

# Wireless Multihop Device-to-Device Caching Networks

Sang-Woon Jeon, *Member, IEEE*, Song-Nam Hong, *Member, IEEE*, Mingyue Ji, *Member, IEEE*, Giuseppe Caire, *Fellow, IEEE*, and Andreas F. Molisch, *Fellow, IEEE*

## Abstract

We consider a wireless device-to-device (D2D) network where  $n$  nodes are uniformly distributed at random over the network area. We let each node with storage capacity  $M$  cache files from a library of size  $m \geq M$ . Each node in the network requests a file from the library independently at random, according to a popularity distribution, and is served by other nodes having the requested file in their local cache via (possibly) multihop transmissions. Under the classical “protocol model” of wireless networks, we characterize the optimal per-node capacity scaling law for a broad class of heavy-tailed popularity distributions including Zipf distributions with exponent less than one. In the parameter regimes of interest, we show that a decentralized random caching strategy with uniform probability over the library yields the optimal per-node capacity scaling of  $\Theta(\sqrt{M/m})$ , which is constant with  $n$ , thus yielding throughput scalability with the network size. Furthermore, the multihop capacity scaling can be significantly better than for the case of single-hop caching networks, for which the per-node capacity is  $\Theta(M/m)$ . The multihop capacity scaling law can be further improved for a Zipf distribution with exponent larger than some threshold  $> 1$ , by using a decentralized random caching uniformly across a subset of most popular files in the library. Namely, ignoring a subset of less popular files (i.e., effectively reducing the size of the library) can significantly improve the throughput scaling while guaranteeing that all nodes will be served with high probability as  $n$  increases.

## Index Terms

Caching, device-to-device networks, multihop transmission, scaling laws.

## I. INTRODUCTION

Internet traffic has grown dramatically in recent years, mainly due to on-demand video streaming [1]. While wireless is by far the preferred way through which users connect to the Internet, today’s cellular technology and service providers do not support seamless cost-effective on-demand video streaming. For example, most monthly cellular data plans would be completely consumed by a *single* streaming session of a standard definition movie from a typical services such as Netflix, iTunes, or Amazon Prime (duration 1h:30, size 2GB). It is evident that in order to fill in the gap between the users’ expectation and the limitations of the provided services, a dramatic technology paradigm shift is required. In this perspective, it has been recently recognized that *caching at the wireless edge*, i.e., caching the content library directly in the wireless nodes (femtocell base stations or user devices), has the potential of solving the problem of network scalability by providing per-node throughput that scales much better than conventional unicast transmission, in a variety of scenarios.

One important feature of on-demand video streaming is that user demands are highly redundant over time and space. As an example, consider a university campus where  $n \approx 10000$  users (distributed over a surface of

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (MEST) [NRF-2013R1A1A1064955] and by the NSF Grants CCF 1161801 and EARS 1444060.

The material in this paper was presented in part at the IEEE Information Theory Workshop (ITW), Jerusalem, Israel, April 2015 and at the IEEE International Conference on Communications (ICC), London, United Kingdom, June 2015.

S.-W. Jeon is with the Department of Information and Communication Engineering, Andong National University, Andong, South Korea (e-mail: swjeon@anu.ac.kr).

S.-N. Hong is with the Ericsson Research Lab., San Jose, CA 95118, USA (e-mail: songnam.hong@ericsson.com).

M. Ji and A. F. Molisch are with the Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089, USA (e-mail: mingyuej@usc.edu; molisch@usc.edu).

G. Caire is with the Department of Telecommunication Systems, Technical University of Berlin, Berlin 10623, Germany, and also with the Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089, USA (e-mail: caire@tu-berlin.de).

$\approx 1km^2$ ) stream movies from a library of  $\approx 100$  files, such as the weekly top-of-the chart titles of Netflix, iTunes, or Amazon Prime. For such scenario, each user demand can be satisfied by local communication from a cache, without cluttering a cellular base station with thousands of unicast sessions, or without requiring to deploy a large number of small cell access points, each requiring costly high-throughput backhaul. Intuitively, caching can effectively take advantage of the inherent redundancy of the user demands, although, differently from live streaming, in on-demand streaming users do not request the same content at the same time (this type of redundancy is referred to in [2]–[4] as *asynchronous content reuse*).

#### A. Related Work

1) *Conventional ad-hoc networks*: Since the seminal work of Gupta and Kumar [5], the capacity scaling laws of wireless ad-hoc networks has been extensively studied (e.g., [6]–[8]). The model introduced in [5] consists of  $n$  nodes placed uniformly at random on a planar region and grouped into source–destination (SD) pairs at random. Assuming an interference avoidance constraint referred to as the *protocol model* (see Section II), it was shown in [5] that the per-node capacity must scale as  $O(\frac{1}{\sqrt{n}})$  (i.e., upper bound). Furthermore, a simple “straight-line” multihop relaying scheme achieves the per-node throughput scaling of  $\Omega(\frac{1}{\sqrt{n \log n}})$ . The same results were confirmed in [6] by using a simpler and more general analysis technique. Later, the  $1/\sqrt{\log n}$  gap factor between converse and achievability was closed in [8], by showing that the per-node throughput scaling of  $\Theta(\frac{1}{\sqrt{n}})$  is indeed achievable by using a more refined multihop strategy based on percolation theory.

Beyond the protocol model, the capacity scaling law of wireless ad-hoc networks has been also studied in an information theoretic sense, considering a *physical model* that includes distance-dependent propagation path-loss, fading, Gaussian noise, and signal interference (e.g., [9]–[13]). While the protocol model is scale-free, the physical model behaves differently depending on whether the network is “extended” (constant node density, with the network area growing as  $\Theta(n)$ ), or “dense” (constant network area, with the node density growing as  $\Theta(n)$ ). In [9], [10], the achievability schemes are based on multihop strategy, point-to-point coding, and treating interference as (Gaussian) noise. For the extended network model, it was shown that if the path-loss exponent is greater than or equal to three, then the scaling law is the same as for the protocol model and the multihop strategy is sufficient to achieve the optimal scaling. In contrast, for the extended network model with the path-loss exponent less than three and for the dense network model (in this case the path-loss exponent is irrelevant) the multihop strategy is suboptimal. In these cases, the *hierarchical cooperation* scheme proposed in [13] (see also improved and optimized hierarchical cooperation scheme in [14], [15]) achieves an almost optimal throughput scaling within a factor of  $n^\epsilon$ , where  $\epsilon$  can be made arbitrarily small as the number of hierarchical stages increases.

2) *Caching networks*: Motivated by the considerations made at the beginning of this section, wireless caching networks have been the subject of recent intensive investigation [2]–[4], [16]–[25]. The single-hop device-to-device (D2D) case was considered in [3], [4], [17], where  $n$  user node request files from a library of  $m$  files according to a common demand or popularity distribution and each node has cache capacity constraint equal to the size of  $M \leq m$  files. The delivery scheme (i.e., the coordination of transmissions in order to serve the users’ requests) is restricted to be one-hop, i.e., either the requested file is found in the cache, or it is directly downloaded from a neighbor node through a D2D wireless link. Under a Zipf popularity distribution [26] with parameter less than one and the protocol model of [5], it was shown in [3], [4], [17] that the per-node throughput scales as  $\Theta(M/m)$ . This can be achieved by an independent and random caching placement and a TDMA-based link scheduling scheme, at the expense of a positive outage probability, due to the random nature of the caching placement scheme. However, in the relevant regime where  $nM \gg m$ , this outage probability can be kept under control, i.e., the system can be designed in order to achieve any target outage probability  $\epsilon > 0$ , for sufficiently large  $n$ . It is remarkable to notice that the per-node throughput in this case scales much better than in the case of general ad-hoc networks under the protocol model. In fact, while in the general case the per-node throughput converges to zero with the size of the network as  $1/\sqrt{n}$ , here it is constant with  $n$  and directly proportional to the fraction of cached files  $M/m$ . This much better scaling can be explained as an effect of the dense spatial spectrum reuse allowed by caching, for which the requested content is found within a short communication radius, and therefore a large number of simultaneous D2D links can be active on the same time slot. Furthermore, an information theoretic study of the one-hop D2D caching network in the case of worst-case arbitrary demands is provided in [27], where the same throughput scaling

of  $\Theta(M/m)$  is achieved through *inter-session network coded multicasting only* scheme, *spatial reuse only* scheme without inter session coding as in [3], or a combination of both schemes.

A different one-hop caching network topology has been studied in [19]–[22], where a single transmitter (i.e., a base station with all files in the library) serves  $n$  user nodes through a common noiseless link of fixed capacity (bottleneck link). The scheme proposed in [19], [20] partitions each file into packets and each node stores subsets of packets from each file. This provides “side information” at each node such that, for the worst-case demands setting, the base station can compute a multicast *network-coded* messages (transmitted via the common link) such that each node can decode its own requested file from the multicast message and its cached side information. Also in this case, the per-node throughput scaling under the worst-case arbitrary demands model is again given by  $\Theta(M/m)$ , which is remarkably identical with the throughput scaling achieved by single-hop D2D caching networks. In this case, the caching gain is explained in terms of “coded multicasting gain”, i.e., in the ability of turning unicast traffic into coded multicast traffic, such that one transmission satisfies multiple nodes. Further, when the user demands are random and follow a Zipf distribution, the order optimal average rate was characterized in [22]. This behaves as a function of all the system parameters including the number of users, the library size, the memory size and the popularity distributions. Remarkably, in all the regimes of system parameters, the cache memory size  $M$  can provide a *multiplicative gain*, which can be linear, sub-linear, or super-linear, depending on the cases. A number of extensions, such as multiple number of requests, hierarchical network structures, and extension to multiple servers under various topology assumptions, can be found in [23]–[25], [28]–[30].

### B. Contributions

In this paper, we study a natural extension of the single-hop D2D network by allowing multihop transmission. As a related work, a multihop transmission scheme for wireless caching networks has been studied in [16] under the protocol model. The key differences between the present paper and [16] are as follows. First, the main objective of [16] is to minimize the average number of flows passing through each node. Such average number of flows is proportional to the reciprocal of the average per-node throughput only for certain network model; on the other hand, we directly derive the optimal scaling law of the per-node throughput. Second, a centralized and deterministic caching placement was proposed in [16] according to the popularity distribution; in contrast, we present a completely decentralized random caching placement according to a uniform distribution over the whole file library, which is “universal” since it is independent of the specific popularity distribution. Remarkably, while the placement and the achievability scheme of [16] would break under a node layout permutation, such that one should re-allocate the cache content when the nodes are in the presence of node mobility, our scheme is robust since any random permutation of the nodes would generate the same caching distribution, and therefore yields the same throughput scaling with high probability. Third, the file delivery scheme in [16] allows for multihop SD paths (i.e., between nodes caching a given file and nodes demanding such file) of the order of  $\sqrt{n}$ , i.e., the delivery paths are allowed to traverse the whole network. In contrast, in this paper we consider a more practical achievability scheme called *local multihop protocol*, where the number of hops between any SD pairs are independent of the number of nodes and decreases when the storage capacity per node increases.

The proposed caching placement and delivery scheme yield a per-node capacity scaling of  $\Theta(\sqrt{M/m})$ , which is order-optimal when the popularity distribution has the “heavy tail” property (see Definition 3 in Section II-B). For example, this is the case of a Zipf distribution with exponent less than one [26].<sup>1</sup> This result shows that multihop yields a much better per-node capacity scaling than single-hop D2D networks, which is given by  $\Theta(M/m)$ . Furthermore, we show that for other popularity distributions, where the “heavy tail” property is not satisfied or the user demands strongly concentrate, a further improvement of the per-node throughput scaling beyond  $\Theta(\sqrt{M/m})$  is achievable, similar to the case of single-hop D2D networks in [16], [22].

### C. Paper Organization

In Section II, we provide our network model and some definitions to be used throughout the paper. Section III states the main results of this paper on the per-node capacity scaling laws for caching wireless D2D networks. In

<sup>1</sup>Throughout the paper, an “order-optimal” scheme means that it achieves the optimal throughput scaling law within a multiplicative gap of  $n^\epsilon$  for any  $\epsilon > 0$ .

Section IV, we present an achievable scheme which is universal independently from a popularity distribution. An upper bound is provided in Section V. In Section VI, we further improve the throughput scaling laws for a Zipf distribution with exponent larger than a certain threshold. Some concluding remarks are provided in Section VII.

## II. PROBLEM FORMULATION

In this section, we provide the model of the network under investigation and define achievable throughput and system scaling regimes. Generally speaking, for a sequence of events  $\{E_n : n = 1, 2, 3, \dots\}$  we say that  $E_n$  “occurs with high probability” (whp) if  $\lim_{n \rightarrow \infty} \mathbb{P}(E_n) = 1$ , where it is understood that these events are defined in an appropriate probability space, with probability measure generally indicated by  $\mathbb{P}(\cdot)$ . For notational convenience, let  $\stackrel{\text{whp}}{\geq}$  and  $\stackrel{\text{whp}}{\leq}$  denote that the corresponding inequalities hold whp. We will also use the following order notations [31].

- $f(n) = O(g(n))$  if there exist  $c > 0$  and  $n_0 > 0$  such that  $f(n) \leq cg(n)$  for all  $n \geq n_0$ .
- $f(n) = \Omega(g(n))$  if  $g(n) = O(f(n))$ .
- $f(n) = \Theta(g(n))$  if  $f(n) = O(g(n))$  and  $g(n) = O(f(n))$ .

### A. Caching in Wireless Multihop D2D Networks

We consider a wireless multihop D2D network consisting of a population  $\mathcal{U}$  of  $n = |\mathcal{U}|$  nodes, distributed uniformly and independently over a unit square area  $[0, 1] \times [0, 1]$ . Let  $d(u, v)$  denote the distance between nodes  $u, v \in \mathcal{U}$ . It is assumed that communication between nodes follows the *protocol model* of [5]: the transmission from node  $u$  to node  $v$  is successful if and only if: i)  $d(u, v) \leq r$ , and ii) no other active transmitter must be in a circle of radius  $(1 + \Delta)r$  from the receiver node  $v$ . Here,  $r, \Delta > 0$  are given protocol parameters. Also, each node sends its packets at some constant rate  $W$  bits/s/Hz.

We consider a library  $\mathcal{F} = \{W_1, \dots, W_m\}$  of  $m = |\mathcal{F}|$  files (information messages), such that messages  $W_f$  are drawn at random and independently with a uniform distribution over a message set  $\mathbb{F}_2^B$  (binary strings of length  $B$ ), for some arbitrary integer  $B$ . It follows that each file in  $\mathcal{F}$  has entropy  $H(W_f) = B$  bits. Consistently with the current information theoretic literature on caching networks (see Section I), a *caching scheme* is formed by two phases: caching placement and delivery. The file library is generated, and then maintained fixed for a long time. Each network node (user) has an on-board cache memory of capacity  $MB$  bits, i.e., expressed in “equivalent file-size” the cache capacity is equal to  $M$  files. The problem consists of storing information in the caches such that the delivery is as efficient as possible. It is important to note that the caching placement phase is performed beforehand, when the file library is generated. Then, each node  $u \in \mathcal{U}$  demands a file with index  $f_u \in \{1, \dots, m\}$ , and the network must coordinate transmissions (in particular, in this paper we consider multihop D2D operations according to the above defined protocol model), such that each demand is satisfied, i.e., each user  $u$  is able to decode its desired files  $f_u$  from the content of its own cache and from what it receives from the other nodes.

In general, the caching phase is defined by a collection of  $n$  maps  $Z_u : \mathbb{F}_2^{Bm} \rightarrow \mathbb{F}_2^{BM}$ , such that  $Z_u(\mathcal{F})$  is the content of the cache at node  $u \in \mathcal{U}$ . Notice that the cache content is independent of the demand vector  $(f_1, \dots, f_n)$ , reflecting the fact that the caching phase is performed beforehand. In this sense, the caching placement can be regarded as part of the “code set-up”. In the achievability strategies considered in this paper we consider only caching of entire files ( $M$  files per node). As a result, as in [2], [4], [16], the parameter  $B$  (file size) is irrelevant for our achievability results.<sup>2</sup>

Restricting caching to entire files, a caching placement realization is uniquely defined by a bipartite graph  $G = (\mathcal{U}, \mathcal{F}, \mathcal{E})$  with “left” nodes  $\mathcal{U}$ , “right” nodes  $\mathcal{F}$  and edges  $\mathcal{E}$  such that  $(u, f) \in \mathcal{E}$  indicates that file  $W_f$  is assigned to the cache of node  $u$ . A bipartite cache placement graph  $G$  is feasible if the degree of each node  $u \in \mathcal{U}$  is not larger than the cache constraint  $M$ . Let  $\mathcal{G}$  denote the set of all feasible bipartite graphs  $G$ . Then, we define a random cache placement as a probability mass function  $\Pi_c$  over  $\mathcal{G}$ . In particular, if  $\Pi_c$  is induced by randomly and independently assigning  $M$  files to each user node  $u \in \mathcal{U}$ , we say that the cache placement is “decentralized”. For a decentralized caching placement, each user node chooses its own  $M$  files independently of the other nodes.

After the caching functions are computed and the result is stored in the user nodes’ caches, the network is repeatedly used in *rounds*. At each round, each node requests a file in the library, and the network must satisfy

<sup>2</sup>Notice that this is not the case for other schemes such as in [19], [20], [27], where the file size plays an important role (see [32]).



such requests. Since the network resets itself at the end of each delivery cycle, by the renewal–reward theorem [33] the per-node throughput is given by the reciprocal of the time needed to deliver the files (up to a multiplicative constant that depends on  $W$ , on the system bandwidth, and on the file size). Two models for the user demands have been investigated in the literature: *arbitrary* and *random*. In the first case, the users' demand vector  $(f_1, \dots, f_n)$  is arbitrary, and the delivery time is defined for the worst-case demand configuration [18]–[20]. In the second case, the demands are generated at random and the delivery time is averaged over the users' demand distribution [2]–[4], [22]. In this paper, we consider the random demands setting. In particular, we assume that the users' demands are independently and uniformly distributed according to a common probability mass function  $\{p_r(f) : f \in \{1, \dots, m\}\}$ . The probability mass function  $p_r(\cdot)$  is referred to in the following as the *popularity distribution*. Without loss of generality, we assume a descending order between request probabilities, i.e.,  $p_r(i) \geq p_r(j)$  if  $i \leq j$  for  $i, j \in \{1, \dots, m\}$ . For instance, a Zipf popularity distribution with exponent  $\gamma > 0$  is defined by  $p_r(i) = \frac{i^{-\gamma}}{\sum_{j=1}^m j^{-\gamma}}$  for  $i \in \{1, \dots, m\}$  [26].

In the following, all events regarding a network of size  $n = 1, 2, 3, \dots$  are defined on a common probability space generated by the random placement of the nodes, indicated by  $P$ , the random placement of the caches, indicated by  $G$ , and the random demand vector, indicated by  $f$ .

### B. Achievable Throughput and System Scaling Regime

In order to study capacity scaling of the caching wireless multihop D2D network defined before, we consider  $m$  and  $M$  expressed as functions of  $n$  as

$$m = a_1 n^\alpha \text{ and } M = a_2 n^\beta, \quad (1)$$

where  $\alpha, a_1, a_2 > 0$  and  $\beta \in [0, \alpha]$ . We assume that  $a_1 > a_2$  if  $\alpha = \beta$  because the delivery phase becomes trivial if  $\alpha = \beta$  and  $a_1 \leq a_2$  (each node is able to store the entire library  $\mathcal{F}$  for this case).

Before entering the analysis, it is important to clearly define the concept of outage event and symmetric throughput. For a given node placement  $P$ , cache placement  $G$ , and demand vector  $f$ , a feasible delivery strategy consists of a sequence of activation sets, i.e., sets of active transmission links,  $\{\mathcal{A}_t : t = 1, 2, 3, \dots\}$ , such that at each time  $t$  the active links in  $\mathcal{A}_t$  do not violate the protocol model. For a given feasible delivery strategy, we let  $T_n$  denote the corresponding per-node symmetric throughput, i.e., the rate (in bit/s/Hz) at which the request of any node in the network can be served with vanishing probability of error, as  $B \rightarrow \infty$ . If for some node  $u \in \mathcal{U}$  the message probability of error is lower bounded by some positive constant for all  $B$ , we say that the network is in outage. In this case, conventionally, we let  $T_n = 0$ .

A sufficient condition for outage is that there exists some  $u \in \mathcal{U}$  for which  $W_{f_u}$  cannot be reconstructed from the whole cache content  $\{Z_v : v \in \mathcal{U}\}$ . Within the assumptions of our model, it is easy to see that the above condition is also necessary. In fact, by contradiction, notice that if for all  $u \in \mathcal{U}$  the requested message  $W_{f_u}$  can be reconstructed from  $\{Z_v : v \in \mathcal{U}\}$ , then there exists some delivery strategy that conveys all the cache messages to all the user nodes by an appropriate multihop schedule, such that all nodes can decode their own desired file. This is an immediate consequence of the fact that the transmission in any single active link of the network is error-free, and that any node can communicate with any other node, by letting the transmission radius  $r$  sufficiently large. Of course, conveying the global cache content to all nodes may take a very long delivery time, yielding low throughput. As a matter of fact, studying the behavior of the optimal  $T_n$  as  $n \rightarrow \infty$  is precisely the goal pursued in the rest of this paper.

From what said above,  $T_n$  is a random variable, function of  $P$ ,  $f$ , and  $G$ . In general, the cumulative distribution function of  $T_n$  takes on the form:

$$F_{T_n}(x) = \mathbb{P}(T_n = 0)u(x) + F_n^+(x)$$

where  $u(x)$  is the (right-continuous) unit-step function with jump at  $x = 0$ , the probability mass at 0,  $\mathbb{P}(T_n = 0)$ , is the outage probability, and  $F_n^+(x)$  is some right-continuous non-decreasing function of  $x$  continuous at  $x = 0$ , such that  $\lim_{x \rightarrow +\infty} F_n^+(x) = 1 - \mathbb{P}(T_n = 0)$ .

For a given delivery strategy, we say that no outage occurs whp if  $\lim_{n \rightarrow \infty} \mathbb{P}(T_n = 0) = 0$ . In addition, we say that a deterministic sequence  $\{g_n^{\text{lb}}\}$  is achievable if  $T_n \stackrel{\text{whp}}{\geq} g_n^{\text{lb}}$ . Also, a throughput upper bound whp is defined by a deterministic sequence  $\{g_n^{\text{ub}}\}$  such that  $T_n \stackrel{\text{whp}}{\leq} g_n^{\text{ub}}$ . This leads to our definition of achievable throughput scaling laws:

*Definition 1 (Throughput Scaling Law: Achievability):* Given a deterministic sequence  $\{g_n^{\text{lb}}\}$ , the scaling law  $T_n = \Omega(g_n^{\text{lb}})$  is achievable whp if there exists a cache placement strategy and delivery protocol such that  $T_n \geq g_n^{\text{lb}}$  whp and  $\lim_{n \rightarrow \infty} \mathbb{P}(T_n = 0) = 0$ .  $\diamond$

*Definition 2 (Throughput Scaling Law: Converse):* Given a deterministic sequence  $\{g_n^{\text{ub}}\}$ , we say that  $T_n = O(g_n^{\text{ub}})$  is a converse throughput scaling law whp if for any cache placement strategy and delivery protocol  $T_n \leq g_n^{\text{ub}}$  whp.  $\diamond$

Obviously, a tight characterization of the throughput scaling law is obtained when  $T_n = \Omega(g_n^{\text{lb}})$  is achievable whp, and we can exhibit a converse whp  $T_n = O(g_n^{\text{ub}})$  such that when  $g_n^{\text{ub}} = \Theta(g_n^{\text{lb}})$ .

### III. MAIN RESULTS

This section states the main results of this paper. We first introduce throughput scaling laws of caching wireless *multihop* D2D networks achievable for any popularity distribution in Theorem 1 and compare with those of caching wireless *single-hop* D2D networks. In Theorem 2, we then establish upper bounds on throughput scaling laws for a class of heavy-tailed popularity distributions. In Theorem 3, we further improve the throughput scaling laws achievable for a Zipf popularity distribution when its exponent is larger than a certain threshold. For ease of exposition, we partition the entire parameter space into five regimes as follows:

- Regime I:  $\alpha - \beta > 1$ .
- Regime II:  $\alpha - \beta = 1$  and  $a_1 > a_2$ .
- Regime III:  $\alpha - \beta = 1$  and  $a_1 \leq a_2$ .
- Regime IV:  $\alpha - \beta \in (0, 1)$ .
- Regime V:  $\alpha - \beta = 0$  and  $a_1 > a_2$ .

Notice that shifting from Regimes I to V tends to increase the relative caching capability at each node, compared to the library size (recall the relation between  $m$  and  $M$  in (1)).

The following scaling laws hold *universally* for any popularity distribution.

*Theorem 1:* For the caching wireless D2D network defined in Section II, the achievable throughput satisfies whp the scaling laws:

$$T_n = \begin{cases} 0 & \text{for Regimes I and II,} \\ \Omega(n^{-\frac{1}{2}-\epsilon}) & \text{for Regime III,} \\ \Omega(n^{-\frac{\alpha-\beta}{2}-\epsilon}) & \text{for Regime IV,} \\ \Omega(n^{-\epsilon}) & \text{for Regime V,} \end{cases} \quad (2)$$

where  $\epsilon > 0$  is arbitrarily small.

*Proof:* The lower bound for Regimes I and II is trivial. For the non-trivial part, the proof for Regimes IV and V is given in Section IV-A and for Regime III in Section IV-B.  $\blacksquare$

*Corollary 1:* Consider the caching wireless D2D network defined in Section II. If the file delivery is restricted to single-hop transmission, then the achievable throughput satisfies whp the scaling laws:

$$T_n = \begin{cases} 0 & \text{for Regimes I and II,} \\ \Omega(n^{-1}) & \text{for Regime III,} \\ \Omega(n^{-(\alpha-\beta)-\epsilon}) & \text{for Regime IV,} \\ \Omega(n^{-\epsilon}) & \text{for Regime V,} \end{cases} \quad (3)$$

where  $\epsilon > 0$  is arbitrarily small.

*Proof:* The proof is given in Section IV-C.  $\blacksquare$

Fig. 1 compares the achievable throughput scaling laws of the caching wireless D2D network between multihop and single-hop file deliveries in (2) and (3), respectively, where we omitted the term  $n^{-\epsilon}$  for simplicity. Regimes I and II correspond to the case where the overall cached files in the entire network is strictly less than the number of files in the library, i.e.,  $Mn < m$ . Thus, an outage is inevitable even if a centralized caching is used, which results in  $T_n = 0$ . As the relative caching capability increases compared to the library size, i.e.,  $\alpha - \beta$  decreases, each node can find its requested file in the network and thus, a non-zero  $T_n$  is achievable for Regimes III and IV. As it

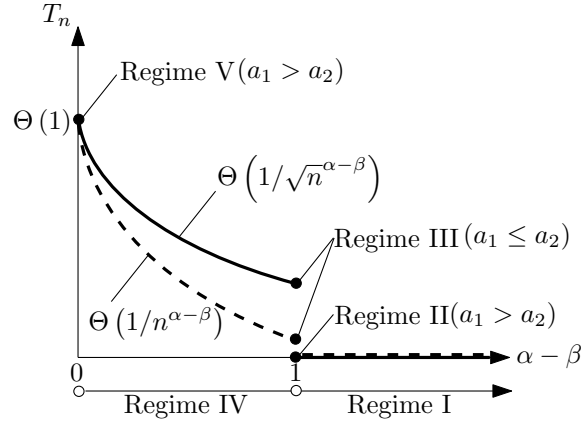


Fig. 1. Achievable throughput scaling laws in (2) for caching wireless multihop D2D networks (solid curve) and (3) for caching wireless single-hop D2D networks (dashed curve).

will be clear from the achievability delivery strategies of Section IV, the geometric interpretation of this behavior is as follows: as  $\alpha - \beta$  decreases (i.e., the storage capacity  $M$  increases), the file delivery distance decreases, such that the network achieves larger and larger spatial reuse (multiple links can be active at the same time, compatibly with the protocol model). As a result,  $T_n$  increases as  $\alpha - \beta$  decreases for both (2) and (3). Finally when  $\alpha = \beta$  (i.e., Regime V), each node can find its requested file from its nearest neighbors. Thus, the delivery distance is  $O(1/\sqrt{n})$  and  $T_n = \Theta(1)$  is achievable.

One of the most important facts is that single-hop file delivery is order-optimal only for Regime V. For almost all parameter space of interest (Regimes III and IV), multihop file delivery significantly improves the throughput by a factor  $\sqrt{\frac{n}{M}}$ . Intuitively, spatial reuse is much more effective with multihop transmissions, namely, we can have more concurrent transmissions in the network. At the same time, the cost of duplicated transmissions by multihop is not very significant comparing with the gains obtained by the simultaneously active links. It is worthwhile to mention that, for a Zipf popularity distribution with  $\gamma < 1$ , our result is well matched with that in [16], even if we use a random caching and a local multihop schemes (see Section IV) rather than a centralized caching and a possibly “whole-network traversing” multihop schemes presented in [16]. Furthermore, due to the universality of the proposed scheme (random and independent caching), the same throughput scaling laws in Theorem 1 and Corollary 1 are achievable for random mobile networks since the network caching distribution is invariant with respect to node permutation.

In order to establish upper bounds on throughput scaling laws, we define a class of popularity distributions with the “heavy tail” property.

*Definition 3 (Heavy-tailed popularity distributions):* Define a class of popularity distributions such that, for any  $0 < c_1 < a_1$ , there exists  $c_2 > 0$  satisfying that

$$\lim_{n \rightarrow \infty} \sum_{i=1}^{c_1 n^\alpha} p_r(i) \leq 1 - c_2, \quad (4)$$

where  $c_1$  and  $c_2$  are some constants and independent of  $n$ .  $\diamond$

*Lemma 1:* The Zipf distribution with exponent less than one (i.e.,  $\gamma \leq 1$ ) [26] satisfies the condition in Definition 3.

*Proof:* Letting  $f(n) = \sum_{i=1}^n i^{-\gamma}$ , we have that

$$\sum_{i=1}^{c_1 n^\alpha} p_r(i) = \sum_{i=1}^{c_1 n^\alpha} \frac{i^{-\gamma}}{\sum_{j=1}^{a_1 n^\alpha} j^{-\gamma}} = \frac{f(c_1 n^\alpha)}{f(a_1 n^\alpha)}.$$

Using the bounds

$$\int_1^n x^{-\gamma} dx \leq f(n) \leq 1 + \int_1^n x^{-\gamma} dx,$$

we have:

$$\lim_{n \rightarrow \infty} \frac{f(c_1 n^\alpha)}{f(a_1 n^\alpha)} \leq \lim_{n \rightarrow \infty} \frac{c_1^{(1-\gamma)} n^{\alpha(1-\gamma)} - \gamma}{a_1^{(1-\gamma)} n^{\alpha(1-\gamma)} - 1},$$

where the upper bound converges to  $\left(\frac{c_1}{a_1}\right)^{(1-\gamma)}$ , which is strictly less than 1 since  $c_1 < a_1$  and  $\gamma < 1$ . ■

For the above class of popularity distributions, ignoring a small portion of requests in the tail of the distribution yields a non-vanishing outage probability. Hence, almost all files in  $\mathcal{F}$  should be cached in the network in order to achieve a non-zero  $T_n$ . This is the main idea underlying the throughput upper bound in the following theorem.

**Theorem 2:** Consider the caching wireless D2D network defined in Section II and assume that demands distribution satisfies the condition in Definition 3. Then the throughput of any scheme must satisfy whp the scaling laws:

$$T_n = \begin{cases} 0 & \text{for Regimes I and II,} \\ O(n^{-\frac{1}{2}+\epsilon}) & \text{for Regime III,} \\ O(n^{-\frac{\alpha-\beta}{2}+\epsilon}) & \text{for Regime IV,} \\ O(1/\log n) & \text{for Regime V,} \end{cases} \quad (5)$$

where  $\epsilon > 0$  is arbitrarily small.

*Proof:* The proof is given in Section V-A for Regimes I and II, Section V-B for Regimes III and IV and Section V-C for Regime V. ■

For all five regimes, the multiplicative gap between the achievable  $T_n$  in Theorem 1 and its upper bound in Theorem 2 is within  $n^\epsilon$  for any arbitrarily small  $\epsilon > 0$ . Therefore, the throughput scaling law depicted in Fig. 1 (solid curve) is order-optimal for the class of heavy-tailed popularity distributions in Definition 3. As we will explain in Section IV, in the parameter regimes of interest, such order-optimal throughput scaling is achievable by fully *decentralized random caching uniformly across  $\mathcal{F}$* . Similarly, from the following corollary, the throughput scaling law depicted in Fig. 1 (dashed curve) is order-optimal for the class of heavy-tailed popularity distributions in Definition 3 when the file delivery is restricted to single-hop transmission.

**Corollary 2:** Consider the caching wireless D2D network defined in Section II and assume that demands distribution satisfies the condition in Definition 3. If the file delivery is restricted to single-hop transmission, then the throughput of any scheme must satisfy whp the scaling laws:

$$T_n = \begin{cases} 0 & \text{for Regimes I and II,} \\ O(n^{-1+\epsilon}) & \text{for Regime III,} \\ O(n^{-(\alpha-\beta)+\epsilon}) & \text{for Regime IV,} \\ O(1/\log n) & \text{for Regime V,} \end{cases} \quad (6)$$

where  $\epsilon > 0$  is arbitrarily small.

*Proof:* The proof is given in Section V-D. ■

As the deviation between the request probabilities in the popularity distribution increases (e.g.,  $\gamma$  increases in a Zipf distribution), the condition in Definition 3 may not be satisfied. In this case, it can be expected that the throughput scaling law may be improved by a more refined caching strategy, biased towards the files requested with higher probability. In particular, we consider caching only an appropriately optimized subset of most popular files, while guaranteeing that the aggregate “tail” probability of the least popular files vanishes, such that we still get no outage whp. In the following, we demonstrate the above statement for a Zipf popularity distribution with  $\gamma > 1 + \frac{1}{\alpha}$ .

**Theorem 3:** Consider the caching wireless D2D network defined in Section II and assume that the demands follow a Zipf popularity distribution with exponent  $\gamma > 1 + \frac{1}{\alpha}$ . Then the achievable throughput satisfies whp the scaling law:

$$T_n = \Omega\left(n^{-\frac{1-\min(1, \beta+1-1/(\gamma-1))}{2}-\epsilon}\right) \quad \text{for Regime IV,} \quad (7)$$

where  $\epsilon > 0$  is arbitrarily small.

*Proof:* The proof is given in Section VI. ■



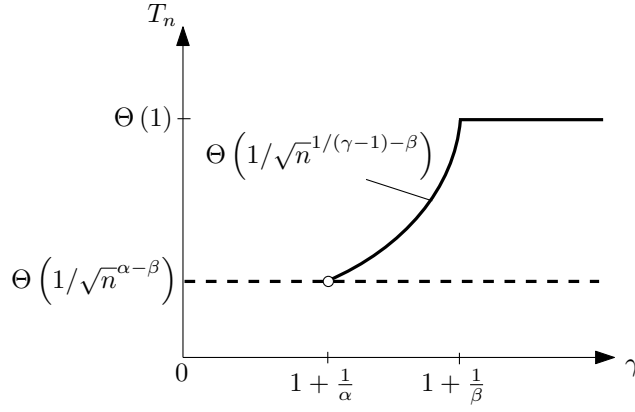


Fig. 2. Achievable throughput scaling laws in (7) (solid curve) and (2) (dashed curve) with respect to  $\gamma$  for Regime IV.

In Fig. 2, we compare the improved scaling laws in (7) and the scaling laws in (2) for Regime IV, where the term  $n^{-\epsilon}$  is omitted. When the demands follow a Zipf popularity distribution, the improved throughput scaling  $\Theta(1/\sqrt{n}^{1/(\gamma-1)-\beta})$  is achievable instead of  $\Theta(1/\sqrt{n}^{\alpha-\beta})$  in (2) if  $\gamma > 1 + \frac{1}{\alpha}$  and eventually  $\Theta(1)$  scaling is achievable when  $\gamma \geq 1 + \frac{1}{\beta}$  (see Fig. 2). As we will explain in Section VI, a fully decentralized random caching still achieves the improved throughput scaling laws in Theorem 3, by appropriately reducing the effective library size, i.e., *decentralized random caching uniformly across a subset of popular files*. Namely, in this regime, we can rule out some files from the library, each of which probability is small enough such that an outage does not occur with probability approaching one as  $n \rightarrow \infty$ .

**Comparison with the results in [16]:** In order to compare our results with these summarized in Table III of [16], we need to let  $\alpha \leq 1$  ( $n = \Omega(m)$ ) and  $M$  be a constant or  $\beta = 0$  ( $M = \Theta(1)$ ), then by ignoring the  $\epsilon$  in the scaling law exponent, we obtain that  $T_n = \Omega\left(\sqrt{\frac{M}{n^{\frac{1}{\gamma-1}}}}\right) = \Omega\left(n^{-\frac{1/(\gamma-1)}{2}}\right)$  under the condition  $\gamma > 1 + \frac{1}{\alpha}$  from Theorem 3, which can be either better or worse than the results in [16]. For example, if we let  $\alpha = 1$  and  $nM - m = \Theta(1)$ , then the throughput in [16] is  $\Omega\left(\frac{1}{\sqrt{n}}\right)$ , which is smaller than  $\Omega\left(n^{-\frac{1/(\gamma-1)}{2}}\right)$  for  $\gamma > 2$ , which is feasible since  $\alpha = 1$ . Remarkably, in this regime, a simple decentralized strategy consisting of caching the files at random with a uniform distribution over the most popular files, while discarding the tail of the distribution, can achieve a better throughput than the centralized caching scheme of [16]. On the other hand, for  $\alpha < 1$ , the throughput in [16] behaves as  $\Omega(1)$ , which is better than  $\Omega\left(n^{-\frac{1/(\gamma-1)}{2}}\right)$ . In this case, the decentralized random caching strategy might not be sufficient to achieve order optimality under Definition 1, i.e., the symmetric rate under no outage. Whether it is possible to achieve order-optimal throughput scaling with decentralized random caching, allowing for more general caching distributions (not just uniform over a subset of most probable files) is an interesting question which is left for future research.

#### IV. UNIVERSALLY ACHIEVABLE THROUGHPUT

In this section, we prove Theorem 1. In particular, we present file placement policies and transmission protocols for Regimes III, IV, and V and analyze their achievable throughput scaling laws.

##### A. Regimes IV and V

In this subsection, we prove that

$$T_n = n^{-\frac{\alpha-\beta}{2}-\epsilon} \quad (8)$$

is achievable whp for Regimes IV and V, where  $\epsilon > 0$  is arbitrarily small.

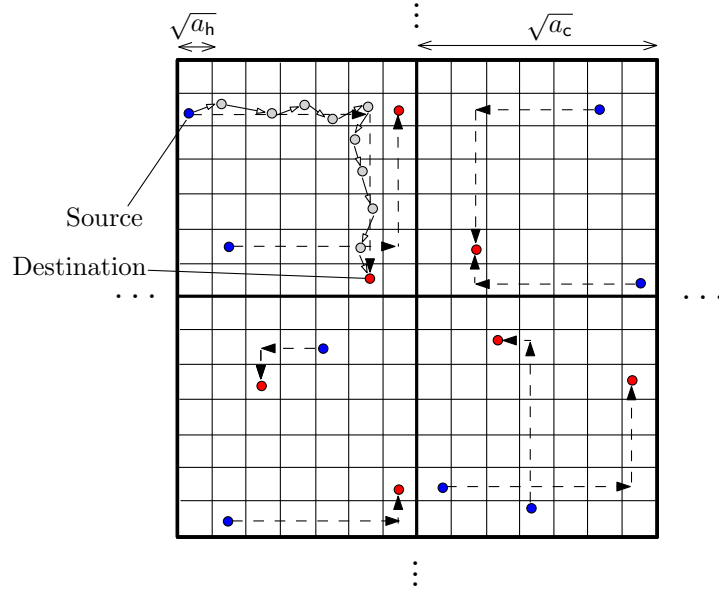


Fig. 3. The proposed multihop routing protocol for file delivery after the source node selection.

1) *File placement policy and transmission protocol*: In these regimes, a *decentralized* file placement and a *local* multihop protocol are proposed as follows.

**Decentralized file placement**: Each node  $u$  stores  $M$  distinct files in its cache, chosen uniformly at random from the library  $\mathcal{F}$ , independently of other nodes.

**Local multihop protocol**: We first explain how each node finds its source node having the requested file (**source node selection**):

- Divide the entire network into square *traffic cells* of area  $a_c = n^{-\eta}$  for some  $\eta \in [0, 1)$ , where  $\eta$  will be determined later on.
- Each node chooses one of the nodes having the requested file in the same traffic cell as its *source* node. If there are multiple candidates, choose one of them uniformly at random.

From Definition 1 and the above source node selection, all nodes should find their source nodes within their own traffic cells whp, in order to achieve a non-zero  $T_n$ . Lemma 3 below characterizes such a condition of the area of traffic cell  $a_c$  (i.e.,  $\eta$ ) such as  $\eta \in [0, 1 - (\alpha - \beta))$ .

For the ease of exposition, we refer to the pair formed by a node and its source node as *source-destination (SD) pair*. Notice that in our model, each SD pair is located in the same traffic cell while in the conventional wireless ad-hoc network, SD pairs are randomly located over the entire network. Thanks to caching, we can reduce the distance of each SD pair (see Lemma 3). Also, differently from the conventional ad-hoc network model, each node can be a source node of multiple destinations, which make the throughput analysis more complicated (see Lemma 5).

Next, we explain the proposed multihop transmission scheme for the file delivery between  $n$  SD pairs, see also Fig. 3 (**multihop transmission**):

- Divide each traffic cell into square *hopping cells* of area  $a_h = \frac{2 \log n}{n}$ .
- Define the horizontal data path (HDP) and the vertical data path (VDP) of a SD pair as the horizontal line and the vertical line connecting a source node to its destination node, respectively. Each source node transmits the requested file to its destination by first hopping to the adjacent hopping cells on its HDP and then on its VDP.<sup>3</sup>

<sup>3</sup>If a source node and its destination node are in the same hopping cell, then the source node directly transmits to its destination.

- Time division multiple access (TDMA) scheme is used with reuse factor  $J$  for which each hopping cell is activated only once out of  $J$  time slots.
- A transmitter node in each active hopping cell sends a file (or fragment of a file) to a receiver node in an adjacent hopping cell. Round-robin is used for all transmitter nodes in the same hopping cell.

In this scheme, each hopping cell should contain at least one node for relaying as in [5], [34], which is satisfied whp since  $a_h = \frac{2 \log n}{n}$  (see Lemma 2 (a)).

*Lemma 2:* The following properties hold whp:

- Partition the network region  $[0, 1] \times [0, 1]$  into cells of area  $\frac{2 \log n}{n}$ . Then the number of nodes in each cell is between 1 and  $4 \log n$ .
- Partition the network region  $[0, 1] \times [0, 1]$  into cells of area  $n^{-a}$ , where  $a \in [0, 1)$ . For any  $\delta > 0$ , the number of nodes in each cell is between  $(1 - \delta)n^{1-a}$  and  $(1 + \delta)n^{1-a}$ .

*Proof:* The proofs of first and second properties are given in [34, Lemma 1] and [13, Lemma 4.1], respectively. ■

*Lemma 3:* Suppose Regimes IV and V. If  $\eta \in [0, 1 - (\alpha - \beta))$ , then all nodes are able to find their source nodes within their traffic cells whp.

*Proof:* Let  $A_i$  denote the event that node  $i$  establishes its source node within its traffic cell, where  $i \in [1 : n]$ . Then, we have:

$$\begin{aligned} \mathbb{P}(\cap_{i \in [1:n]} A_i) &= 1 - \mathbb{P}(\cup_{i \in [1:n]} A_i^c) \\ &\geq 1 - \sum_{i \in [1:n]} \mathbb{P}(A_i^c) \\ &\stackrel{\text{whp}}{\geq} 1 - n \left( \frac{m - M}{m} \right)^{(1-\delta)na_c}, \end{aligned} \quad (9)$$

where the first inequality follows from the union bound and the second inequality is due to the fact that the number of nodes in each traffic cell is lower bounded by  $(1 - \delta)na_c$  whp (see Lemma 2 (b)) and hence,  $\mathbb{P}(A_i^c) \stackrel{\text{whp}}{\leq} \left( \frac{m - M}{m} \right)^{(1-\delta)na_c}$ .

Thus, for Regime IV,

$$\mathbb{P}(\cap_{i \in [1:n]} A_i) \stackrel{\text{whp}}{\geq} 1 - n \left( \left( 1 - \frac{a_2}{a_1} \frac{1}{n^{\alpha-\beta}} \right)^{\frac{a_1}{a_2} n^{\alpha-\beta}} \right)^{\frac{a_2(1-\delta)}{a_1} n^{1-\eta-\alpha+\beta}} \quad (10)$$

and from the fact that

$$\lim_{n \rightarrow \infty} \left( 1 - \frac{a_2}{a_1} \frac{1}{n^{\alpha-\beta}} \right)^{\frac{a_1}{a_2} n^{\alpha-\beta}} = \frac{1}{e}, \quad (11)$$

$\mathbb{P}(\cap_{i \in [1:n]} A_i) \rightarrow 1$  as  $n \rightarrow \infty$ , since  $\eta < 1 - \alpha + \beta$  is assumed in this lemma. Similarly, for Regime V,

$$\mathbb{P}(\cap_{i \in [1:n]} A_i) \stackrel{\text{whp}}{\geq} 1 - n \left( \frac{a_1 - a_2}{a_1} \right)^{(1-\delta)n^{1-\eta}}, \quad (12)$$

which again converges to one as  $n \rightarrow \infty$ , since  $a_1 > a_2$  for this regime and  $\eta < 1 - \alpha + \beta$  is assumed in this lemma. In conclusion, all nodes are able to find their source nodes within their traffic cells whp under the condition where  $\eta \in [0, 1 - (\alpha - \beta))$ . ■

2) *Achievable throughput:* We now show that the proposed scheme in Section IV-A1 achieves (8) whp for Regimes IV and V. From Lemma 3, we assume  $\eta \in [0, 1 - (\alpha - \beta))$  to achieve a non-zero  $T_n$  by the proposed scheme in Section IV-A1. The following lemmas are instrumental to proof.

*Lemma 4:* Suppose Regimes IV and V and assume that  $\eta \in [0, 1 - (\alpha - \beta))$ . Let  $R_n$  denote the aggregate rate achievable for any hopping cell. If  $J \geq (2\lceil(1 + \Delta)\sqrt{5}\rceil + 1)^2$ , then  $R_n = \frac{W}{J}$  is achievable.

*Proof:* This lemma is a well-known property, e.g., see [34, Lemma 2]. For completeness, we briefly review proof steps here. Consider an arbitrary transmission pair consisting of a transmitter node and its receiver node illustrated in Fig. 4. Clearly, the hopping distance is upper bounded by  $\sqrt{5a_h}$  and hence, we choose the transmission range

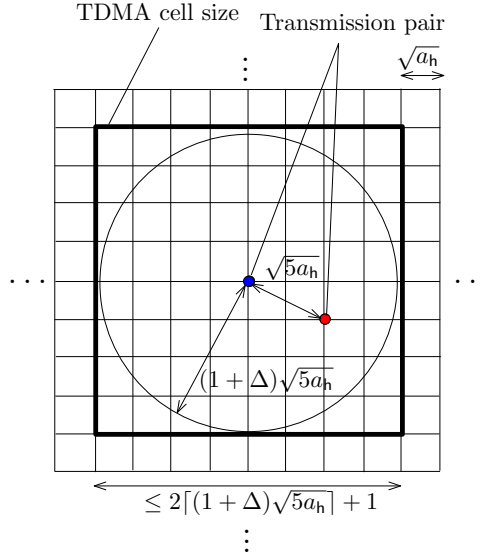


Fig. 4. TDMA cell size from the protocol model.

$r = \sqrt{5a_h}$  in the protocol model. Thus, the transmission is successful if there is no node simultaneously transmitting within the distance of  $(1 + \Delta)\sqrt{5a_h}$  from the receiver node. This is satisfied if  $J \geq (2\lceil(1 + \Delta)\sqrt{5}\rceil + 1)^2$ . That is, the aggregate rate of  $\frac{W}{J}$  is achievable if  $J \geq (2\lceil(1 + \Delta)\sqrt{5}\rceil + 1)^2$ . Since this holds for all hopping cells,  $R_n = \frac{W}{J}$  is achievable if  $J \geq (2\lceil(1 + \Delta)\sqrt{5}\rceil + 1)^2$ . ■

**Lemma 5:** Suppose Regimes IV and V and assume that  $\eta \in [0, 1 - (\alpha - \beta))$ . For  $\epsilon > 0$  arbitrarily small, each node can be a source node of at most  $n^{1-\eta-(\alpha-\beta)+\epsilon}$  nodes in its traffic cell whp.

*Proof:* Let  $B_i(k)$  denote the event that node  $i$  becomes a source node for less than  $k$  nodes. Denote  $n_1 = (1 + \delta)n^{1-\eta}$ . Then, we have

$$\begin{aligned}
 \mathbb{P}(\cap_{i \in [1:n]} B_i(k)) &= 1 - \mathbb{P}(\cup_{i \in [1:n]} B_i^c(k)) \\
 &\stackrel{\text{whp}}{\geq} 1 - n \sum_{j=k}^{n_1} \binom{n_1}{j} \left(\frac{M}{m}\right)^j \left(1 - \frac{M}{m}\right)^{n_1-j} \\
 &\geq 1 - n \exp\left(-n_1 D\left(\frac{k}{n_1} \parallel \frac{M}{m}\right)\right) \\
 &= 1 - n \exp\left(-k \log\left(\frac{km}{n_1 M}\right) - (n_1 - k) \log\left(\frac{m(n_1 - k)}{n_1(m - M)}\right)\right) \\
 &= 1 - n \underbrace{\exp(-k) \left(\frac{km}{n_1 M}\right)^{-\ln(2)}}_{:=A} \underbrace{\exp(-(n_1 - k)) \left(\frac{m(n_1 - k)}{n_1(m - M)}\right)^{-\ln(2)}}_{:=B} \quad (13)
 \end{aligned}$$

if  $\frac{M}{m} < \frac{k}{n_1} < 1$ , where  $D(a \parallel b) = a \log(\frac{a}{b}) + (1 - a) \log(\frac{1-a}{1-b})$  denotes the relative entropy for  $a, b \in (0, 1)$ . Here the first inequality follows from the union bound and holds whp since the number of nodes in each traffic cell is upper bounded by  $n_1$  whp from Lemma 2 (b), and the second inequality is due to the fact that for  $X \sim B(n, p)$ ,

$$\mathbb{P}(X \geq k) \leq \exp(-nD(k/n \parallel p)) \text{ if } p < k/n < 1. \quad (14)$$

First consider Regime IV. Suppose that  $k = n^\tau$  for  $\tau \in (0, 1]$ . Then the condition  $\frac{M}{m} < \frac{k}{n_1} < 1$  is given by  $\frac{a_2(1+\delta)}{a_1} n^{1-\eta-(\alpha-\beta)} < n^\tau < (1 + \delta)n^{1-\eta}$ , which is satisfied as  $n$  increases if

$$1 - \eta - (\alpha - \beta) < \tau \leq 1 - \eta. \quad (15)$$

Since  $\tau > 0$ ,

$$A = n \exp(-n^\tau) \left( \frac{a_1}{a_2(1+\delta)} n^{\tau-1+\eta+(\alpha-\beta)} \right)^{-\ln(2)} \quad (16)$$

converges to zero as  $n$  increases. Furthermore

$$B = \exp(-((1+\delta)n^{1-\eta} - n^\tau)) \left( \frac{a_1 n^\alpha ((1+\delta)n^{1-\eta} - n^\tau)}{(1+\delta)n^{1-\eta}(a_1 n^\alpha - a_2 n^\beta)} \right)^{-\ln(2)} \quad (17)$$

converges to zero as  $n$  increases if  $\tau \leq 1 - \eta$ . In summary,  $\mathbb{P}(\cap_{i \in [1:n]} B_i(n^\eta))$  converges to zero as  $n$  increases if (15) holds. Therefore,  $\mathbb{P}(\cap_{i \in [1:n]} B_i(n^{1-\eta-(\alpha-\beta)+\epsilon})) \rightarrow 0$  as  $n \rightarrow \infty$  by setting  $\tau = 1 - \eta - (\alpha - \beta) + \epsilon$  for  $\epsilon > 0$  arbitrarily small, implying that each node becomes a source node of at most  $n^{1-\eta-(\alpha-\beta)+\epsilon}$  nodes whp for Regime IV.

Now consider Regime V. Suppose again that  $k = n^\tau$  for  $\tau \in (0, 1]$ . Then the condition  $\frac{M}{m} < \frac{k}{n_1} < 1$  is given by  $\frac{a_2(1+\delta)}{a_1} n^{1-\eta} < n^\tau < (1+\delta)n^{1-\eta}$ , which is satisfied by setting  $\tau = 1 - \eta$  since  $a_1 > a_2$  for Regime V so that we can find  $\delta > 0$  satisfying  $\frac{a_2(1+\delta)}{a_1} < 1$ , see Lemma 2 (b). For this case, we have

$$A = n \exp(-n^{1-\eta}) \left( \frac{a_1}{a_2(1+\delta)} \right)^{-\ln(2)} \quad (18)$$

and

$$B = \exp(-\delta n^{1-\eta}) \left( \frac{a_1 \delta}{(a_1 - a_2)(1+\delta)} \right)^{-\ln(2)}.$$

Hence,  $A \rightarrow 0$  and  $B \rightarrow 0$  as  $n \rightarrow \infty$  since  $\eta < 1$ . Therefore, each node becomes a source node of at most  $n^{1-\eta}$  nodes whp for Regime V. ■

Based on Lemma 5, we derive an upper bound on the number of data paths that should be carried by each hopping cell in the following lemma, which is directly related to achievable throughput scaling laws.

**Lemma 6:** Suppose Regimes IV and V and assume that  $\eta \in [0, 1 - (\alpha - \beta))$ . For  $\epsilon > 0$  arbitrarily small, each hopping cell is required to carry at most  $n^{\frac{3(1-\eta)}{2} - (\alpha-\beta) + \epsilon}$  data paths whp.

*Proof:* First consider the number of HDPs that must be carried by an arbitrary hopping cell, denoted by  $N_{\text{hdp}}$ . By assuming that all HDPs of the nodes in the hopping cells located at the same horizontal line pass through the considered hopping cell, we have an upper bound on  $N_{\text{hdp}}$ . Since the total area of these cells is given by

$$\sqrt{a_c a_h} = \sqrt{n^{-\eta} \frac{2 \log n}{n}} = n^{\frac{1-\eta}{2}} \frac{1}{\sqrt{2 \log n}} \frac{2 \log n}{n}, \quad (19)$$

the number of nodes in that area is upper bounded by

$$n^{\frac{1-\eta}{2}} \frac{1}{\sqrt{2 \log n}} 2 \log n = n^{\frac{1-\eta}{2}} \sqrt{2 \log n} \quad (20)$$

whp from Lemma 2 (a). Moreover, each of these nodes may become a source node of multiple nodes within the same traffic cell. Therefore, from Lemma 5 and (20)

$$\begin{aligned} N_{\text{hdp}} &\stackrel{\text{whp}}{\leq} n^{1-\eta-(\alpha-\beta)+\epsilon'} n^{\frac{1-\eta}{2}} \sqrt{2 \log n} \\ &= n^{\frac{3(1-\eta)}{2} - (\alpha-\beta) + \epsilon'} \sqrt{2 \log n} \end{aligned} \quad (21)$$

for  $\epsilon' > 0$  arbitrarily small. The same analysis holds for VDPs. In conclusion, each hopping cell carries at most  $n^{\frac{3(1-\eta)}{2} - (\alpha-\beta) + \epsilon}$  data paths whp for  $\epsilon > 0$  arbitrarily small, which completes the proof. ■

We are now ready to prove that (8) is achievable whp for Regimes IV and V. Let  $\epsilon' > 0$  be an arbitrarily small constant satisfying that  $1 - (\alpha - \beta) - \epsilon' > 0$ , which is valid for Regimes IV and V since  $\alpha - \beta \in [0, 1)$ . Then set  $\eta = 1 - (\alpha - \beta) - \epsilon'$ , which determines the size of each traffic cell. From Lemma 3, every node can find its source node within its traffic cell whp. From Lemma 4, setting  $J = (2\lceil(1 + \Delta)\sqrt{5}\rceil + 1)^2$ , each hopping cell is able to achieve the aggregate rate of

$$R_n = W / \left( 2\lceil(1 + \Delta)\sqrt{5}\rceil + 1 \right)^2. \quad (22)$$



Furthermore, from Lemma 6, the number of data paths that each hopping cell needs to perform is upper bounded by

$$n^{\frac{3(1-\eta)}{2} - (\alpha-\beta) + \epsilon'} = n^{\frac{\alpha-\beta}{2} + \frac{5}{2}\epsilon'} \quad (23)$$

whp, where we used  $\eta = 1 - (\alpha - \beta) - \epsilon'$ .

Since each hopping cell serves multiple data paths using round-robin fashion, each data path is served with a rate of at least (22) divided by (23) whp. Therefore, an achievable per-node throughput is given by

$$T_n = \frac{W}{(2\lceil(1+\Delta)\sqrt{5}\rceil + 1)^2} n^{-\frac{\alpha-\beta}{2} - \frac{5}{2}\epsilon'} \geq n^{-\frac{\alpha-\beta}{2} - \epsilon} \quad (24)$$

whp for  $\epsilon > 0$  arbitrarily small. In conclusion, (8) is achievable whp for Regimes IV and V.

### B. Regime III

In this subsection, we prove that

$$T_n = n^{-\frac{1}{2} - \epsilon} \quad (25)$$

is achievable whp assuming that  $\alpha - \beta = 1$  and  $a_1 = a_2$ , where  $\epsilon > 0$  is arbitrarily small. Hence the same  $T_n$  is also achievable whp for  $\alpha - \beta = 1$  and any  $a_1 \leq a_2$ , which corresponds to Regime III.

From now on, assume that  $\alpha - \beta = 1$  and  $a_1 = a_2$ . For this case, the total number of files that can be stored by  $n$  nodes (i.e., the total number of files stored in the network) is exactly the same as the number of files in the library (i.e.,  $nM = m$ ). We propose a *centralized* file placement and a *globally* multihop protocol as follows.

**Centralized file placement:** It can be seen that a distributed file placement might result in an outage, as seen from the analysis in Lemma 3. Instead, we employ a simple centralized file placement for which all distinct  $m$  files (in the library) are randomly stored in the total memories of  $n$  nodes. Hence, the network can contain all  $m$  files, thus being able to avoid an outage.

**Globally multihop protocol:** As explained before, the traffic cell should be equal to the entire network (i.e.,  $\eta = 0$  in Section IV-A), in order to avoid an outage. Namely,  $n$  SD pairs are located over the entire network. Hence, we can expect the same scaling result with the conventional wireless ad-hoc network in [5], namely, no caching gain is expected.

We briefly explain how to achieve (25) whp, since the procedures of proof are almost similar to Regimes IV and V. Similarly to Lemma 5, we can show that each node is able to be a source node of at most  $n^\epsilon$  nodes whp for  $\epsilon > 0$  arbitrarily small. Then, following the analysis in Section IV-A2, we can easily prove that (25) is achievable whp for Regime III.

### C. Single-Hop File Delivery

In this subsection, we prove Corollary 1. First consider Regimes IV and V. We apply the same file placement and source node selection policy described in Section IV-A1. Then Lemma 3 holds guaranteeing no outage whp by setting  $\eta = 1 - (\alpha - \beta) - \epsilon'$ , where  $\epsilon' > 0$  be an arbitrarily small constant satisfying that  $1 - (\alpha - \beta) - \epsilon' > 0$ . Consider the file delivery. Instead of multihop routing within each traffic cell, each source directly transmits the required file to its destination within each traffic cell. Then, from the same analysis in Lemma 4, each traffic cell achieves the aggregate rate of  $R_n = \frac{W}{(2\lceil(1+\Delta)\sqrt{5}\rceil + 1)^2}$  by TDMA between traffic cells with reuse factor  $(2\lceil(1+\Delta)\sqrt{5}\rceil + 1)^2$ . Since there are at most  $(1+\delta)n^{(\alpha-\beta)+\epsilon'}$  nodes in each traffic cell whp (Lemma 2 (b)), the rate of  $\frac{R_n}{1+\delta}n^{-(\alpha-\beta)-\epsilon'}$  is achievable whp for each file delivery. Therefore,  $T_n = n^{-(\alpha-\beta)-\epsilon}$  is achievable whp for Regimes IV and V, where  $\epsilon > 0$  is arbitrarily small.

Now consider Regime III. As the same reason in IV-B, we assume  $\alpha - \beta = 1$  and  $a_1 = a_2$ , and then apply the same file placement and source node selection policy described in Section IV-B, which guarantees no outage. Then, from the direct file delivery by time-sharing between  $n$  SD pairs,  $T_n = \frac{1}{n}$  is achievable whp for Regime III.

## V. CONVERSE

In this section, we prove the upper bounds in Theorem 2 assuming that the popularity distribution satisfies the condition in Definition 3. We first introduce the following technical lemma.

*Lemma 7:* Let  $X$  follow a binomial distribution with parameters  $l$  and  $p$ , i.e.,  $X \sim B(l, p)$ . Then, for  $k \in [0 : lp]$ ,

$$\mathbb{P}(X \leq k) \leq \exp\left(-\frac{1}{2p} \frac{(lp - k)^2}{l}\right). \quad (26)$$

*Proof:* The proof follows immediately from the Chernoff bound.  $\blacksquare$

### A. Regimes I and II

We introduce the following lemma, which demonstrates that a non-vanishing outage probability is inevitable for Regimes 1 and 2 even if centralized caching were allowed. Therefore, a non-vanishing outage probability implied by Lemma 8 yields that  $T_n = 0$  whp for Regimes I and II.

*Lemma 8:* Suppose Regimes I and II. Let  $N_{\text{out},1}$  denote the number of nodes that they cannot find their requested files in the entire network. Then, we have  $N_{\text{out},1} \geq c_3 n$  whp for some constant  $c_3 > 0$  independent of  $n$ .

*Proof:* The total number of files that are able to be stored by the entire network is given by  $nM = a_2 n^{1+\beta}$ . Hence the probability that each node cannot find its requested file in the entire network is lower bounded by

$$1 - \sum_{i=1}^{a_2 n^{1+\beta}} p_r(i) := p_{\text{out},1}. \quad (27)$$

Then, for  $\mu \in [0, p_{\text{out},1}]$ , we have

$$\begin{aligned} \mathbb{P}(N_{\text{out},1} \geq \mu n) &\stackrel{(a)}{\geq} \sum_{i=\mu n}^n \binom{n}{i} p_{\text{out},1}^i (1 - p_{\text{out},1})^{n-i} \\ &\geq 1 - \sum_{i=1}^{\mu n} \binom{n}{i} p_{\text{out},1}^i (1 - p_{\text{out},1})^{n-i} \\ &\stackrel{(b)}{\geq} 1 - \exp\left(-\frac{(p_{\text{out},1} - \mu)^2}{2p_{\text{out},1}} n\right), \end{aligned} \quad (28)$$

where (a) follows from (27) and the fact that each node requires a file independent of other nodes and (b) follows from Lemma 7. Here, the condition  $\mu \in [0, p_{\text{out},1}]$  is required to apply Lemma 7.

Now consider  $p_{\text{out},1}$  defined in (27). Notice that  $a_2 n^{1+\beta} < a_1 n^\alpha$  as  $n \rightarrow \infty$  for both regimes because  $\alpha - \beta > 1$  for Regime I and  $\alpha - \beta = 1$  and  $a_1 > a_2$  for Regime II. Hence, from Definition 3,  $\lim_{n \rightarrow \infty} p_{\text{out},1} \geq c_4$  for some constant  $c_4 > 0$  independent of  $n$ . Then setting  $\mu = \frac{c_4}{2}$  in (28), which satisfies  $\mu \in [0, p_{\text{out},1}]$  as  $n \rightarrow \infty$ , yields that  $\mathbb{P}(N_{\text{out},1} \geq \frac{c_4}{2} n) \rightarrow 1$  as  $n \rightarrow \infty$ . Therefore,  $N_{\text{out},1} \geq c_3 n$  whp for some constant  $c_3 > 0$  independent of  $n$ .  $\blacksquare$

### B. Regimes III and IV

The key ingredient to establish the upper bounds in Theorem 2 for Regimes III and IV is to characterize the minimum distance for file transmission that a non-zero fraction of SD pairs must go through, which is given in Lemma 9 below. Then, as a consequence of the protocol model which does not allow concurrent transmission within a circle of radius  $(1 + \Delta)r$  around each intended receiver, we are able to determine how many SD pairs can be simultaneously activate at a given time slot, which is directly related to the desired throughput upper bounds.

*Lemma 9:* Suppose Regimes III and IV. For  $\epsilon > 0$  arbitrarily small, let  $N_{\text{out},2}$  denote the number of nodes that they cannot find their requested files within the distance of  $n^{-\frac{1-(\alpha-\beta)}{2}-\epsilon}$  from their positions. Then, we have  $N_{\text{out},2} \geq c_5 n$  whp for some constant  $c_5 > 0$  independent of  $n$ .

*Proof:* Let  $\epsilon' > 0$  be an arbitrarily small constant satisfying that  $1 - (\alpha - \beta) + \epsilon' \in [0, 1]$ , which is valid for Regimes III and IV since  $\alpha - \beta \in (0, 1]$ . For simplicity, denote  $\zeta = \frac{1-(\alpha-\beta)+\epsilon'}{2}$ . Let  $N_{\text{file}}$  be the total number of files that are able to be stored by the area of radius  $n^{-\zeta}$ . From Lemma 2 (b), the number of nodes in that area is

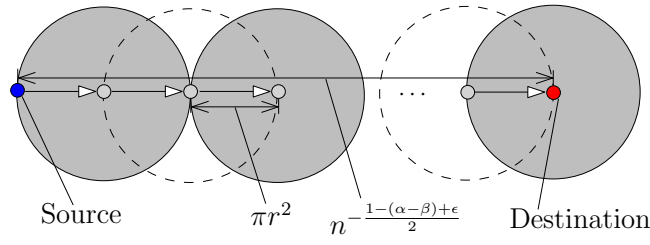


Fig. 5. A lower bound on the exclusive area occupied by the multihop transmission of a SD pair with distance  $n^{-\frac{1-(\alpha-\beta)+\epsilon}{2}}$ .

upper bounded by  $(1 + \delta)n^{1-2\zeta}$  whp because  $2\zeta \in [0, 1)$ . Hence  $N_{\text{file}} \leq (1 + \delta)n^{1-2\zeta}M = a_2(1 + \delta)n^{\alpha-\epsilon'}$  whp. Then the probability that each node cannot find its requested file within the radius of  $n^{-\zeta}$  is lower bounded by

$$1 - \sum_{i=1}^{N_{\text{file}}} p_r(i) \stackrel{\text{whp}}{\geq} 1 - \sum_{i=1}^{a_2(1+\delta)n^{\alpha-\epsilon'}} p_r(i) := p_{\text{out},2}. \quad (29)$$

Then similarly to (28), we have

$$\mathbb{P}(N_{\text{out},2} \geq \mu n) \stackrel{\text{whp}}{\geq} 1 - \exp\left(-\frac{(p_{\text{out},2} - \mu)^2}{2p_{\text{out},2}}n\right) \quad (30)$$

for  $\mu \in [0, p_{\text{out},2}]$ . From Definition 3,  $\lim_{n \rightarrow \infty} p_{\text{out},2} \geq c_6$  for some constant  $c_6 > 0$  independent of  $n$ . More specifically, we can apply Definition 3 because  $a_2(1 + \delta)n^{\alpha-\epsilon'} < a_1n^\alpha$  as  $n \rightarrow \infty$ . Hence setting  $\mu = \frac{c_6}{2}$  in (30), which satisfies  $\mu \in [0, p_{\text{out},2}]$  as  $n \rightarrow \infty$ , yields that  $\mathbb{P}(N_{\text{out},2} \geq \frac{c_6}{2}n) \rightarrow 1$  as  $n \rightarrow \infty$ . Therefore,  $N_{\text{out},2} \geq c_5n$  whp for some constant  $c_5 > 0$  independent of  $n$ . ■

Based on Lemma 9, we can prove that the throughput of any scheme must satisfy

$$T_n \stackrel{\text{whp}}{\leq} n^{-\frac{\alpha-\beta}{2}+\epsilon} \quad (31)$$

for Regimes III and IV, where  $\epsilon > 0$  is arbitrarily small. Specifically, from Lemma 9, for  $\epsilon' > 0$  arbitrarily small, there are at least  $c_5n$  SD pairs whose distances are larger than  $n^{-\frac{1-(\alpha-\beta)}{2}-\epsilon'}$  whp, where  $c_5 > 0$  is some constant and independent of  $n$ . Then, we restrict only on the delivery of the requests of such SD pairs, obtaining clearly an upper bound on the per-node throughput. First, we consider the exclusive area (i.e., the area to prohibit the transmission for other SD pairs) occupied by the multihop transmission of a SD pair with distance  $n^{-\frac{1-(\alpha-\beta)}{2}-\epsilon'}$ . In order to obtain a lower bound on such area, we assume that  $\Delta = 0$  and each receiver node is located at the distance of  $r$  from its transmitter node along with the SD line (see Fig. 5). Then, the exclusive area is lower bounded (i.e., only taking the shaded areas in Fig. 5) such as

$$\frac{2\pi r^2 n^{-\frac{1-(\alpha-\beta)}{2}-\epsilon'}}{2r} = \pi r n^{-\frac{1-(\alpha-\beta)}{2}-\epsilon'}. \quad (32)$$

Hence, the maximum number of SD pairs guaranteeing a rate of  $W$  over the entire network of a unit area is upper bounded by  $\frac{1}{\pi r} n^{\frac{1-(\alpha-\beta)}{2}+\epsilon'}$  whp. As a result, the sum throughput  $S_n$  (summing the rate of all users) is upper bounded by  $S_n \stackrel{\text{whp}}{\leq} \frac{W}{\pi r} n^{\frac{1-(\alpha-\beta)}{2}+\epsilon'}$ . Notice that for a given sum throughput  $S_n$ , the symmetric per-user rate is trivially upper bounded by  $T_n \leq S_n/n$ . Hence, we have

$$T_n \leq \frac{S_n}{n} \stackrel{\text{whp}}{\leq} \frac{W}{\pi r} n^{\frac{1-(\alpha-\beta)}{2}+\epsilon'}. \quad (33)$$

That the above bound on  $T_n$  increases as  $r$  decreases. On the other hand, it was shown in [5, Section V] that the absence of isolated nodes is a necessary condition for a non-zero  $T_n$  requiring that

$$r \stackrel{\text{whp}}{\geq} c_7 \sqrt{\log n/n} \quad (34)$$

for some constant  $c_7 > 0$  independent of  $n$ . Therefore, from (33) and (34), we have an upper bound on the per-node throughput as

$$\begin{aligned} T_n &\stackrel{\text{whp}}{\leq} \frac{W}{\pi c_7} n^{-\frac{\alpha-\beta}{2}-\frac{\log \log n}{2 \log n}+\epsilon'} \\ &\leq n^{-\frac{\alpha-\beta}{2}+\epsilon} \end{aligned} \quad (35)$$

for  $\epsilon > 0$  arbitrarily small. In conclusion, the upper bound in (31) holds whp for Regimes III and IV.

### C. Regime V

In this subsection, we prove that the throughput of any scheme must satisfy

$$T_n \stackrel{\text{whp}}{\leq} \frac{c_8}{\log n} \quad (36)$$

for Regime V, where  $c_8 > 0$  is some constant independent of  $n$ . The following lemma shows that at least a constant fraction of nodes have to download their requested files from other nodes, which will be used as the key ingredient to prove the upper bound in (36).

*Lemma 10:* Suppose Regime V. Let  $N_{\text{out},3}$  denote the number of nodes that they cannot find their requested files in their own cache memories. Then, we have  $N_{\text{out},3} \geq c_9 n$  whp for some constant  $c_9 > 0$  independent of  $n$ .

*Proof:* Similar to the proof in Lemmas 8 and 9, we have

$$\mathbb{P}(N_{\text{out},3} \geq \mu n) \geq 1 - \exp\left(-\frac{(p_{\text{out},3} - \mu)^2}{2p_{\text{out},3}}n\right) \quad (37)$$

for  $\mu \in [0, p_{\text{out},3}]$ , where  $p_{\text{out},3} = 1 - \sum_{i=1}^{a_2 n^\alpha} p_r(i)$ . Since  $a_2 < a_1$  for Regime V,  $\lim_{n \rightarrow \infty} p_{\text{out},3} \geq c_{10}$  for some constant  $c_{10} > 0$  independent of  $n$  from Definition (3). Hence setting  $\mu = \frac{c_{10}}{2}$  in (37) yields that  $\mathbb{P}(N_{\text{out},3} \geq \frac{c_{10}}{2}n) \rightarrow 1$  as  $n \rightarrow \infty$ . Therefore,  $N_{\text{out},3} \geq c_9 n$  whp for some constant  $c_9 > 0$  independent of  $n$ . ■

From Lemma 10, a non-vanishing fraction of nodes have to download their requested files from other nodes and, as a result, (34) should be satisfied for successful file delivery, see [5, Section V]. From the protocol model, then, the rate of each file delivery is upper bounded by  $W$  bits/sec/Hz and there are at most  $\frac{1}{\pi c_7^2} \frac{n}{\log n}$  concurrent file deliveries in the network whp, from the bound in (34). Therefore,  $S_n \stackrel{\text{whp}}{\leq} \frac{W}{\pi c_7^2} \frac{n}{\log n}$  and  $T_n \leq \frac{S_n}{n} \stackrel{\text{whp}}{\leq} \frac{W}{\pi c_7^2} \frac{1}{\log n}$ . In conclusion the upper bound in (36) holds whp for Regime V.

### D. Single-Hop File Delivery

In this subsection, we prove Corollary 2. For Regimes I and II, Lemma 8 still holds, resulting that  $T_n = 0$  whp for these regimes. Also, the same argument in Section V-C holds, resulting that (36) whp for Regime V.

Now consider Regimes III and IV. From Lemma 9, if the file delivery for SD pairs with distance at least  $n^{-\frac{1-(\alpha-\beta)}{2}-\epsilon}$  is restricted to single-hop transmission, the exclusive area occupied by each of those SD pairs is lower bounded by

$$\pi n^{-(1-(\alpha-\beta))-\epsilon'} \quad (38)$$

whp for  $\epsilon' > 0$  arbitrarily small. Then, as the same analysis in Section V-B, we have  $S_n \stackrel{\text{whp}}{\leq} \frac{W}{\pi} n^{1-(\alpha-\beta)+\epsilon'}$ , resulting that  $T_n \stackrel{\text{whp}}{\leq} n^{-(\alpha-\beta)+\epsilon}$  for  $\epsilon > 0$  arbitrarily small for these regimes.

## VI. IMPROVED ACHIEVABLE THROUGHPUT

In this section, we prove Theorem 3 by assuming that user demands follow a Zipf popularity distribution with exponent  $\gamma > 1 + \frac{1}{\alpha}$ .

### A. File Placement and Delivery

Similar to the case of Regime IV in Section IV-A, i.e.,  $\alpha - \beta \in (0, 1)$ , a *distributed* file placement and a *local* multihop protocol are performed. In order to describe the proposed file placement, let  $\epsilon_c > 0$  be an arbitrarily small constant satisfying that

$$\beta + 1 - \frac{1}{\gamma} - \epsilon_c > 0, \quad (39)$$

which is valid because  $\beta + 1 - \frac{1}{\gamma-1} > 1 - (\alpha - \beta) > 0$ , where the first inequality holds from the assumption  $\gamma > 1 + \frac{1}{\alpha}$  and the second inequality holds for Regime IV since  $\alpha - \beta \in (0, 1)$ . Then, define

$$n_2 = n^{1 - \min(1, \beta + 1 - 1/(\gamma-1)) + \epsilon_c/2} \quad (40)$$

and let  $\mathcal{F}_{\text{sub}} \subseteq \mathcal{F}$  denote the subset of the first (most probable)  $(Mn_2)$  files in the library. During the file placement phase, each node stores  $M$  distinct files in its cache, chosen uniformly at random from  $\mathcal{F}_{\text{sub}}$  independently of other nodes.

During the file delivery phase, the same local mulithop described in Section IV-A is performed. To determine the size of each traffic cell, we set

$$\eta = \min \left( 1, \beta + 1 - \frac{1}{\gamma-1} \right) - \epsilon_c, \quad (41)$$

which is valid since  $\eta \in (0, 1)$  from (39). Then the number of nodes in each traffic cell is upper bounded by

$$(1 + \delta)n_2n^{\epsilon_c/2} \quad (42)$$

whp and lower bounded by

$$(1 - \delta)n_2n^{\epsilon_c/2} \quad (43)$$

whp from Lemma 2 (b).

### B. Achievable Throughput

In this subsection, we prove that

$$T_n = n^{-\frac{1 - \min(1, \beta + 1 - 1/(\gamma-1))}{2} - \epsilon} \quad (44)$$

is achievable whp for Regime IV, where  $\epsilon > 0$  is arbitrarily small. The overall procedure is similar to the case of Regime IV in Section IV-A. In the following, we first show that all nodes can find their required files within their traffic cells whp by setting  $\eta$  as in (41).

*Lemma 11:* Suppose Regime IV and  $\eta = \min \left( 1, \beta + 1 - \frac{1}{\gamma-1} \right) - \epsilon_c$ . Then all nodes are able to find their sources within their traffic cells whp.

*Proof:* Denote  $P = \sum_{i=1}^{Mn_2} p_r(i)$ , where the definition of  $n_2$  is given by (40). For  $i \in [1 : n]$ , denote  $\mathcal{N}_i \subseteq [1 : n]$  as the set of nodes in the traffic cell that node  $i$  is included and  $A_i$  as the event that node  $i$  establishes its source node in  $\mathcal{N}_i$ . Then the outage probability  $\mathbb{P}(A_i^c)$  is given by

$$\begin{aligned} \mathbb{P}(A_i^c) &= \mathbb{P}(\text{node } i \text{ requests } f_i \in \mathcal{F}_{\text{sub}}) \mathbb{P}(f_i \notin \cup_{j \in \mathcal{N}_i} \mathcal{M}_j | \text{node } i \text{ requests } f_i \in \mathcal{F}_{\text{sub}}) \\ &\quad + \mathbb{P}(\text{node } i \text{ requests } f_i \notin \mathcal{F}_{\text{sub}}) \\ &= P \left( \frac{Mn_2 - M}{Mn_2} \right)^{|\mathcal{N}_i|} + (1 - P) \\ &\stackrel{\text{whp}}{\leq} P \left( 1 - \frac{1}{n_2} \right)^{(1-\delta)n_2n^{\epsilon_c/2}} + (1 - P), \end{aligned} \quad (45)$$

where  $|\mathcal{N}_i|$  denotes the cardinality of  $\mathcal{N}_i$ . Here, the second equality holds since each node  $i$  stores  $M$  distinct files in its local memory  $\mathcal{M}_i$ , chosen uniformly at random from  $\mathcal{F}_{\text{sub}}$  independently of other nodes and the inequality holds from (43).

Then, following the analysis in (9), we have:

$$\begin{aligned} \mathbb{P} \left( \cap_{i \in [1:n]} A_i \right) &\stackrel{\text{whp}}{\geq} 1 - n \left( P \left( 1 - \frac{1}{n_2} \right)^{(1-\delta)n_2n^{\epsilon_c/2}} + (1 - P) \right) \\ &\geq 1 - n \left( 1 - \frac{1}{n_2} \right)^{(1-\delta)n_2n^{\epsilon_c/2}} - n(1 - P) \\ &= 1 - n \left( \left( 1 - \frac{1}{n_2} \right)^{n_2} \right)^{(1-\delta)n^{\epsilon_c/2}} - n(1 - P). \end{aligned} \quad (46)$$



From (11) and the fact that  $n_2 \rightarrow \infty$  as  $n \rightarrow \infty$ , the term  $n \left( \left(1 - \frac{1}{n_2}\right)^{n_2} \right)^{(1-\delta)n^{\epsilon_c/2}}$  in (46) converges to zero as  $n$  increases. Furthermore,

$$\begin{aligned} n(1-P) &\stackrel{(a)}{=} n \left( \frac{\sum_{i=Mn_2+1}^m i^{-\gamma}}{\sum_{i=1}^m i^{-\gamma}} \right) \\ &\stackrel{(b)}{\leq} n \left( \frac{\int_{Mn_2}^m x^{-\gamma} dx}{\int_1^m x^{-\gamma} dx} \right) \\ &\stackrel{(c)}{=} \frac{n^{\alpha(1-\gamma)+1} - n^{(\beta+1-\min(1,\beta+1-1/(\gamma-1))+\epsilon_c/2)(1-\gamma)+1}}{n^{\alpha(1-\gamma)} - 1}, \end{aligned} \quad (47)$$

where (a) follows from the definition of  $P$ , (b) follows because  $\sum_{i=a+1}^b i^{-\gamma} \leq \int_a^b x^{-\gamma} dx$  and  $\sum_{i=a}^b i^{-\gamma} \geq \int_a^b x^{-\gamma} dx$ , and (c) follows from the definition of  $n_2$ . Notice that  $n^{\alpha(1-\gamma)}$  and  $n^{\alpha(1-\gamma)+1}$  in (47) converge to zero as  $n$  increases since  $\gamma > 1 + \frac{1}{\alpha}$ . Also,

$$\begin{aligned} n^{(\beta+1-\min(1,\beta+1-1/(\gamma-1))+\epsilon_c/2)(1-\gamma)+1} &\leq n^{(\beta+1-(\beta+1-1/(\gamma-1))+\epsilon_c/2)(1-\gamma)+1} \\ &= n^{\frac{\epsilon_c/2}{1-\gamma}} \end{aligned} \quad (48)$$

converges to zero as  $n$  increases because  $\frac{\epsilon_c/2}{1-\gamma} < 0$ . Therefore, the term  $n(1-P)$  in (46) also converges to zero as  $n$  increases.

In conclusion, from (46),  $\mathbb{P}(\cap_{i \in [1:n]} A_i)$  converges to zero as  $n$  increases. ■

As proved in Lemma 11, we set  $\eta = \min\left(1, \beta + 1 - \frac{1}{\gamma-1}\right) - \epsilon_c$  from now on, which determines the size of each traffic cell guaranteeing no outage at all nodes whp. Notice that Lemma 4 holds regardless of the file popularity distribution. Hence, a non-vanishing aggregate rate is achievable for any hopping cell by TDMA between hopping cells with some constant reuse factor. We then derive the same statement in Lemma 5 in the following lemma.

*Lemma 12:* Suppose Regime IV and  $\eta = \min\left(1, \beta + 1 - \frac{1}{\gamma-1}\right) - \epsilon_c$ . Then each node can be a source node of at most  $n^{\epsilon_c}$  nodes in its traffic cell whp.

*Proof:* Let  $B_i(k)$  denote the event that node  $i$  becomes a source node for less than  $k$  nodes. From the same analysis in (13), we have

$$\begin{aligned} \mathbb{P}(\cap_{i \in [1:n]} B_i(k)) &\stackrel{\text{whp}}{\geq} 1 - n \sum_{j=k}^{(1+\delta)n_2n^{\epsilon_c/2}} \binom{(1+\delta)n_2n^{\epsilon_c/2}}{j} \left(\frac{1}{n_2}\right)^j \left(1 - \frac{1}{n_2}\right)^{(1+\delta)n_2n^{\epsilon_c/2}-j} \\ &\geq 1 - n \exp\left(- (1+\delta)n_2n^{\epsilon_c/2} D\left(\frac{k}{(1+\delta)n_2n^{\epsilon_c/2}} \parallel \frac{1}{n_2}\right)\right) \\ &= 1 - n \exp\left(-k \log\left(\frac{k}{(1+\delta)n^{\epsilon_c/2}}\right)\right) \\ &\quad \cdot \exp\left(-((1+\delta)n_2n^{\epsilon_c/2} - k) \log\left(\frac{(1+\delta)n_2n^{\epsilon_c/2} - k}{(1+\delta)n_2n^{\epsilon_c/2} - (1+\delta)n^{\epsilon_c/2}}\right)\right) \\ &= 1 - n \underbrace{\exp(-k) \left(\frac{k}{(1+\delta)n^{\epsilon_c/2}}\right)^{-\ln(2)}}_{:=C} \\ &\quad \cdot \underbrace{\exp(-((1+\delta)n_2n^{\epsilon_c/2} - k)) \left(\frac{(1+\delta)n_2n^{\epsilon_c/2} - k}{(1+\delta)n_2n^{\epsilon_c/2} - (1+\delta)n^{\epsilon_c/2}}\right)^{-\ln(2)}}_{:=D} \end{aligned} \quad (49)$$

if  $\frac{1}{n_2} < \frac{k}{(1+\delta)n_2n^{\epsilon_c/2}} < 1$ , where  $D(a||b) = a \log(\frac{a}{b}) + (1-a) \log(\frac{1-a}{1-b})$  denotes the relative entropy for  $a, b \in (0, 1)$ .

Suppose that  $k = n^{\epsilon_c}$ . Then the condition  $\frac{1}{n_2} < \frac{k}{(1+\delta)n_2n^{\epsilon_c/2}} < 1$  is satisfied because  $(1+\delta)n^{\epsilon_c/2} < n^{\epsilon_c} < (1+\delta)n^{1-\min(1,\beta+1-1/(\gamma-1))+\epsilon_c}$ . Furthermore, we have

$$C = n \exp(-n^{\epsilon_c}) \left(\frac{n^{\epsilon_c/2}}{1+\delta}\right)^{-\ln(2)} \rightarrow 0 \text{ as } n \rightarrow \infty. \quad (50)$$

Similarly, from the definition of  $n_2$  in (40),

$$D = \exp(-((1 + \delta)n^{1-\min(1, \beta+1-1/(\gamma-1))+\epsilon_c} - n^{\epsilon_c})) \cdot \left( \frac{(1 + \delta)n^{1-\min(1, \beta+1-1/(\gamma-1))+\epsilon_c} - n^{\epsilon_c}}{(1 + \delta)n^{1-\min(1, \beta+1-1/(\gamma-1))+\epsilon_c} - (1 + \delta)n^{\epsilon_c/2}} \right)^{-\ln(2)} \rightarrow 0 \text{ as } n \rightarrow \infty \quad (51)$$

because  $((1 + \delta)n^{1-\min(1, \beta+1-1/(\gamma-1))+\epsilon_c} - n^{\epsilon_c}) \rightarrow \infty$  as  $n \rightarrow \infty$  and  $\frac{(1+\delta)n^{1-\min(1, \beta+1-1/(\gamma-1))+\epsilon_c} - n^{\epsilon_c}}{(1+\delta)n^{1-\min(1, \beta+1-1/(\gamma-1))+\epsilon_c} - (1+\delta)n^{\epsilon_c/2}} \rightarrow 1$  as  $n \rightarrow \infty$ . Therefore,  $\mathbb{P}(\cap_{i \in [1:n]} B_i(n^{\epsilon_c})) \rightarrow 1$  as  $n \rightarrow \infty$ , meaning that each node becomes a source node of at most  $n^{\epsilon_c}$  nodes whp. ■

**Lemma 13:** Suppose Regime IV and  $\eta = \min\left(1, \beta + 1 - \frac{1}{\gamma-1}\right) - \epsilon_c$ . For  $\epsilon > 0$  arbitrarily small, each hopping cell is required to carry at most  $n^{\frac{1-\min(1, \beta+1-1/(\gamma-1))}{2} + \epsilon}$  data paths whp.

*Proof:* Let  $N_{\text{hdp}}$  denote the number of HDPs that must be carried by an arbitrary hopping cell. From the same analysis in (19) and (20) and Lemma 12, we have

$$N_{\text{hdp}} \stackrel{\text{whp}}{\leq} n^{\frac{1-\min(1, \beta+1-1/(\gamma-1))}{2} + \epsilon_c} \sqrt{2 \log n}. \quad (52)$$

The same analysis holds for VDPs. In conclusion, each hopping cell carries at most  $n^{\frac{1-\min(1, \beta+1-1/(\gamma-1))}{2} + \epsilon}$  data paths whp for  $\epsilon > 0$  arbitrarily small, which completes the proof. ■

We are now ready to prove that (44) is achievable whp for Regime IV. From Lemma 11, every node can find its source node within its traffic cell whp. From Lemma 4, setting  $J = (2\lceil(1 + \Delta)\sqrt{5}\rceil + 1)^2$ , each hopping cell is able to achieve the aggregate rate in (22). Furthermore, from Lemma 13, the number of data paths that each hopping cell needs to perform is upper bounded by

$$n^{\frac{1-\min(1, \beta+1-1/(\gamma-1))}{2} + \epsilon'} \quad (53)$$

whp for  $\epsilon'$  arbitrarily small. Therefore, an achievable per-node throughput is given by at least (22) divided by (53) whp. In conclusion, (44) is achievable whp for Regime IV.

## VII. CONCLUDING REMARKS

We considered a wireless ad-hoc network in which nodes have cached information from a library of possible files. For such network, we proposed an order-optimal caching policy (i.e., file placement policy) and multihop transmission protocol for a broad class of heavy-tailed popularity distributions including a Zipf distribution with exponent less than one. Interestingly, we showed that a distributed uniform random caching is order-optimal for the parameter regimes of interest as long as the total number of files in the library is less than the overall caching memory size in the network. i.e.,  $\alpha - \beta \in (0, 1]$ . Also, it was shown that a multihop transmission provides a significant throughput gain over one-hop direct transmission as in the conventional wireless ad-hoc networks. As a future work, the complete characterization of the optimal throughput scaling laws for this network with random demands following a Zipf distribution with an arbitrary exponent  $\gamma$  (in particular, with  $\gamma \geq 1$ ) remains to be determined. In this regime, decentralized uniform random caching over a subset of most probable files is generally not order-optimal, and gains can be achieved by more refined random decentralized caching policies. Whether these can achieve the same scaling laws of the deterministic centralized strategy of [16] in all regimes remains also to be seen.

## REFERENCES

- [1] “Cisco visual networking index: Global mobile data traffic forecast update, 2-12-2017,” *Cisco Public Information*, 2015.
- [2] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, “Femtocaching: Wireless content delivery through distributed caching helpers,” *IEEE Trans. Inf. Theory*, vol. 59, pp. 8402–8413, Dec. 2013.
- [3] M. Ji, G. Caire, and A. F. Molisch, “Optimal throughput-outage trade-off in wireless one-hop caching networks,” in *Proc. IEEE Int. Symp. on Information Theory (ISIT)*, Istanbul, Turkey, Jul. 2013.
- [4] —, “The throughput-outage tradeoff of wireless one-hop caching networks,” *arXiv preprint arXiv:1312.2637*, 2013.
- [5] P. Gupta and P. R. Kumar, “The capacity of wireless networks,” *IEEE Trans. Inf. Theory*, vol. 46, pp. 388–404, Mar. 2000.
- [6] S. Kulkarni and P. Viswanath, “A deterministic approach to throughput scaling in wireless networks,” *IEEE Trans. Inf. Theory*, vol. 50, pp. 1041–1049, Jun. 2004.
- [7] X.-Y. Li, “Multicast capacity of wireless ad hoc networks,” *IEEE/ACM Transactions on Networking*, vol. 17, pp. 950–961, Jun. 2009.

- [8] M. Franceschetti, O. Dousse, D. Tse, and P. Thiran, "Closing the gap in the capacity of wireless networks via percolation theory," *IEEE Trans. Inf. Theory*, vol. 53, pp. 1009–1018, Mar. 2007.
- [9] L.-L. Xie and P. R. Kumar, "A network information theory for wireless communication: Scaling laws and optimal operation," *IEEE Trans. Inf. Theory*, vol. 50, pp. 748–767, May 2004.
- [10] F. Xue, L.-L. Xie, and P. R. Kumar, "The transport capacity of wireless networks over fading channels," *IEEE Trans. Inf. Theory*, vol. 51, pp. 834–847, Mar. 2005.
- [11] S. Shakkottai, X. Liu, and R. Srikant, "The multicast capacity of large multihop wireless networks," *IEEE/ACM Trans. Networking*, vol. 18, pp. 1691–1700, Dec. 2010.
- [12] U. Niesen, P. Gupta, and D. Shah, "The balanced unicast and multicast capacity regions of large wireless networks," *IEEE Trans. Inf. Theory*, vol. 56, pp. 2249–2271, May 2010.
- [13] A. Özgür, O. Lévêque, and D. Tse, "Hierarchical cooperation achieves optimal capacity scaling in ad hoc networks," *IEEE Trans. Inf. Theory*, vol. 53, pp. 3549–3572, Oct. 2007.
- [14] A. Özgür and O. Lévêque, "Throughput-delay tradeoff for hierarchical cooperation in ad hoc wireless networks," *IEEE Trans. Inf. Theory*, vol. 56, pp. 1369–1377, Mar. 2010.
- [15] S.-N. Hong and G. Caire, "Beyond scaling laws: On the rate performance of dense device-to-device wireless networks," *IEEE Trans. Inf. Theory*, vol. 61, pp. 4735–4750, Sep. 2015.
- [16] S. Gkitzenis, G. S. Paschos, and L. Tassiulas, "Asymptotic laws for joint content replication and delivery in wireless networks," *IEEE Trans. Inf. Theory*, vol. 59, pp. 2760–2776, May 2013.
- [17] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *arXiv preprint arXiv:1305.5216*, 2013.
- [18] —, "Fundamental limits of distributed caching in D2D wireless networks," in *Proc. IEEE Information Theory Workshop (ITW)*, Seville, Spain, Sep. 2013.
- [19] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, pp. 2856–2867, May 2014.
- [20] —, "Decentralized coded caching attains order-optimal memory-rate tradeoff," *IEEE/ACM Trans. Networking*, vol. 23, pp. 1029–1040, Aug. 2014.
- [21] U. Niesen and M. A. Maddah-Ali, "Coded caching with nonuniform demands," *arXiv preprint arXiv:1308.0178*, 2013.
- [22] M. Ji, A. Tulino, J. Llorca, and G. Caire, "Order-optimal rate of caching and coded multicasting with random demands," *arXiv preprint arXiv:1502.03124*, 2015.
- [23] —, "Caching and coded multicasting: Multiple groupcast index coding," in *Proc. IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Atlanta, GA, Dec. 2014.
- [24] N. Karamchandani, U. Niesen, M. A. Maddah-Ali, and S. Diggavi, "Hierarchical coded caching," *arXiv preprint arXiv:1403.7007*, 2014.
- [25] J. Hachem, N. Karamchandani, and S. Diggavi, "Multi-level coded caching," *arXiv preprint arXiv:1404.6563*, 2014.
- [26] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: Evidence and implications," in *Proc. IEEE INFOCOM*, New York, NY, Mar. 1999.
- [27] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *arXiv preprint arXiv:1405.5336*, 2014.
- [28] M. Ji, A. Tulino, J. Llorca, and G. Caire, "Caching-aided coded multicasting with multiple random requests," in *Proc. IEEE Information Theory Workshop (ITW)*, Jerusalem, Israel, Apr. 2015.
- [29] M. Ji, M. F. Wong, A. Tulino, J. Llorca, G. Caire, M. Effros, and M. Langberg, "On the fundamental limits of caching in combination networks," in *Proc. IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Stockholm, Sweden, Jun. 2015.
- [30] S. P. Shariatpanahi, S. A. Motahari, and B. H. Khalaj, "Multi-server coded caching," *arXiv preprint arXiv:1503.00265*, 2015.
- [31] D. E. Knuth, "Big Omicron and big Omega and big Theta," *ACM SIGACT News*, vol. 8, pp. 18–24, Apr.-Jun. 1976.
- [32] K. Shanmugam, M. Ji, A. Tulino, J. Llorca, and A. G. Dimakis, "Finite length analysis of caching-aided coded multicasting," in *Proc. 52nd Annu. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Oct. 2014.
- [33] H. Tijms, *A first course in stochastic models*. John Wiley & Sons, 2003.
- [34] A. El Gamal, J. Mammen, B. Prabhakar, and D. Shah, "Optimal throughput-delay scaling in wireless networks—Part I: The fluid model," *IEEE Trans. Inf. Theory*, vol. 52, pp. 2568–2592, Jun. 2006.