, we use (2) for the throughput-delay analysis and optimization problem.

For concreteness, we consider the content delivery mechanism embodied in the NDN architecture \cite{8}. Specifically, requests for content objects are submitted using Interest Packets, which are forwarded toward Content Stores caching the requested content object. When the Interest Packet reaches a node caching the requested content object, a Data Packet containing the requested content object is transmitted in the reverse direction along the path taken by the corresponding Interest Packet, back to the requesting node.\footnote{Assume that routing (topology discovery and data reachability) has already been accomplished in the network, so that each node knows to which other nodes it can forward an Interest Packet to reach a Content Store caching the requested object. Equivalently, in an NDN network, the Forward Information Base (FIB) has already been populated at each node for each content object. Note that Interest Packets are usually much smaller in size than the corresponding Data Packet.} In this paper, we assume that an Interest Packet requesting object $m$ is forwarded along the shortest path from the requesting node toward the closest (in Euclidean distance) holder of content object $m$, using multi-hop communication.\footnote{If a node requests a content object which is cached in its local Content Store, the request can be satisfied immediately and there is no need to generate an Interest Packet. Since the Content Store has limited cache space, this is not usually the case. For ease of analysis, we assume in this paper that if the requested content is in the local cache, the node still generates an Interest Packet for it, transmits it to the nearest holder excluding itself, and uses the network to retrieve the content object.}

Transmissions in wireless networks are subject to multi-user interference. Our model for a successful wireless transmission is the Physical Model \cite{1}. Since these two models are essentially equivalent (assuming a path loss exponent of greater than 1 and equal node transmission powers in the Physical Model) \cite{1}, we focus on the Protocol Model in this paper.

To simplify our analysis, we adopt the fluid model for packet transmission considered in \cite{5}. In the fluid model, we allow the size of the content unit, and therefore the sizes of the Interest Packets and Data Packets, to be arbitrarily small. Thus, the time required for transmitting an Interest Packet or Data Packet is much smaller than a time slot. Nevertheless, a packet received by a node in a given time slot cannot be transmitted by the node until the next time slot. Thus, all packets waiting for transmission at a given node will be transmitted by the node in one time slot. The fluid model makes unnecessary detailed analysis of the scheduling of individual packets.

## III. Throughput and Delay

Transmission and caching in the wireless network are coordinated and controlled by a scheme. More precisely, a scheme $\pi$ is a sequence of policies $\{\pi_n\}$, where $\pi_n$ determines the (static) caching allocation, as well as the scheduling of transmissions in each time slot, for a network of $n$ nodes. For a given scheme, the throughput and delay are defined as follows:

**Definition 1** (Throughput). For a given scheme $\pi_n$, let $B_{\pi_n}(i,t)$ be the total number of bits of all content objects received by the requesting node $i$ up to time $t$. The long-term throughput of node $i$ is \begin{equation} \lim_{t \to \infty} \frac{1}{t} B_{\pi_n}(i,t). \end{equation}

The average throughput over all nodes is \begin{equation} \lambda_{\pi_n}(n) = \frac{1}{n} \sum_{i=1}^{n} \lim_{t \to \infty} \frac{1}{t} B_{\pi_n}(i,t). \end{equation}

The throughput of $\pi_n$, is defined as the expectation over all realizations of node positions $\{U_1,U_2,\ldots,U_n\}$, of the corresponding average throughput:

\begin{equation} \lambda_{\pi_n}(n) \triangleq E\left[ \lambda_{\pi_n}(n) \right]. \end{equation}

**Definition 2** (Delay). For a given $\pi_n$, let $D_{\pi_n}(i,k)$ be the delay of the $k$-th request for any content object by node $i$ (measured from the moment the Interest Packet leaves $i$ for the closest holder until the corresponding Data Packet arrives at $i$ from the holder). The delay (over all content requests) for node $i$ is \begin{equation} \lim_{r \to \infty} \sup_{k=1}^{r} \frac{1}{r} \sum_{k=1}^{r} D_{\pi_n}(i,k). \end{equation}

The average delay over all nodes is \begin{equation} D_{\pi_n}(n) = \frac{1}{n} \sum_{i=1}^{n} \lim_{r \to \infty} \sup_{k=1}^{r} \frac{1}{r} \sum_{k=1}^{r} D_{\pi_n}(i,k). \end{equation}

The delay of $\pi_n$ is defined as the expectation over all realizations of node positions $\{U_1,U_2,\ldots,U_n\}$, of the corresponding average delay:

\begin{equation} D_{\pi_n}(n) \triangleq E\left[ D_{\pi_n}(n) \right]. \end{equation}

The throughput and delay quantities $\lambda_{\pi_n}(n)$ and $D_{\pi_n}(n)$ are random variables, since they depend on the realization of node positions. The quantities $\lambda_{\pi_n}(n)$ and $D_{\pi_n}(n)$ are ensemble averages. To study the asymptotical behavior of $\lambda_{\pi_n}(n)$ and $D_{\pi_n}(n)$, we will let the number of nodes $n$ go to infinity.

To study the throughput-delay tradeoff of the content-centric wireless network, we use a multi-hop, TDM, cellular
transmission scheme combined with a caching allocation scheme. The unit torus is divided into square cells, each with area \(a(n)\). We use the following sequence of lemmas to construct the transmission and caching scheme yielding the main throughput-delay tradeoff result.

The following lemma from [5] shows that with an appropriately chosen cell area \(a(n)\), each cell has at least one node w.h.p., so that multi-hop relaying of packets through adjacent cells is possible.

**Lemma 1.** [5] If \(a(n) \geq 2 \log n / n\), then each cell has at least one node w.h.p.

For \(a(n)\) satisfying Lemma 1 we set the transmission radius to be \(r(n) = \sqrt{8a(n)}\). This allows each node to transmit to nodes within its cell and to the 8 neighboring cells. It is then clear that multi-hop packet relaying through adjacent cells can take place w.h.p.

The next lemma from [5] makes possible the establishment of an interference-free TDM transmission schedule where each cell becomes active (i.e. any of the nodes in the cell transmits) regularly once every \(N + 1\) time slots, where \(N\) is specified in Lemma 2 and no two simultaneously active cells interfere with each other. Here, two simultaneously active cells interfere if the transmission of a node in one active cell affects the success of a simultaneous transmission by a node in the other active cell.

**Lemma 2.** [5] Under the Protocol model, the number of cells that interfere with any given cell is bounded above by a constant \(N = 16(1 + \Delta)^2\), independent of \(n\).

Recall that from any given node, requests for content \(m\) are forwarded to the holder of content \(m\) which is closest in Euclidean distance. The next lemma computes the expected Euclidean distance from a given node requesting content object \(m\) to the closest holder of content \(m\).

**Lemma 3.** Let \(\chi_m\) be the set of holders of content \(m\), uniformly distributed in the unit-sized network area, where \(X_m = |\chi_m|\). For any node requesting content \(m\), the average Euclidean distance from the requesting node to the closest holder of content \(m\) is \(\Theta(\frac{1}{\sqrt{X_m}})\).

**Proof:** Since the holders are independently and uniformly distributed, the probability that no holder is within distance less than or equal to \(\tau\) of the requester is \(\text{Pr}(d \geq \tau) = (1 - \pi \tau^2)^{X_m}\) for \(0 \leq \tau \leq 1 / \sqrt{\pi}\). Therefore, the average distance from the requester to the closest holder is

\[
E[d] = \int_0^\infty \text{Pr}(d \geq \tau) d\tau = \int_0^{1 \over \sqrt{\pi}} (1 - \pi \tau^2)^{X_m} d\tau.
\]

Using a change of variable \(\sqrt{\pi} \tau = \cos \theta\) and applying integration by parts, we have

\[
E[d] = \frac{1}{\sqrt{\pi}} \int_0^\pi (\sin \theta)^{2X_m+1} d\theta
\]

\[
= \frac{1}{\sqrt{\pi}} \frac{2X_m}{2X_m + 1} \frac{2X_m - 2}{2X_m - 1} \cdots \frac{2}{3} \int_0^\pi \sin \theta d\theta
\]

\[
= \frac{1}{\sqrt{\pi}} \frac{2X_m}{2X_m + 1} \frac{2X_m - 2}{2X_m - 1} \cdots \frac{2}{3}
\]

\[
= \Theta(\frac{1}{\sqrt{X_m}}).
\]

where (3) is derived from

\[
\int \sin^n x \, dx = -\frac{1}{n} \sin^{n-1} x \cos x + \frac{n-1}{n} \int \sin^{n-2} x \, dx.
\]

and (5) is proved in Appendix A. Assume \(a(n) \geq 2 \log n / n\) and \(r(n) = \sqrt{8a(n)} \geq 4\sqrt{\log n} / n\). Consider a fixed node \(i\) requesting content object \(m\). Let \(L_{H,R}(i,m)\) be the straight line connecting \(i\) to the closest holder of content \(m\). From Lemma 3

\[
E[|L_{H,R}(i,m)|] = \Theta \left( \frac{1}{\sqrt{X_m}} \right).
\]

where \(|L|\) denotes the Euclidean length of line \(L\). Let \(H_i,m\) be the number of hops along a path (sequence of nodes) which belongs to the set of cells intersecting the \(L_{H,R}(i,m)\) line between requester \(i\) and the closest holder of content \(m\), where there is exactly one node per cell along the path. The following lemma characterizes \(E[H_i,m]\) as \(n\) becomes large.

**Lemma 4.** For \(a(n) \geq 2 \log n / n\), and each \(m = 1, \ldots, M\),

\[
E[H_i,m] = \Theta \left( \frac{1}{\sqrt{a(n)X_m}} \right) \text{ w.h.p.}
\]

**Proof:** To compute \(E[H_i,m]\), we consider the case where the holder is within one hop of the requester, and the case where the holder is farther than one hop away. We have

\[
E[H_i,m] = E[H_i,m]|L_{H,R}(i,m)| \leq \sqrt{a(n)} \text{ Pr}(|L_{H,R}(i,m)| \leq \sqrt{a(n)})
\]

\[
+ E[H_i,m]|L_{H,R}(i,m)| > \sqrt{a(n)} \text{ Pr}(|L_{H,R}(i,m)| > \sqrt{a(n)})
\]

Clearly, \(E[H_i,m]|L_{H,R}(i,m)| \leq \sqrt{a(n)} = 1\). Also, since the side-length of each cell is \(\sqrt{a(n)}\), it can be shown that \(E[H_i,m]|L_{H,R}(i,m)| > \sqrt{a(n)} = \Theta(\sqrt{a(n)})\).

Letting \(\alpha(n) = \text{ Pr}(|L_{H,R}(i,m)| > \sqrt{a(n)})\), it follows that

\[
E[H_i,m] = \Theta \left( 1 + \frac{1}{\sqrt{a(n)X_m}} - 1 \right) \alpha(n).
\]

Note that \(\alpha(n) = \text{ Pr}(d > \sqrt{a(n)}) = (1 - \pi \alpha(n))^{X_m}\).

Expanding \(\alpha(n)\) using the binomial form, and noting that \(\binom{n}{k} \leq \binom{\alpha(n)}{k} \leq n^k\), for \(n \geq k \geq 1\), we have

\[
1 + \sum_{i=1}^{X_m} \frac{(\pi a(n)X_m)^i}{i!} \leq \alpha(n) \leq e^{-\pi a(n)X_m}.
\]
Now, as \( n \to \infty \), for \( X_m = \omega(1/a(n)) \), \( e^{-\alpha(n)X_m} \to 0 \), and hence \( \alpha(n) \to 0 \), implying that \( E[H_{i,m}] = 1 \). For \( X_m = \Theta(1/a(n)) \), both bounds in (10) and consequently \( \alpha(n) \), are constant, leading to \( E[H_{i,m}] = \Theta(1) \). On the other hand, for \( X_m = o(1/a(n)) \), \( a(n)X_m \to 0 \), resulting in both bounds in (10) converging to 1, as \( n \to \infty \). Substituting \( \alpha(n) = 1 \) in (10) gives \( E[H_{i,m}] = \Theta(1) \). Therefore, the average number of hops can be re-written as

\[
E[H_{i,m}] = \Theta \left( \max \left\{ \frac{1}{\sqrt{a(n)X_m}}, 1 \right\} \right) \text{ w.h.p.} \tag{11}
\]

We now prove a key lemma, characterizing the number of \( L_{H,R}(i, m) \) lines passing through each cell as \( n \) becomes large. The result may be seen as an analogue of Lemma 3 in [5] for the content-centric wireless environment.

**Lemma 5.** For \( a(n) \geq 2\log n/n \), the number of \( L_{H,R} \) lines passing through each cell is

\[
\Theta \left( n \sum_{m=1}^{M} p_m \max \left\{ \frac{1}{\sqrt{a(n)X_m}}, 1 \right\} \right) \text{ w.h.p.}
\]

**Proof:** For a given content request vector \((m_1, m_2, \ldots, m_n)\) at time \( t \) and a given node \( i \), we know that \( H_{i,m_1} = H_{i,m} \), w.p. \( p_m \), for \( m = 1, 2, \ldots, M \). Therefore,

\[
E[H_{i,m}] = \sum_{m=1}^{M} p_m E[H_{i,m}]
\]

\[
= \Theta \left( \sum_{m=1}^{M} p_m \max \left\{ \frac{1}{\sqrt{a(n)X_m}}, 1 \right\} \right) \tag{12}
\]

There are \( 1/a(n) \) cells. Fix a cell \( j \) and let \( Y_{i,m} \) be the indicator of the event that the \( L_{H,R}(i, m) \) line passes through cell \( j \). That is,

\[
Y_{i,m} = \begin{cases} 1, & \text{if } L_{H,R}(i, m) \text{ passes through cell } j \\ 0, & \text{otherwise} \end{cases}
\]

for \( 1 \leq i \leq n, 1 \leq j \leq 1/a(n) \) and \( 1 \leq m \leq M \). We know that \( Y_{i,m} = Y_{j,m} \), w.p. \( p_m \), for \( m = 1, 2, \ldots, M \). Hence, we obtain \( E[Y_{i,m}] = \sum_{m=1}^{M} p_mE[Y_{j,m}] \). Summing up the total number of hops for any \( m \) in two different ways gives us:

\[
\sum_{i=1}^{n} \sum_{j=1}^{1/a(n)} Y_{i,m} = \sum_{i=1}^{n} H_{i,m}. \tag{13}
\]

Taking the expectation on the both sides of (13), and noting that \( E[H_{i,m}] \) is the same for each node \( i \) and \( E[Y_{i,m}] \) is equal for every \( i \) and \( j \) due to symmetry of the torus, we have

\[
\sum_{i=1}^{n} \sum_{j=1}^{1/a(n)} E[Y_{i,m}] = \sum_{i=1}^{n} E[H_{i,m}].
\]

\[
nE[Y_{i,m}]/a(n) = nE[H_{i,m}].
\]

Therefore,

\[
E[Y_{i,m}] = a(n) \cdot E[H_{i,m}]
\]

\[
= \Theta \left( \max \{ \sqrt{a(n)/X_m}, a(n) \} \right). \tag{14}
\]

Now,

\[
E[Y_{i,m}] = \sum_{m=1}^{M} p_mE[Y_{i,m}]
\]

\[
= \Theta \left( \sum_{m=1}^{M} p_m \max \{ \sqrt{a(n)/X_m}, a(n) \} \right). \tag{15}
\]

The average total number of \( L_{H,R} \) lines passing through a fixed cell \( j \), is given by \( Y = \sum_{i=1}^{n} Y_{i,m} \). Hence, \( E[Y] = \Theta(n \sum_{m=1}^{M} p_m \max \{ \sqrt{a(n)/X_m}, a(n) \}) \). Note that nodes are uniformly and independently distributed in the network and \( p_m \) is invariant for all nodes. Therefore, \( Y \) is the sum of independent random variables and \( 0 \leq Y_{i,m} \leq 1 \) for each \( i \) and \( j \). Applying the Chernoff bound yields [17]

\[
P\{Y > (1+\delta)E[Y]\} \leq \exp \left( -\frac{\delta^2E[Y]}{3} \right). \tag{16}
\]

Choosing \( \delta = \sqrt{6 \log n/E[Y]} \), we are guaranteed that \( \delta = o(1) \). This is true as we are assuming that \( a(n) = \Omega(\log n/n) \). Also, as explained later, there is no need for any content object to have more than \( \Theta(1/a(n)) \) holders. Due to the total caching capacity constraint, \( \sum_{m=1}^{M} X_m \leq nK \), and the fact that \( M = \Theta(n^2) \), where \( 0 < \beta < 1 \), we are assured that \( E[Y] = \omega(na(n)) \), or equivalently, \( E[Y] = \omega(\log n) \), resulting in \( \delta = o(1) \). Substituting \( \delta \) in (16), we have

\[
P\{Y > (1+\delta)E[Y]\} \leq 1/n^2. \tag{17}
\]

Therefore, \( Y = O(E[Y]) \) with probability \( \geq 1-1/n^2 \). Similarly, by applying the Chernoff bound to the lower tail [17], we have

\[
P\{Y < (1-\delta)E[Y]\} \leq \exp \left( -\frac{\delta^2E[Y]}{2} \right). \tag{18}
\]

Applying similar techniques as above, we can show that \( Y = \Omega(E[Y]) \) with probability \( \geq 1-1/n^2 \). Now applying the union bound over all \( 8/r^2(n) \) cells, we see that the number of \( L_{H,R} \) lines passing through each cell of the network is

\[
\Theta(E[Y]) = \Theta \left( \sum_{m=1}^{M} p_m \max \{ \sqrt{a(n)/X_m}, a(n) \} \right)
\]

with probability \( \geq 1-1/n \).

We now present an achievable transmission and caching scheme. The achievable scheme can be seen as an analogue of Scheme 1 in [5], for the content-centric wireless environment. The scheme is parameterized by the cell area \( a(n) \), where \( a(n) = \Omega(\log n/n) \) and \( a(n) \leq 1 \).

1) Divide the unit torus using a square grid into square cells, each with area \( a(n) \).
2) For the given realization of the random network, check if the following conditions are satisfied:

- Condition 1: No cell is empty.
Condition 2: The number of $L_{H,R}$ lines passing through each cell is of the order
\[
\Theta \left( n \sum_{m=1}^{M} p_m \max \{ \sqrt{a(n)/X_m}, a(n) \} \right).
\]

3) If either of the above conditions is not satisfied, then use a time-division policy, where each of the $n$ requesters communicates directly with the closest holder of the requested content object, in a round-robin fashion.

4) If both conditions are satisfied, use the following policy $\pi_n$:

a) Each cell becomes active regularly once every $1 + N$ time-slots (Lemma 2). Cells which are sufficiently far apart become active simultaneously. That is, the scheme uses TDM between neighboring cells.

b) Requesting nodes transmit Interest Packets to the closest holders by hops along the adjacent cells intersecting the $L_{H,R}$ lines. Similarly, the holders transmit Data Packets to the requesting nodes along the same path taken by their corresponding Interest Packets, in the reverse direction.

c) Each time slot is split into two sub-slots. In the first sub-slot, each active cell transmits a single Interest Packet for each of the $L_{H,R}$ lines passing through the cell toward the closest holder. In the second sub-slot, the active cell transmits a single Data Packet for each of the $L_{H,R}$ lines passing through the cell toward the requesting node.

We now derive the throughput and delay performance of the achievable transmission and caching scheme described above.

**Theorem 1.** For $a(n) \geq 2 \log n/n$, the throughput and delay of the achievable scheme, and their trade-off, are given by
\[
\lambda(n) = \Theta \left( \frac{1}{n \sum_{m=1}^{M} p_m \max \{ \sqrt{a(n)/X_m}, a(n) \} } \right) \text{ w.h.p.} \tag{19}
\]
\[
D(n) = \Theta \left( \sum_{m=1}^{M} p_m \max \left\{ \frac{1}{\sqrt{a(n)/X_m}}, 1 \right\} \right) \text{ w.h.p.} \tag{20}
\]
\[
D(n)\lambda(n) = \Theta((na(n))^{-1}) \text{ w.h.p.} \tag{21}
\]

**Proof:** First note that if the time-division policy with direct communication is used, then the throughput is $W/n$ with a delay of 1. But since this happens with a vanishingly low probability, as shown by Lemmas 1 and 5, the throughput and delay for the achievable scheme are determined by that of policy $\pi_n$. When policy $\pi_n$ is used, since Condition 1 is satisfied, each cell has at least one node w.h.p. This assures us that requestor-holder pairs can communicate with each other by hops along adjacent cells on their $L_{H,R}$ lines w.h.p. From Lemma 2, each cell gets to transmit packets every $1 + N$ time-slots. Hence, the cell throughput is $\Theta(1)$. The total traffic through each cell is due to all the $L_{H,R}$ lines passing through the cell, which is $\Theta(n \sum_{m=1}^{M} p_m \max \{ \sqrt{a(n)/X_m}, a(n) \})$ w.h.p., since Condition 2 is also satisfied. This shows that
\[
\lambda(n) = \Theta \left( \frac{1}{n \sum_{m=1}^{M} p_m \max \{ \sqrt{a(n)/X_m}, a(n) \} } \right) \text{ w.h.p.} \tag{22}
\]

Recall that by Lemma 2 each cell can be active once every $N + 1$ time-slots, where $N$ is constant and independent of $n$. As we are assuming that packets scale in proportion to the throughput $\lambda(n)$ (fluid model), each packet arriving at a node in the cell departs in the next active time-slot of the cell. Hence, the packet delay is $N + 1$ times the number of hops from the requester to the holder. For a given realization of the random network, where node $i$ is requesting $m_i$ for $i = 1, 2, \ldots, n$, and $m_i \in \{1, 2, \ldots, M\}$, let $h_{i,m_i}$ be the number of hops from the requester $i$ to its closest holder of content $m_i$ in the given realization. Furthermore, since the Data Packet takes the same path as the corresponding Interest Packet in reverse, the average delay of the network realization is given by two times the mean sample of the $h_{i,m_i}$’s, i.e.
\[
\frac{2}{n} \sum_{i=1}^{n} h_{i,m_i} \approx 2E[H_{i,m_i}]. \tag{23}
\]

Using (12), equation (20) follows. Finally, combining (20) and (19) gives us the throughput-delay tradeoff (21).

Note that the throughput-delay trade-off for the achievable scheme holds for any feasible set $(X_m)_{m=1}^{M}$. In the next section, we find the optimized set $(X_m)_{m=1}^{M}$ which minimizes the delay and maximizes the throughput.

**IV. Optimized Caching**

We now optimize the delay and throughput of the achievable transmission and caching scheme described in Section III by selecting the appropriate $(X_m)_{m=1}^{M}$ subject to caching constraints. We first relax the integer constraint on $(X_m)_{m=1}^{M}$, thus allowing $X_m$ to be a non-negative real number. Furthermore, we enforce only the total caching constraint in (2) which is a relaxation of the per node caching constraint.

To illustrate the optimization process, we focus on the commonly used Zipf distribution as the content popularity distribution $[10]$, $[11]$. Let $p_m = m^{-\alpha}/H_\alpha(M)$, where $\alpha$ is the Zipf’s law exponent, and $H_\alpha(M) = \sum_{i=1}^{M} i^{-\alpha}$ is a normalization constant, given by $[10]$

\[
H_\alpha(M) = \begin{cases} 
\Theta(1), & \alpha > 1 \\
\Theta(\log M), & \alpha = 1 \\
\Theta(M^{1-\alpha}), & \alpha < 1 
\end{cases} \tag{24}
\]

As can be seen, for the case $X_m = \Omega(a^{-1}(n))$, $\lambda(n)$ and $D(n)$ are independent of the number of holders. Hence, there is no need to cache more than one copy of any given content object in any one cell. Also, note that by (21), minimizing the delay is equivalent to maximizing the throughput. We

It can easily be shown that the integer constraint relaxation does not change the order of the optimal delay and throughput scaling.
may obtain the minimum delay by solving the following optimization problem:

\[
\begin{align*}
\min \{ x_m \} \quad & \sum_{m=1}^{M} \frac{p_m}{\sqrt{a(n)x_m}} \\
\text{subject to:} & \\
\sum_{m=1}^{M} x_m & \leq nK \\
1 & \leq x_m \leq a^{-1}(n) \quad \text{for } m = 1, 2, \ldots, M
\end{align*}
\]  

(25)

As the objective function is strictly convex, we are assured that there is a unique global minimum. Defining the non-negative Lagrange multipliers \( \lambda \) for the constraint \( \sum_{m=1}^{M} x_m \leq nK \), and taking into account the constraint \( 1 \leq x_m \leq a^{-1}(n) \), the necessary conditions for a minimum of \( D \) with respect to \( x_m \), \( \forall m \in M \) are given by

\[
\frac{\partial D}{\partial x_m} \begin{cases} 
-\lambda & \text{if } x_m = a^{-1}(n) \\
-\lambda & \text{if } 1 < x_m < a^{-1}(n) \\
\geq \lambda & \text{if } x_m = 1 
\end{cases}
\]  

(26)

For the Zipf distribution, it is clear that \( p_m \) is strictly decreasing in \( m \) and therefore so is \( x_m \). Hence, let \( M_1 = \{1, 2, \ldots, m_1 - 1\} \) be the set of content objects such that \( x_m = a^{-1}(n) \) for \( m \in M_1 \). Similarly, let \( M_2 = \{m_1, m_1 + 1, \ldots, m_2 - 1\} \) and \( M_3 = \{m_2, m_2 + 1, \ldots, M\} \) be the set of contents such that \( 1 < x_m < a^{-1}(n) \) for \( m \in M_2 \), and \( x_m = 1 \) for \( m \in M_3 \), respectively. From (26), we have \( \forall m \in M \)

\[
\frac{p_m}{2\sqrt{a(n)x_m}} \begin{cases} 
\geq \lambda & \forall m \in M_1 \\
= \lambda & \forall m \in M_2 \\
\leq \lambda & \forall m \in M_3
\end{cases}
\]  

(27)

Using the equality for the case \( m \in M_2 \), we obtain

\[
\frac{m_1}{m_2} \simeq (a(n))^{2/3}. 
\]  

(28)

Clearly from (27), we have \( \lambda > 0 \) and hence, \( \sum_{m=1}^{M} x_m = nK \). Combining this with (28), we can derive \( m_1 \) and \( m_2 \). The optimal number of holders of content \( m \), \( X_m^* \), is then given by

\[
X_m^* = \begin{cases} 
a^{-1}(n), & m = 1, 2, \ldots, m_1 - 1 \\
\frac{p_{m_1/3}}{\sum_{j=m_1}^{M} p_j} K', & m = m_1, \ldots, m_2 - 1 \\
1, & m = m_2, \ldots, M
\end{cases}
\]  

(29)

where \( K' = K - (m_1 - 1)a^{-1}(n) - \frac{(M-m_2+1)}{n} \). The average delay is then w.h.p.:

\[
D^*(n) = \Theta \left( \frac{\sum_{j=1}^{m_1} p_j + \left( \sum_{j=m_1}^{M} p_j/3 \right)^{3/2}}{\sqrt{nK'a(n)}} + \frac{\sum_{j=m_2}^{M} p_j}{\sqrt{a(n)}} \right).
\]  

(30)

To gain more insight on the structure of the optimal solution, we have the following lemma.

**Lemma 6.** As \( n \to \infty \), the scaling of indices \( m_1 \) and \( m_2 \) is given by

\[
m_1 = \begin{cases} 
\Theta(\min\{M, na(n)\}), & \alpha > 3/2 \\
\Theta(\min\{M, na(n)/\log n\}), & \alpha = 3/2 \\
\Theta(\max\{1, \min\{M, na(n), \frac{na(n)}{M^{\frac{1}{\alpha}} - 1}\}\}), & \alpha < 3/2
\end{cases}
\]  

(31)

\[
m_2 = \begin{cases} 
\min\{M + 1, 2\alpha - 3/2a \sqrt{nK(a(n))^{1 - \frac{1}{3\alpha}}}\}, & \alpha > 3/2 \\
M + 1, & \alpha \leq 3/2
\end{cases}
\]  

(32)

**Proof:** As \( M = o(n) \), then \( M - m_2 = o(n) \). Therefore, \( K' \to K - (m_1 - 1)a^{-1}(n)/n \) as \( n \to \infty \). Clearly, \( K' = \Theta(1) \), hence, \( m_1 = O(na(n)) \). Now, by definition, \( m_1 \) is the smallest index for which the number of holders is less than \( a^{-1}(n) \). That is, \( X_m < a^{-1}(n) \). Using (29), it follows that

\[
nK'a(n) < m_1^{2/3} \left[ H_{2\alpha}(m_2 - 1) - H_{2\alpha}(m_1 - 1) \right].
\]  

(33)

Now, if \( m_1 > 1 \), attempting to decrease the index \( m_1 \) by one would result in

\[
\frac{2/3}{\sum_{j=m_1}^{M} p_j^{2/3}} nK' \geq a^{-1}(n).
\]

Hence, we have

\[
nK'a(n) \geq (m_1 - 1)^{2/3} \left[ H_{2\alpha}(m_2 - 1) - H_{2\alpha}(m_1 - 2) \right].
\]  

(34)

Hence, for \( m_1 > 1 \), an approximation of \( m_1 \) can be obtained from:

\[
nK'a(n) \simeq m_1^{2/3} \left[ H_{2\alpha}(m_2 - 1) - H_{2\alpha}(m_1 - 1) \right].
\]  

(35)

Similarly, by the definition of \( m_2 \), we know \( X_{m_2-1} > 1 \)

\[
nK' > (m_2 - 1)^{\frac{2\alpha}{3}} \left[ H_{2\alpha}(m_2 - 1) - H_{2\alpha}(m_1 - 1) \right].
\]  

(36)

Now if \( m_2 \leq M \), attempting to increase the index \( m_2 \) by one would lead to

\[
\frac{2/3}{\sum_{j=m_2}^{M} p_j^{2/3}} nK' < 1.
\]

Thus, it follows that

\[
nK' \leq m_2^{2/3} \left[ H_{2\alpha}(m_2 - 1) - H_{2\alpha}(m_1 - 1) \right].
\]  

(37)

Therefore, for \( m_2 \leq M \), \( m_2 \) can be computed approximately by:

\[
nK' \simeq (m_2 - 1)^{\frac{2\alpha}{3}} \left[ H_{2\alpha}(m_2 - 1) - H_{2\alpha}(m_1 - 1) \right].
\]  

(38)

For \( \alpha > 3/2 \): Using (35), we have

\[
n(a(n)K - (m_1 - 1) \simeq (m_1 - 1)^{\frac{2\alpha}{3}} \frac{[-(m_1 - 1)^{1 - \frac{2\alpha}{3}}]}{1 - \frac{2\alpha}{3}}.
\]  

(39)

which leads to

\[
m_1 \simeq 1 + \frac{2\alpha - 3}{2\alpha} na(n)K.
\]  

(40)

Now if \( m_2 \leq M \), following (28), we have

\[
m_2 \simeq m_1(a(n))^{-\frac{3}{3\alpha}} \simeq \frac{2\alpha - 3}{2\alpha} nK(a(n))^{1 - \frac{3}{3\alpha}}.
\]  

(41)
For $\alpha = 3/2$: Assuming $m_2 \leq M$, and by using (38) and (28), we have
\[ m_2 - 1 \simeq \frac{nK - (m_1 - 1) a^{-1}(n)}{\log m_2}. \] (42)
It follows that
\[ m_2 - 1 \simeq \frac{nK}{\log m_2}. \] (43)
This contradicts $m_2 = O(n^\beta)$, where $\beta < 1$. Hence $m_2 = M + 1$. Assuming $m_1 > 1$, and using (35), we have
\[ m_1 - 1 \simeq \frac{nKa(n) - (m_1 - 1)}{\log m_2}. \] (44)
resulting in $m_1 = \Theta\left(\frac{nKa(n)}{\log m_2}\right)$.

For $\alpha < 3/2$: Assuming $m_2 \leq M$, and by using (38), it follows that
\[ m_2 - 1 \simeq \frac{nK - (m_1 - 1)}{1 - 2\alpha/3} \simeq nK^\prime. \] (45)
Clearly, this contradicts the $m_2 \leq M$ assumption. Therefore, $m_2 = M + 1$. Now using (35) we have
\[ (m_1 - 1) \frac{2}{\alpha} \simeq \frac{nKa(n) - (m_1 - 1)}{H_{\frac{2}{\alpha}}(m_2 - 1) - H_{\frac{2}{\alpha}}(m_1 - 1)} \simeq \frac{nKa(n)}{M^{1-2\alpha/3}}. \] (46)
leading to $m_1 = \Theta\left(\frac{nKa(n)}{M^{\frac{2}{\alpha}-1}}\right)$.

We can now compute the optimized delay and throughput for the achievable scheme, assuming $M = \Theta(n^\beta)$ where $0 < \beta < 1$, under the Zipf popularity distribution.

**Theorem 2.** For $a(n) \geq 2 \log n/n$, the throughput and delay of the proposed scheme using Zipf distribution are w.h.p.:

\[
D^*(n) = \begin{cases}
\Theta(1), & \alpha > 3/2 \\
\Theta\left(\frac{1}{a(n)}\right), & \alpha = 3/2 \\
\Theta\left(\frac{1}{\sqrt{\log M a(n)}}\right), & 1 < \alpha < 3/2 \\
\Theta\left(\frac{1}{\sqrt{M a(n)}}\right), & \alpha = 1 \\
\Theta\left(\frac{1}{\sqrt{M a(n)}}\right), & \alpha < 1
\end{cases}
\] (47)

\[
\lambda^*(n) = \begin{cases}
\Theta\left(\frac{1}{a(n)}\right), & \alpha > 3/2 \\
\Theta\left(\frac{1}{\sqrt{\log M a(n)}}\right), & \alpha = 3/2 \\
\Theta\left(\frac{1}{\sqrt{M a(n)}}\right), & 1 < \alpha < 3/2 \\
\Theta\left(\frac{1}{\sqrt{M a(n)}}\right), & \alpha = 1 \\
\Theta\left(\frac{1}{\sqrt{M a(n)}}\right), & \alpha < 1
\end{cases}
\] (48)

**Proof:** Please see Appendix B.

To get more intuition about these results, we can substitute $a(n) = 2 \log n/n$, in (47) and (48). We have
\[
D^*(n) = \begin{cases}
\Theta(1), & \alpha > 3/2 \\
\Theta\left(\frac{1}{\log M}\right), & \alpha = 3/2 \\
\Theta\left(\frac{1}{\sqrt{\log M}}\right), & 1 < \alpha < 3/2 \\
\Theta\left(\frac{1}{\sqrt{M}}\right), & \alpha = 1 \\
\Theta\left(\frac{1}{\sqrt{M}}\right), & \alpha < 1
\end{cases}
\] (49)
\[
\lambda^*(n) = \begin{cases}
\Theta\left(\frac{1}{a(n)}\right), & \alpha > 3/2 \\
\Theta\left(\frac{1}{\sqrt{\log M a(n)}}\right), & \alpha = 3/2 \\
\Theta\left(\frac{1}{\sqrt{M a(n)}}\right), & 1 < \alpha < 3/2 \\
\Theta\left(\frac{1}{\sqrt{M a(n)}}\right), & \alpha = 1 \\
\Theta\left(\frac{1}{\sqrt{M a(n)}}\right), & \alpha < 1
\end{cases}
\] (50)

In Figure 1 we have plotted the theoretical results given in (49), to demonstrate the scaling of the network delay under the achievable scheme for various values of $\alpha$. Similarly, in Figure 2 the scaling of per-node throughput of the achievable scheme is depicted, following (50), for different values of $\alpha$. For both figures, $\beta = 0.9$. It is clear from the figures that the throughput and delay advantages offered by caching are both enhanced as $\alpha$ increases. This makes intuitive sense: for larger $\alpha$, the content popularity distribution is dominated by the most popular content objects, which when cached, can lead to increased throughput and decreased delay at the same time.

**V. HETEROGENEOUS WIRELESS NETWORKS**

Thus far, we have considered a pure ad hoc wireless network with caching, in which there are no base stations. We now consider a more general heterogeneous wireless network environment with caching. Consider a heterogeneous wireless network where, in addition to uniformly distributed wireless nodes, there are a number of base stations which are also uniformly distributed at random in the network area. This
models the scenario where smaller cells, e.g. femtocells, are deployed with random placement of base stations inside the network area [16]. The base stations are distinguished from the wireless nodes in that they are assumed to connect to the wired backbone, and thus are assumed to have access to all $M$ content objects. Let $f(n)$ be the number of base stations, where $f(n)$ is a non-decreasing function of $n$. For our analysis, we assume $f(n) = \Theta(n^\mu)$, where $0 \leq \mu < 1$.

We assume that each wireless node is assigned to the closest base station in Euclidean distance. Thus, the network area is divided into $f(n)$ cellular regions. If the size of each cellular region is large compared to the transmission range $r(n)$ (equivalently $a(n)$) of the wireless nodes, then a wireless node transmits to its assigned base station via multi-hop relaying through other wireless nodes.

We now consider a transmission and caching scheme for the heterogeneous wireless network, which is similar to the scheme considered for the ad hoc case. That is, the network area is divided into $a^{-1}(n)$ squared cells each with area $a(n)$. Based on a TDM scheme, each node, including base stations, transmit packets over the shared channel, subject to the Protocol Model. For simplicity, we assume all the nodes, including base stations, have the same transmission range, $r(n)$. Note that this is a reasonable assumption when considering femtocells.

Each wireless node can request contents from its assigned base station through multi-hop relaying. Each wireless node requests content $m$ with probability $p_m$. If the closest wireless holder of content $m$ is closer to the requesting node than the node’s assigned base station, then the content is retrieved from the closest wireless holder. Otherwise, it is retrieved from the base station.

Similar to the previous sections, we assume that the $X_m$ wireless holders of content $m$ are uniformly distributed in the network area. Since we are interested in evaluating the performance of the wireless network, we assume that all requests for content, upon reception at base stations, are satisfied immediately (i.e. a Data Packet is generated immediately). In other words, we do not consider the delay within the wired backbone network.

Unlike the pure ad hoc case in which we need to have at least one copy of each content object in the caches of the wireless nodes to satisfy all the requests, for the proposed heterogeneous network we relax this restriction due to the presence of the base stations. As a result, the number of content types can exceed the number of nodes. i.e., $\beta$ can be $\geq 1$.

As in Lemma [3] we can show that the average length of the $L_{H,R}(i, m)$ line connecting the requesting node $i$ to the closest cache of content $m$ (either a wireless holder or a base station) is given by:

$$E[|L_{H,R}(i, m)|] = \Theta\left(\frac{1}{\sqrt{\lambda + f(n)}}\right). \quad (51)$$

Consequently, the average of number of hops along the $L_{H,R}$ line is w.h.p.

$$E[H_{i,m}] = \Theta\left(\max\left\{1, \frac{1}{\sqrt{a(n)(\lambda + f(n))}}\right\}\right). \quad (52)$$

Using an approach similar to that in the proof of Lemma [5] we see that for $a(n) \geq 2\log n/n$, the number of $L_{H,R}$ lines passing through each cell (of area $a(n)$) is

$$\Theta\left(n \max_{m=1}^M p_m \max\left\{a(n), \sqrt{\frac{a(n)}{\lambda + f(n)}}\right\}\right) \text{ w.h.p.}$$

Therefore, the throughput and the delay of the achievable scheme for the heterogeneous network model are given by:

$$\lambda(n) = \Theta\left(\frac{1}{n \sum_{m=1}^M p_m \max\{a(n), \sqrt{\frac{a(n)}{\lambda + f(n)}}\}}\right) \text{ w.h.p.} \quad (53)$$

$$D = \Theta\left(\frac{1}{\sum_{m=1}^M p_m \max\left\{1, \frac{1}{\sqrt{a(n)(\lambda + f(n))}}\right\}}\right) \text{ w.h.p.} \quad (54)$$

Combining the equations (53) and (54), we obtain the same throughput-delay trade-off as in the ad hoc case given in (21).

Next, we optimize the throughput and delay of the achievable scheme for the heterogeneous network scenario by choosing the appropriate $(X_m)_{m=1}^M$. Note that here the constraints on $X_m$ are $0 \leq X_m \leq a^{-1}(n) - f(n)$, as larger $X_m$‘s do not change the order of the throughput or delay. Thus, the optimization problem is

$$\min_{(X_m)} \sum_{m=1}^M \frac{p_m}{\sqrt{a(n)(\lambda + f(n))}}$$

subject to:

$$(X_m)_{m=1}^M$$

$$\sum_{m=1}^M X_m \leq n\lambda$$

$$0 \leq X_m \leq a^{-1}(n) - f(n) \quad \text{for} \quad m = 1, 2, \ldots, M. \quad (55)$$

Since the objective function is strictly convex, we are assured that there is a unique global minimum. Defining the non-negative Lagrange multipliers $\lambda$ for the constraint

![Fig. 2: The logarithmic scaling of network throughput for various values of $\alpha$ vs. number of nodes, for $\beta = 0.9$.](image)
\[ \sum_{m=1}^{M} X_m \leq nK, \] and taking into account the constraint 0 \leq X_m \leq a^{-1}(n) - f(n), the necessary conditions for a minimum of D with respect to X_m, \( \forall m \in M \) are given by

\[
\frac{\partial D}{\partial X_m} \begin{cases} 
-\lambda & \text{if } X_m = a^{-1}(n) - f(n) \\
-\lambda & \text{if } 0 < X_m < a^{-1}(n) - f(n) \\
\geq -\lambda & \text{if } X_m = 0 
\end{cases} \tag{56}
\]

Given the Zipf distribution, let \( M_1 = \{1, 2, \ldots, m_1 - 1\} \) be the set of content objects such that \( X_m = a^{-1}(n) - f(n) \) for \( m \in M_1 \). Similarly, let \( M_2 = \{m_1, m_1 + 1, \ldots, m_2 - 1\} \) and \( M_3 = \{m_2, m_2 + 1, \ldots, M\} \) be the set of contents such that \( 0 < X_m < a^{-1}(n) - f(n) \) for \( m \in M_2 \), and \( X_m = 0 \) for \( m \in M_3 \), respectively. From (56), we have \( \forall m \in M \)

\[
\frac{p_m}{2\sqrt{a(n)(X_m + f(n))^3}} \begin{cases} 
\geq \lambda & \forall m \in M_1 \\
= \lambda & \forall m \in M_2 \\
\leq \lambda & \forall m \in M_3 \tag{57}
\end{cases}
\]

Using the equality for the case \( \forall m \in M_2 \), we obtain

\[
m_1/m_2 \simeq (a(n)f(n))^m. \tag{58}
\]

From (57), we have \( \lambda > 0 \) and hence, \( \sum_{m=1}^{M} X_m = nK \). Combining this with (58), we can derive \( m_1 \) and \( m_2 \). The optimal number of holders of content \( m, X^*_m \), is then given by

\[
X^*_m = \begin{cases} 
a^{-1}(n) - f(n), & m = 1, 2, \ldots, m_1 - 1 \\
p_2^{2/3}/\sum_{j=m_1}^{m_2-1} p_j, & m = m_1, \ldots, m_2 - 1 \\
0, & m = m_2, \ldots, M \tag{59}
\end{cases}
\]

where \( K' \equiv K - (m_1 - 1)\frac{a^{-1}(n)}{n} + (m_2 - 1)\frac{f(n)}{n} \). Hence, the average delay is w.h.p.

\[
D^*(n) = \Theta \left( \left( \frac{\sum_{j=m_1}^{m_2-1} 2^{j/3}}{\sqrt{a(n)nK'}} + \frac{\sum_{j=m_2}^{M} p_j}{\sqrt{f(n)a(n)}} \right)^{2/3} \right) \tag{60}
\]

We can now apply techniques similar to the one used in the ad hoc case in order to estimate the indices \( m_1 \) and \( m_2 \), and then compute the scalings of the delay and throughput. So far we have considered \( a(n) \geq 2\log n/n \) to be a general parameter resulting in a trade-off between the throughput and delay of the network: as \( a(n) \) increases (decreases), both throughput and delay of the network decrease (increase). In this section, we consider a single point of this trade-off where \( a(n) = 2\log n/n \), as this will give us more intuitive formulas for delay and throughput. The generalization of this result is a straightforward calculation following the approach of the ad hoc case. Following this, we can estimate the indices \( m_1 \) and \( m_2 \) as follows.

**Lemma 7.** Taking \( n \to \infty \), \( m_1 \) and \( m_2 \) scales as:

\[
m_1 = \begin{cases} 
\Theta(\log n) & \alpha > 3/2 \\
\Theta(1) & \alpha = 3/2 \\
\text{converging to 1} & \alpha < 3/2 \tag{61}
\end{cases}
\]

\[
m_2 = \begin{cases} 
\min \{ M + 1, \Theta \left( \left( \frac{n}{f(n)} \right)^{\frac{\alpha}{2}} \left( \log n \right)^{1-\frac{\alpha}{2}} \right) \} & \alpha > 3/2 \\
\min \{ M + 1, \Theta \left( \frac{n}{f(n)} \right) \} & \alpha = 3/2 \\
\min \{ M + 1, \Theta \left( \frac{n}{f(n)} \right) \} & \alpha < 3/2 \tag{62}
\end{cases}
\]

**Proof:** Since \( \mu < 1 \), \( K' \to K - (m_1 - 1)/(2\log n) \) as \( n \to \infty \). By definition, \( m_1 \) is the smallest index for which the number of holders is less than \( a^{-1}(n) - f(n) \). Using (59), it follows that

\[
2K' \log n < m_1^2 \left( H_{2\alpha'}(m_2 - 1) - H_{2\alpha'}(m_1 - 1) \right). \tag{63}
\]

Now, if \( m_1 > 1 \), attempting to decrease the index \( m_1 \) by one would result in

\[
\frac{p_{m_1-1}^{2/3}}{\sum_{j=m_1-1}^{m_2-1} p_j^{2/3}} nK' - f(n) \geq a^{-1}(n) - f(n).
\]

Hence, we have

\[
2K' \log n \geq (m_1 - 1)^{2\alpha'} [H_{2\alpha'}(m_2 - 1) - H_{2\alpha'}(m_1 - 2)]. \tag{64}
\]

For \( m_1 > 1 \), an approximation of \( m_1 \) can be obtained from:

\[
2K' \log n \simeq m_1^{2\alpha'} [H_{2\alpha'}(m_2 - 1) - H_{2\alpha'}(m_1 - 1)]. \tag{65}
\]

Similarly, by the definition of \( m_2 \), we know \( X_{m_2-1} > 0 \). Using (59), it follows that

\[
\frac{nK'}{f(n)} > (m_2 - 1)^{2\alpha'} [H_{2\alpha'}(m_2 - 1) - H_{2\alpha'}(m_1 - 1)]. \tag{66}
\]

If \( m_2 \leq M \), attempting to increase the index \( m_2 \) by one would lead to

\[
\frac{nK'}{f(n)} \leq m_2^{2\alpha'} [H_{2\alpha'}(m_2 - 1) - H_{2\alpha'}(m_1 - 1)]. \tag{67}
\]

It follows that

\[
\frac{nK'}{f(n)} \simeq (m_2 - 1)^{2\alpha'} [H_{2\alpha'}(m_2 - 1) - H_{2\alpha'}(m_1 - 1)]. \tag{68}
\]

For \( \alpha > 3/2 \): By using (65), we have

\[
2\log n(K - (m_1 - 1)/2\log n) \simeq (m_1 - 1)^{2\alpha} \left[ (m_2 - 1)^{2\alpha'} - 1 - \frac{2\alpha}{\alpha'} \right]. \tag{69}
\]

which leads to

\[
m_1 \simeq 1 + \frac{2\alpha - 3}{\alpha} K \log n. \tag{70}
\]

Now if \( m_2 \leq M \), following (58) we have

\[
m_2 = \Theta \left( \left( \frac{n}{f(n)} \right)^{\frac{\alpha}{2}} \left( \log n \right)^{1-\frac{\alpha}{2}} \right). \tag{71}
\]

For \( \alpha = 3/2 \): using (68), we have

\[
m_1 - 1 \simeq 2\log n \left( K - \frac{(m_1 - 1)}{2\log n} \right). \tag{72}
\]
leading to $m_1 - 1 \sim \frac{K \log n}{\log m_2}$. Now, if $m_2 = M + 1$, then $m_1 = \Theta(1)$. Otherwise, if $m_2 \leq M$, combining this result with (58), we have

$$m_1 = \Theta(1).$$

$$m_2 = \Theta\left(\frac{n}{f(n) \log n}\right).$$

For $\alpha < 3/2$: using (64) we have

$$(m_1 - 1)^{2a} \leq \frac{2 \log n (K - \frac{(m_1 - 1)}{2 \log n})}{m_2^2 - 2a/3}. \quad (73)$$

Using straightforward calculations, it follows that

$$(m_1 - 1)^{2a} \leq \frac{2K \log n}{m_2^2 - 2a/3}. \quad (74)$$

If $m_2 = M + 1$ then clearly, the RHS converges to zero. Therefore, $m_1 \to 1$ as $n$ grows. Otherwise, if $m_2 \leq M$, by using (68) we have

$$m_2 - 1 \sim \frac{n(K - \frac{(m_1 - 1)}{2 \log n})}{f(n)} = \Theta\left(\frac{n}{f(n)}\right). \quad (75)$$

By plugging in this result in (74), the RHS converges to zero, as previously. Thus, $m_1 \to 1$ as $n$ grows.

We now compute the throughput and delay of the proposed heterogeneous network model as follows. Note that part 1 of Theorem 2 considers the case where $m_2 = M + 1$. For $\alpha \leq 3/2$ this happens when $\beta < 1 - \mu$, or equivalently $f(n) = o\left(\frac{\log n}{n}\right)$ and $f(n) \geq 1$. For $\alpha > 3/2$, $m_2 = M + 1$ if $\beta \leq \frac{3}{2a} (1 - \mu)$, or equivalently $f(n) = O\left(\frac{\log n}{n^{\frac{1}{3}}}\right)$ and $f(n) \geq 1$. On the other hand, part 2 of Theorem 3 shows the performance of the network when $m_2 \leq M$. For $\alpha \leq 3/2$ this happens when $\beta \geq 1 - \mu$, or equivalently $f(n) = \Omega\left(\frac{n^{\frac{1}{2}}}{\log n}\right)$ and $f(n) \geq 1$. In addition, for $\alpha > 3/2$, $m_2 \leq M$ if $\beta > \frac{3}{2a} (1 - \mu)$, or equivalently $f(n) = o\left(\frac{\log n}{n^{\frac{1}{3}}}\right)$ and $f(n) \geq 1$. Note that for any value of $\alpha$, if $f(n) = \Omega\left(\frac{n^{\frac{1}{2}}}{\log n}\right)$ and $f(n) \geq 1$ (or equivalently $\mu \geq \max\{0, 1 - \beta\}$), then the heterogeneous network performance follows (76) and (77).

**Theorem 3.** For $a(n) = 2 \log n/n$,

1. The throughput and delay performance of the achievable scheme for the heterogeneous network, when $m_2 = M + 1$ and the content popularity distribution follows the Zipf distribution, is the same as given in (48) and (47), respectively.

2. The throughput and delay of the achievable scheme, when $m_2 \leq M$, are w.h.p.:

$$D^*(n) = \begin{cases} 
\Theta(1) & \alpha > 3/2 \\
\Theta(\log n) & \alpha = 3/2 \\
\Theta\left(\frac{n^{\frac{1}{3}}}{\log n^{\frac{1}{3}}}\right) & 1 < \alpha < 3/2 \\
\Theta\left(\frac{n}{f(n) \log n}\right) & \alpha \leq 1 
\end{cases} \quad (76)$$

**Proof:** The proof is given in Appendix C.

Comparing the results for the heterogeneous network in Theorem 3 with those for the pure ad hoc network given in Theorem 2 for $a(n) = \Theta\left(\frac{\log n}{n}\right)$, we conclude that the number of base stations in the network needs to be greater than $n^{\frac{3}{\beta}} = n^{1 - \beta}$ to improve the order of the performance metrics (throughput and delay). For the scenario where $\beta \geq 1$, this condition reduces to $f(n) \geq 1$. In other words, if $\beta \geq 1$, the heterogeneous network always outperforms the pure ad hoc network. Also, note that for $\alpha \geq 3/2$, the performance of the heterogeneous network is the same as that for the pure ad hoc case. Intuitively, this is because for large $\alpha$’s, the majority of content requests are for the most popular content objects, hence, caching the most popular content objects will almost eliminate the need for base stations.

We have plotted the theoretical results given in (76) and (77) in Figures 3 and 4, respectively, to demonstrate the scaling of the network delay and per-node throughput for $\alpha = 0.8$ and $\alpha = 1.2$. In addition, we have plotted the performance of the ad hoc network model for the same values of $\alpha$. In both figures, $\beta = 0.9$ and $f(n) = n^{0.4}$. Note that for $\alpha > 3/2$, the performance of the heterogeneous network is the same in order as that for the ad hoc case. In Figures 5 and 6, the scaling of the per-node throughput and network delay is shown for $\alpha = 0.8$, $\beta = 0.9$, and various values of $\mu$, along with the corresponding scaling for the pure ad hoc case. As predicted, by adding more base stations to the network, the performance of the network, both in terms of throughput and delay, is improved.

![Fig. 3: The scaling of delay for the heterogeneous and ad hoc network models for various values of $\alpha$, vs. number of nodes, for $\beta = 0.9$, and for $\mu = 0.4$.](image-url)
Fig. 4: The logarithmic scaling of per-node throughput for the heterogeneous and ad hoc network models for various values of α, vs. number of nodes, for β = 0.9, and, for μ = 0.4.

Fig. 5: The logarithmic scaling of per-node throughput for the heterogeneous and ad hoc network models for various values of μ, vs. number of nodes, for β = 0.9, and, for α = 0.8.

Fig. 6: The scaling of delay for the heterogeneous and ad hoc network models for various values of μ, vs. number of nodes, for β = 0.9, and, for α = 0.8.

VI. Conclusions

We have investigated the asymptotic behavior of wireless networks with an content-centric networking architecture. We derived the throughput-delay tradeoff of the proposed wireless ad hoc network model for a general content popularity distribution, and optimized the caching strategy to simultaneously minimize the average network delay and maximize the network throughput. Using the optimal caching strategy, we evaluated the network performance under a Zipf content popularity distribution. Furthermore, we investigated heterogeneous wireless networks where, in addition to wireless nodes, there are a number of base stations uniformly distributed at random in the network area. We showed that in order to achieve a better performance in a heterogeneous network in the order sense, the number of base stations needs to be greater than the ratio of the number of nodes to the number of content types. For the case where the number of content objects is greater than the number of wireless nodes, this condition reduces to having at least one base station in the network. In addition, we demonstrated that for the Zipf content popularity distribution with exponent α ≥ 3/2, the performance of the wireless ad hoc network is of the same order as for the heterogeneous wireless network, independent of number of base stations.

Appendix

A. Proof of Lemma 3

In order to prove (5), we define g(n) as

\[ g(n) = \frac{n-1}{n} \cdot \frac{n-3}{n-2} \cdot \ldots \cdot \frac{2}{3}. \]  

We want to show that \( g(2X_m + 1) = \Theta(1/\sqrt{m}) \), or equivalently, \( g(n) \) scales as \( 1/\sqrt{n} \) where \( n \) is an odd number. Let \( n_1 \) and \( n_2 \) be two odd numbers and \( n_1 > n_2 \). Then, we can write

\[ g(n_1) = \frac{n_1-1}{n_1} \cdot \frac{n_1-3}{n_1-2} \cdot \ldots \cdot \frac{n_2+1}{n_2+2} \cdot g(n_2). \]

Reordering the sequence in RHS and dividing both sides by \( g(n_2) \),

\[ g(n_1) = \frac{n_2+1}{n_1} \cdot \frac{n_1-1}{n_1} \cdot \frac{n_1-3}{n_1-2} \cdot \ldots \cdot \frac{n_2+3}{n_2+2}. \]

Hence, we have

\[ \frac{g(n_1)}{g(n_2)} = \frac{g(n_1-1)}{g(n_2+1)} = \frac{n_2+1}{n_1}. \]  

Also, it can easily be shown that \( g(n) \) is a non-increasing function of \( n \). Thus,

\[ \frac{g(n_1-1)}{g(n_2+1)} \geq \frac{g(n_1)}{g(n_2)}. \]
Therefore,
\[
\left(\frac{g(n_1)}{g(n_2)}\right)^2 \leq \frac{g(n_1)}{g(n_2)} \cdot \frac{g(n_1 - 1)}{g(n_2 + 1)} = \frac{n_2 + 1}{n_1}.
\] (81)

Also,
\[
\left(\frac{g(n_1 - 1)}{g(n_2 + 1)}\right)^2 \geq \frac{g(n_1)}{g(n_2)} \cdot \frac{g(n_1 - 1)}{g(n_2 + 1)} = \frac{n_2 + 1}{n_1}.
\] (82)

Combining (81) and (82), we have the required bound for the proof:
\[
\frac{n_2}{n_1 + 1} \leq \left(\frac{g(n_1)}{g(n_2)}\right)^2 \leq \frac{n_2 + 1}{n_1}.
\] (83)

**B. Proof of Theorem 2**

We prove that the average delay is given by (47). The average throughput given in (48) can be calculated easily by using equation (21). Substituting for the \(p_j\)'s in equation (30) using the Zipf distribution, we obtain
\[
D = \frac{H_o(m_1 - 1)}{H_o(M)} + \frac{H_o(m_2 - 1)}{\sqrt{nK\alpha(a(n)n)H_o(M)}} - \frac{H_o(M) - H_o(m_2 - 1)}{\alpha(a(n))H_o(M)}.
\] (84)

where \(K' = K - \frac{(m_1 - 1)}{2\log n} - \frac{(M - m_2 + 1)}{2\log n} = \Theta(1).\) Let the three expressions on the RHS of (84) be denoted by \(D_1, D_2,\) and \(D_3,\) respectively.

Clearly, \(D_1 = \Theta(1), \forall \alpha > 0.\) Also, if \(a(n) = 1\) then \(m_2 = m_1 + 1,\) and \(D_2 = 0.\) It can easily be shown that \(D_1 = 1,\) and \(\lambda = W/n,\) which coincides with the result of time-division with direct communication policy. Hence, we assume here that \(a(n) < 1.\) By Lemma [6] we know that for \(\alpha \leq 3/2, m_2 = M + 1.\) Therefore, \(D_3\) is zero, and \(D = \Theta(\max\{D_1, D_2\})\).

For \(\alpha < 1:\)
\[
D_2 = \Theta\left(\frac{(m_2 - 1)^{3/2 - \alpha}}{\alpha(a(n))\sqrt{n}M^{1 - \alpha}}\right).
\] (85)

For \(\alpha = 1:\)
\[
D_2 = \Theta\left(\frac{(m_2 - 1)^{1/2}}{\sqrt{n}a(n)\sqrt{M}}\right).
\] (86)

For \(1 < \alpha < 3/2:\) similarly, we have
\[
D_2 = \Theta\left(\frac{M^{3/2 - \alpha}}{\sqrt{n}a(n)}\right).
\] (87)

For \(\alpha = 3/2:\)
\[
D_2 = \Theta\left(\frac{\log M}{\sqrt{n}a(n)}\right).
\] (88)

For \(\alpha > 3/2:\) \(D_2 = \Theta\left(\frac{1}{\sqrt{n}a(n)}\right) = o(1).\) Also, as shown in the following, \(D_3 = o(1)\). Therefore, \(D = \Theta(D_1) = \Theta(1)\).

Now, if \(m_2 = M + 1\) then \(D_3 = 0.\) Otherwise, \(m_2 \approx \frac{2\alpha - 1}{2\alpha}K\alpha(n(a(n)))^{1 - \frac{1}{2\alpha}}.\) Using straightforward calculation, it follows that
\[
D_3 = \Theta\left(\frac{m_2^{1 - \alpha}}{\sqrt{n}a(n)}\right) = o(1).
\] (89)

**C. Proof of Theorem 3**

We compute the average delay. The average throughput follows by (21). Substituting for the \(p_j\)'s in equation (60) using the Zipf distribution, we have
\[
D = \frac{H_o(m_1)}{H_o(M)} + \frac{H_o(m_2) - H_o(m_1)}{\sqrt{K\log nH_o(M)}} + \frac{1}{\sqrt{f(n)\log n}}\left|H_o(M) - H_o(m_2)\right|.
\] (90)

where \(K' \rightarrow K - \frac{(m_1 - 1)}{2\log n} \approx 3\) as \(n \rightarrow \infty.\) Similar to the proof of Theorem 2, let the three expressions on the RHS of (90) be denoted by \(D_1, D_2,\) and \(D_3,\) respectively. Moreover, when \(m_2 = M + 1, D_3 = 0.\) Hence, the equation (90) is simplified to equation (84), given that \(m_2 = M + 1.\) As shown in (31), this always holds for \(\alpha \leq 3/2.\) In addition, for \(\alpha > 3/2,\) if we assign \(m_2 = M + 1,\) we still get the same result as shown in Appendix B.

Now we prove the results for the second part of the theorem, where \(m_2 \leq M.\) By Lemma [7] we know for \(\alpha \leq 3/2, K' \rightarrow K\) and \(D_1 = o(1).\) For \(\alpha > 3/2, D_1 = \Theta(1).\)

For \(\alpha < 1: \) by (62), \(m_2 = \Theta\left(\frac{1}{\alpha(n)}\right).\) Following (90),
\[
D_2 \approx \frac{n^{1-\mu}(3/2 - \alpha)}{\sqrt{\log nM^{1-\alpha}}}. \quad \text{(91)}
\]
\[
D_3 \approx \frac{n^{1-\mu}}{\sqrt{\log n}}. \quad \text{(92)}
\]

It can easily be shown that \(D_2 = o(D_3).\) Thus, \(D = \Theta(D_3).\)

For \(\alpha = 1: \) Similarly, by using (90), it follows that \(D_3 \) is given by (92). For \(D_2\) we have
\[
D_2 \approx \frac{n^{1-\mu}(3/2 - \alpha)}{\sqrt{\log nM^{1-\alpha}}}. \quad \text{(93)}
\]

Now since \(\log M = \Theta(\log n),\) we have \(D_2 = \Theta\left(\frac{n^{1-\mu}}{(\log n)^{3/2}}\right).\) Clearly, \(D_2 = o(D_3).\) Hence, \(D = \Theta(D_3).\)

For \(1 < \alpha < 3/2: \) By using the same technique as in the previous part, we can see that \(D_3 = \Theta(D_2)\) and therefore, \(D = \Theta(D_2);\) we have
\[
D_2 \approx \frac{m_2^{3/2 - \alpha}}{\sqrt{\log n}} = \Theta\left(\frac{n^{1-\mu}(3/2 - \alpha)}{\sqrt{\log n}}\right). \quad \text{(94)}
\]
\[
D_3 \approx \frac{m_2^{1-\alpha}n^{1-\mu}}{\sqrt{\log n}} = \Theta\left(\frac{n^{1-\mu}(3/2 - \alpha)}{\sqrt{\log n}}\right). \quad \text{(95)}
\]

For \(\alpha = 3/2: \) using Lemma [7] it follows from (90) that
\[
D_2 \approx \left(\frac{n^{1-\mu}}{\log n}\right)^{3/2} = \Theta(\log n). \quad \text{(96)}
\]
\[
D_3 \approx \frac{m_2^{1-\alpha}n^{1-\mu}}{\sqrt{\log n}} = \Theta(1). \quad \text{(97)}
\]

Therefore, \(D = \Theta(D_2).\)

For \(\alpha > 3/2: \) using a similar calculation, we have
\[
D_2 \approx \frac{m_2^{3/2 - \alpha}}{\sqrt{\log n}} = o(1).
\]
and,

$$D_3 \simeq \frac{m_{12}^{1-\alpha_n} \sqrt{\log n}}{n^\alpha_n} = o(1).$$  \hspace{1cm} (98)

To show the last equation in (98), let’s consider the power of $n$ in $D_3$: 

$$\frac{3}{2\alpha_n} (1 - \mu)(1 - \alpha_n) + \frac{\mu}{n} = (1 - \mu) \left(\frac{3}{2\alpha_n} - 1\right) < 0. $$

Hence, $D_3 \to 0$ as $n \to \infty$. Therefore, $D = \Theta(D_1) = \Theta(1)$.

REFERENCES

[1] P. Gupta, P.R. Kumar, The capacity of wireless networks, IEEE Transactions on Information Theory 46, 2000, no. 2, pp. 388–404.

[2] Xiaojun Lin and Ness B. Shroff, The fundamental capacity-delay tradeoff in large mobile ad hoc networks, Proc. Third Annual Mediterranean Ad Hoc Networking Workshop, 2004.

[3] M.J. Neely and E. Modiano, Capacity and delay tradeoffs for ad hoc mobile networks, IEEE Transactions on Information Theory 51, 2005, no. 6, pp. 1917–1937.

[4] Matthias Grossglauser and D.N.C. Tse, Mobility increases the capacity of ad hoc wireless networks, IEEE/ACM Transactions on Networking 10, 2002, no. 4, pp. 477–486.

[5] A. El Gamal, I. Mammen, B. Prabhakar, D. Shah, Optimal throughput-delay scaling in wireless networks - part i: the fluid model, IEEE Transactions on Information Theory 52, 2006, no. 6, pp. 2568–2592.

[6] A. El Gamal, I. Mammen, B. Prabhakar, D. Shah, Throughput-delay scaling in wireless networks with constant-size packets, Proc. IEEE International Symposium on Information Theory, 2003, pp. 1329–1333.

[7] Z. Wang, H.R. Sadjadpour, J.J. Garcia-Luna-Aceves, S. Karande, Fundamental limits of information dissemination in wireless ad hoc networks - part i: single-packet reception, IEEE Transactions on Wireless Communications 8, 2009, no. 12, pp. 5749–5754.

[8] L. Zhang, D. Estrin, J. Burke, V. Jacobson, J. Thornton, D. Smetters, G. Tsudik B. Zhang, kc cliffy, D. Krioukov, D. Massey, C. Papadopoulos, T. Abdelzaher, L. Wang and P. Crowley, E. Yeh, Named data networking (ndn) project, 2010.

[9] V. Jacobson, D. Smetters, J. Thornton, M. Plass, N. Briggs, R. Braynard, Networking named content, Proc. International Conference on Emerging Networking Experiments and Technologies (ACM CoNEXT), 2009, pp. 1–12.

[10] S. Gitzenis, G.S. Paschos, L. Tassiulas, Asymptotic laws for joint content replication and delivery in wireless networks, IEEE Transactions on Information Theory 59, 2013, no. 5, pp. 2760–2776.

[11] G. Alfano, M. Garetto, E. Leonardi, Content-centric wireless networks with limited buffers: When mobility hurts, Proc. IEEE INFOCOM, 2013, pp. 1815–1823.

[12] B. Azimdoost, C. Westphal, H.R. Sadjadpour, On the throughput capacity of content-centric networks, Proc. International Teletraffic Congress (ITC), 2013, pp. 1–9.

[13] K. Shannugam, N. Golrezaei, A. Dimakis, A. Molisch, G. Caire, Femtocaching: Wireless content delivery through distributed caching helpers, IEEE Transactions on Information Theory 59, 2013, no. 12, pp. 8402–8413.

[14] M. Ji, G. Caire, A. Molisch, Wireless device-to-device caching networks: Basic principles and system performance, 2013, CoRR abs/1305.5216.

[15] E. Yeh, T. Ho, Y. Cui, M. Burd, R. Liu, D. Leong, VIP: A framework for joint dynamic forwarding and caching in named data networks, Proc. International Conference on Information-Centric Networking (ACM ICN), 2014, pp. 117–126.

[16] A. Ghosh, N. Mangalvedhe, R. Ratasuk, B. Mondal, M. Cudak, E. Vitsotsky, T.A. Thomas, J.G. Andrews, P. Xia, H.S. Jo, H.S. Dhillon, T.D. Novlan, Heterogeneous cellular networks: From theory to practice, IEEE Communications Magazine 50, 2012, no. 6, pp. 54–64.

[17] Robert Tarjan, Probability and computing, https://www.cs.princeton.edu/courses/archive/fall09/cos521/Handouts/probabilityandcomparing.pdf.