Throughput–Outage Analysis of Cache-Aided Wireless Multi-Hop D2D Networks

Ming-Chun Lee, *Member, IEEE*, Mingyue Ji, *Member, IEEE*, and Andreas F. Molisch, *Fellow, IEEE*

***Abstract*—Cache-aided wireless device-to-device (D2D) networks have demonstrated promising performance improvement for video distribution compared to conventional distribution methods. Understanding the fundamental scaling behavior of such networks is thus importance. Recently, based on real-world data, it has been observed that the popularity distribution should be modeled by a Mandelbrot-Zipf (MZipf) distribution, instead of the common Zipf distribution. We thus in this work investigate the throughput-outage performance for cache-aided wireless D2D network adopting multi-hop communications, with the MZipf popularity distribution for file requests and Poisson point process for user distribution. Considering the case that Zipf factor is larger than one, we first propose an achievable content caching and delivery scheme and analyze its performance. Then, by showing that the achievable performance is tight to the proposed outer bound, we show that an optimal scaling law for cache-aided wireless multi-hop D2D networks is obtained.**

I. Introduction

Over the last years, progress of semiconductor technology has made memory one of the cheapest hardware resources. Accordingly, caching at the wireless edge has emerged as a promising approach for significantly improving the efficiency and quality of video distribution [2]. The idea of caching is to trade cheap memory resources for expensive bandwidth resources by caching video files close to the prospective users. This principle, combined with the *asynchronous content reuse* and *concentrated popularity* that are general characteristics of video requests [2], has rendered caching at the wireless edge a widely explored method for video distribution [3].

Among the commonly discussed scenarios of caching at the wireless edge, cache-aided wireless D2D networks have been shown to effectively improve the video distribution [4]–[7], and have been widely discussed in recent years [3], [8]. While most papers are devoted to improving cache-aided D2D networks in practical settings, there exist papers that focus on understanding the fundamental properties and limits of cache-aided D2D networks [4]–[7], [9]–[12]. These papers use scaling law analysis to characterize how the network/user performance scales as the number of users N tends to infinity. Their results thus provide us the performance trend as well as means of comparison between fundamentally different communication frameworks. This paper provides a contribution to this range of investigation.

A. Related Literature

The throughput scaling law analysis for wireless D2D (or ad-hoc) networks has been subject to many investigations since the seminal work of Gupta and Kumar [13]. Since then, papers have been published for understanding the throughput scaling law [14], [15]. Their results showed that the throughput of wireless D2D networks scales with $\Theta\left(\frac{1}{\sqrt{N}}\right)$. Meanwhile, cache-aided D2D/ad-hoc networks have also been studied by the computer science community, e.g., [16], [17]. However, the fundamental scaling laws and optimality considerations did not draw much attention.

Only recently did the fundamental properties of cache-aided D2D/ad-hoc networks start to draw more attention. In [18], the scaling law of the maximum expected throughput was characterized for single-hop cache-aided D2D networks considering a Zipf popularity distribution and a protocol model for transmission between nodes; however, it did not characterize the outage probability. To resolve this limitation, [4] showed that in single-hop networks, the throughput per node can scale with $\Theta\left(\frac{S}{M}\right)$ with negligibly small outage probability when a heavy-tailed Zipf popularity distribution is considered, where M is the file library size and S is the per-user memory size. This result was later generalized in [7] by adopting a more practical and general modeling for popularity distribution, namely the Mandelbrot-Zipf (MZipf) distribution, where the distribution is characterized by the Zipf factor γ and the plateau factor q. In [9], the throughput scaling law of networks with multi-hop communications was characterized with the assumption of user locations on a grid, while the tradeoff between throughput and outage was not explicitly investigated. Ref. [6] provided an achievable throughput scaling law for networks with multi-hop D2D under the condition that the outage is vanishing. In [12], an upper bound for the throughput was proposed. However, similar to [9], the tradeoff between throughput and outage performance was not characterized.

B. Contributions

In this work, we focus on the the scaling law analysis for the throughput-outage performance of uncoded cache-aided D2D with multi-hop communications. As mentioned above, recent

M.-C. Lee was with Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089, USA. He is now with Institute of Communications Engineering, National Chiao Tung University, Hsinchu 30010, Taiwan (email: mingchunlee@nctu.edu.tw).

A. F. Molisch is with Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089, USA (email: molisch@usc.edu).

M. Ji is with Department of Electrical and Computer Engineering, University of Utah, Salt Lake City, UT 84112, USA (email: mingyue.ji@utah.edu).

This work was supported in part by the National Science Foundation (NSF) under grants CNS-1816699, CCF-1423140, CCF-1817154 and SpecEES-1824558, and by the Ministry of Science and Technology (MOST) of Taiwan under grant MOST 109-2218-E-009-002.

An extended journal version of this paper is provided in [1].

work [7] has shown that the MZpif distribution better fits a large real-world data set than the Zpif distribution assumed in [6], [9], [12]. Thus, our paper aims to provide scaling law analysis under the MZipf distribution assumption. Note that this paper is the first one to provide scaling law analysis for cache-aided wireless multi-hop D2D networks considering the MZipf popularity distribution.

In this work, we use Poisson point process (PPP) to model the user distribution and use the MZipf distribution to model the popularity distribution of video requests [7]. We focus on the case that the Zipf factor of the MZipf distribution is larger than 1, i.e., $\gamma > 1$, indicating the light-tailed popularity distribution.[1] We assume a decentralized random caching policy [19] and derive the achievable scaling law and its outer bound for throughput-outage performance in the regimes that the outage probability is either negligibly small or converging to zero, corresponding to the practical requirement that the desirable outage probability of the network should be small.

We show that the derived achievable per-user throughput scaling law and its outer bound are tight. Thus, our achievable throughput-outage scaling law is optimum. Specifically, we show that when the outage probability is negligibly small, the throughput per user scales according to $\Theta\left(\sqrt{\frac{S}{q}}\right)$.[2] Such result is interesting as it indicates that the performance is dominated by the plateau factor q, i.e., the total number of very popular files, instead of the total number of files M. Note that according to the dataset in [7], q is much smaller than M in practice.

II. Network Setup

We consider a random dense network where users are placed according to a PPP within a unit square-shaped area $[0,1] \times [0,1]$. We assume that the density of the PPP is N. As a result, the average number of users in the network is N and the number of users n in the network is a random variable following the Poisson distribution. Accordingly, the probability that the network has n users is:

$$\mathbb{P}_N(\mathsf{n}=n) = \frac{N^n}{n!}e^{-N}. \tag{1}$$

We assume each device in the network can cache S files. We consider a library consisting of M files and assume that each file has equal size. We assume that users request the files from the library independently according to a request distribution modeled by the MZipf distribution [7], [20]:

$$P_r(f;\gamma,q) = \frac{(f+q)^{-\gamma}}{\sum_{m=1}^{M}(m+q)^{-\gamma}}, \tag{2}$$

where γ is the Zipf factor and q is the plateau factor of the distribution. The major difference between MZipf and Zipf distributions lies in that q creates a plateau regime in which files in such regime have similar probabilities. Such plateau regime is larger when q is larger, and the MZipf distribution degenerates to a Zipf distribution when $q = 0$. To simplify the notation, we will in this paper use $P_r(f)$ instead of $P_r(f;\gamma,q)$ as the short-handed expression.

We consider the decentralized random caching policy for all users [19], in which users cache files independently according to the same caching policy. Denoting $P_c(f)$ as the probability that a user caches file f, the caching policy is fully described by $P_c(1), P_c(2), ..., P_c(M)$, where $0 \leq P_c(f) \leq 1, \forall f$; thus users cache files according to the caching policy $\{P_c(f)\}_{f=1}^{M}$. To satisfy the cache space constraint, we have $\sum_{f=1}^{M} P_c(f) = S$. In this paper, we assume that S and γ are some constants.

We consider the asymptotic analysis in this paper, in which we assume that $N \to \infty$ and $M \to \infty$. We consider $\gamma > 1$, $q = o(M) \to \infty$, and $M = o(N)$. The reasons for these considerations are as follows. First, we consider $q \to \infty$ to ensure that q is impactful for scaling law analysis. This is because if q is a constant while $M \to \infty$, the MZipf distribution behaves like a Zipf distribution in terms of the scaling law performance, as indicated in [7]. Second, when q goes to infinity, it is sufficient to consider $q = \mathcal{O}(M)$. This is because the MZipf distribution would behave like a uniform distribution asymptotically as $q = \omega(M)$. Furthermore, when $\gamma > 1$, it is more interesting to consider the case that $q = o(M)$ because it gives a clear distinction over the case that $\gamma < 1$ in terms of the scaling law performance. As a matter of practice, we see from the measurment results in [7] that q is much smaller than M when $\gamma > 1$, which supports the consideration of $q = o(M)$. Finally, the assumption that $q = o(M)$ and $M = o(N)$ when $\gamma > 1$ can render the users of the network the sufficient ability to cache the most popular q files (orderwise); otherwise the outage probability could go to 1 [7]. We note that since S is a constant, the probability that a user can find the desired file from its own cache goes to zero as q and M go to infinity. This prevents the possible gain of trivial self-caching; we thus concentrate on the analysis of D2D collaborative caching gain. Moreover, similar to [4], we assume that different users making the requests on the same file would request different segments of the file, which avoids the gain from the naive multicasting.

We consider the physical model and define that the link rate between two users i and j follows the well-known physical model [14]:

$$R(i,j) = \begin{cases} R(\vartheta), \log_2\left(1+\frac{P_i l(i,j)}{N_0 + \sum_{k\neq j} P_k l(k,j)}\right) \geq \vartheta \\ 0, \log_2\left(1+\frac{P_i l(i,j)}{N_0 + \sum_{k\neq j} P_k l(k,j)}\right) < \vartheta \end{cases},$$

where $R(\vartheta) = \log_2(1+\vartheta)$ and ϑ is some constant according to the delivery mechanism; N_0 is the noise power spectral density; P_i is the power of user i; and $l(i,j) = \frac{\chi}{d_{ij}^{\alpha}}$ is the power attenuation between users i and j, where d_{ij} is the distance between users i and j, $\chi > 0$ is some constant, and

[1] The comprehensive analysis covering other cases can be found in our extended journal version [1].

[2] The scaling law order order notations used in this paper follow the conventional definitions in complexity analysis territory. They are carefully defined in the journal version [1].

$\alpha > 2$ is the pathloss factor. We note that this model will not be directly used in this paper. However, it is necessary when we want to leverage the results in [14] and [15] later.

We consider multi-hop D2D delivery for the network. Users can only obtain their desired files through either multi-hop D2D delivery or self-caching. In other words, users can only obtain files from caches of the users in the network. Note that since S is a constant but M goes to infinity, we can assume without loss of generality that the throughput per user of using self-caching is identical to that using D2D-caching; thus we do not distinguish between users retrieving the desired files from their own caches and from caches of other users. We define an outage as an occurrence that a user cannot obtain its desired file through either the multi-hop D2D delivery or self-caching. Suppose we are given a realization of number of users n in the network with a realization of the placement of the user locations P. In addition, we are given a realization of file requests F and a realization of file placement G of users according to the popularity distribution $P_r(\cdot)$ and caching policy $P_c(\cdot)$, respectively. We can define T_u as the throughput of user $u \in \mathcal{U}$ under a feasible multi-hop file delivery scheme. We then define the average throughput of user u with a given number of users n and location placement of users r as $\overline{T}_u(n,r) = \mathbb{E}[T_u \mid \mathsf{n} = n, \mathsf{P} = r]$, where the expectation is taken over the file requests F of users, the file placement of users G, and the file delivery scheme. Subsequently, we define

$$T_{\text{user}}(n,r) = \min_{u\in\mathcal{U}} \overline{T}_u(n,r). \tag{3}$$

Finally, the expected minimum average throughput of a user in the network is defined as

$$\overline{T}_{\text{user}} = \mathbb{E}_{\mathsf{n},\mathsf{P}}[T_{\text{user}}(n,r)], \tag{4}$$

where the expectation is taken over n and P.

When the number of users in the network is n, we define

$$N_o(n) = \sum_{u\in\mathcal{U}} \mathbf{1}\{\mathbb{E}[T_u \mid \mathsf{P},\mathsf{F},\mathsf{G}] = 0\} \tag{5}$$

as the number of users that in outage, where $\mathbf{1}\{\mathbb{E}[T_u \mid \mathsf{P},\mathsf{F},\mathsf{G}] = 0\}$ is the indicator function such that the value is 1 if $\mathbb{E}[T_u \mid \mathsf{P},\mathsf{F},\mathsf{G}] = 0$; otherwise the value is 0. Intuitively, $\mathbf{1}\{\mathbb{E}[T_u \mid \mathsf{P},\mathsf{F},\mathsf{G}] = 0\}$ is equal to zero when the file delivery scheme cannot deliver the desired file to user u. We note that the expectation of $\mathbb{E}[T_u \mid \mathsf{P},\mathsf{F},\mathsf{G}]$ is taken over the file delivery scheme and $\mathbf{1}\{\mathbb{E}[T_u \mid \mathsf{P},\mathsf{F},\mathsf{G}] = 0\}$ is a random variable with the distribution being the function of P, F, and G. The outage probability in the case of n users is then defined as

$$p_o(n) = \frac{1}{n}\mathbb{E}_{\mathsf{P},\mathsf{F},\mathsf{G}}[N_o(n)] = \frac{1}{n}\sum_{u\in\mathcal{U}} \mathbb{P}\left(\mathbb{E}[T_u \mid \mathsf{P},\mathsf{F},\mathsf{G}] = 0\right). \tag{6}$$

Consequently, the network outage probability is defined as

$$p_o = \mathbb{E}_{\mathsf{n}>0}[p_o(\mathsf{n} = n)] + \mathbb{P}_N(n = 0). \tag{7}$$

Note that since we consider $N \to \infty$, $\mathbb{P}_N(n = 0)$ is actually negligible for the asymptotic analysis. In the following, we will aim to analyze the throughput-outage performance in terms of $\overline{T}_{\text{user}}$ and p_o. We will be especially interested in the regime that the outage probability p_o is small, i.e., the regime that $p_o = \epsilon$, where ϵ is a negligibly small number or converges to zero.

III. Achievable Throughput-outage Performance

In this section, we derive the achievable throughput-outage performance of the network, in which we say $(T(P_o), P_o)$ is achievable if there exists a caching and multi-hop file delivery scheme such that $\overline{T}_{\text{user}} \geq T(P_o)$ and $p_o \leq P_o$. We will in the following first provide the achievable file delivery scheme, and then propose the achievable caching scheme. Accordingly, the achievable throughput-outage performance will be derived.

A. Achievable Caching and File Delivery Scheme

We consider the following achievable multi-hop file delivery scheme. We let $g_c(M)$ be a function of M which goes to infinity as $M \to \infty$. Then, a clustering approach is used to split the cell into equally-sized square clusters, in which each cluster has the side length $\sqrt{\frac{g_c(M)}{N}}$, and $g_c(M)$ is thus denoted as the cluster size. Different clusters could be activated simultaneously. The inter-cluster interference is avoided by a Time Division Multiple Access (TDMA) scheme with reuse factor K [21]. Such a reuse scheme evenly applies K colors to the clusters, and only the clusters with the same color can be activated on the same time-frequency resource for file delivery. We assume that a user in a cluster can only access files cached by users in the same cluster via either accessing its own cache or using (multi-hop) D2D communications following the multi-hop approach proposed in [15]. Specifically, denoting $\mathcal{V}_f$ as the set of users in a cluster that cache file f, we consider the following transmission policy: for each user u in the cluster, if the requested file f can be found in the caches of users in $\mathcal{V}_f$, then a user v_f, randomly selected from $\mathcal{V}_f$, is set as the source to deliver the requested (real) file f to user u; if the requested file cannot be found from the caches of any users in the cluster, user u would be matched with a randomly selected user v from users in the cluster, and then user v is set as the source for delivering a *virtual* file to user u. Note that it does not matter what file is delivered in this case, as the user is actually in outage.

After the establishment of the matching of the sources and destinations, to deliver (both real and virtual) files, the multi-hop approach proposed in [15] directly applies. Note that the delivery of virtual files cannot generate throughput for the network because users receiving virtual files are indeed in outage and the desired files are not actually received. However, we would still assume them to be included in the multi-hop D2D communications for the convenience of the mathematical analysis. Such scheme is suboptimal. Nevertheless, when the outage probability is either negligibly small or converging to zero, this scheme will be orderwise optimal because the performance degradation caused by delivering virtual files is negligible. Also note that, since the multi-hop approach in [15] can provide the per-user symmetric throughput for all users, this delivery scheme can thus provide the per-user symmetric throughput for users that are not in outage. Furthermore, since users cache files independently, the matching of the source-destination pairs here is equivalent to the uniformly random matching. Finally, we note that the assumption that a user may get a desired file from only its own cluster seems rather restrictive. However, the fact that this scheme can achieve (in the order sense) the outer bound shows that inclusion of inter-cluster communication cannot change the scaling law.

By adopting the aforementioned scheme, due to the symmetry of the network and the thinning property of PPP, the throughput-outage performance for each cluster is the same as the throughput-outage performance for the whole network. We will thus in the following focus on the analysis of a cluster to derive $\overline{T}_{\text{user}}$ and p_o. In addition, since a user is in outage only if this user cannot find the desired file from any users in the same cluster, the outage probability p_o is then equivalent to the probability that a user cannot find the desired file from users in the same cluster. Accordingly, when we denote the probability that a user can find the desired file in the cluster, i.e., the file hit-rate, as P_h, it is then clear that $P_h = 1 - p_o$.

To obtain the achievable caching scheme, we first provide Lemma 1 for the closed-form expression of p_o. Then, serving as the achievable caching scheme, the caching policy which minimizes p_o is proposed in Theorem 1.

Lemma 1: Considering the proposed file delivery scheme, cluster size $g_c(M)$, and the caching distribution $P_c(\cdot)$, the outage probability of the proposed achievable scheme is

$$p_o = \sum_{f=1}^{M} P_r(f) e^{-g_c(M)P_c(f)}. \tag{8}$$

Proof. See proof in Appendix A in [1]. □

Theorem 1: Let $N \to \infty$, $M \to \infty$, $q \to \infty$, and $g_c(M) \to \infty$. Denote m^* as the smallest index such that $P_c^*(m^*+1) = 0$. Let $C_2 = \frac{q\gamma}{Sg_c(M)}$; C_1 is the solution of the equation: $C_1 = 1 + C_2 \log\left(1 + \frac{C_1}{C_2}\right)$. The caching distribution $P_c^*(\cdot)$ that minimizes the outage probability p_o is as follows:

$$P_c^*(f) = \left[\log\left(\frac{z_f}{\nu}\right)\right]^+, f = 1, ..., M, \tag{9}$$

where $\nu = \exp\left(\frac{\sum_{f=1}^{m^*}\log z_f - S}{m^*}\right)$, $z_f = (P_r(f))^{\frac{1}{g_c(M)}}$, $[x]^+ = \max(x, 0)$, and

$$m^* = \Theta\left(\min\left(\frac{C_1 S g_c(M)}{\gamma}, M\right)\right). \tag{10}$$

Proof. See proof in Appendix B in [1]. □

Remark 1: Similar to the results in [7], Theorem 1 indicates that the number of files with non-zero probability to be cached by users is at least on the same order as the plateau factor q – if $q = \mathcal{O}(g_c(M))$, then $m^* = \Theta(g_c(M))$; if $q = \omega(g_c(M))$, then $m^* = \Theta(q)$. This is intuitive when we look at the shape of the MZipf distribution: the most popular q files (orderwise) have similar request probabilities, and we need to cache them to have the minimal outage probability.

Remark 2: Since Theorem 1 gives the optimal caching policy that minimizes the outage probability for a given cluster size, this implies that such a caching policy requires the smallest cluster size for a given outage probability. Consequently, with a given outage probability, the network throughput for the *clustering network* can be maximized by the caching policy in Theorem 1 because the number of activated clusters is maximized.

Based on the achievable caching and file delivery scheme in this subsection, we subsequently characterize the achievable throughput-outage performance considering $\gamma > 1$.

B. Throughput-Outage Performance

In this section, the achievable throughput-outage performance is characterized with $\gamma > 1$, $q = \omega(1)$, and $q = o(M)$.

Proposition 1: Let $M \to \infty$, $N \to \infty$, and $q \to \infty$. Suppose $\gamma > 1$ and $g_c(M) \to \infty$. Consider $g_c(M) = o(M)$ and $q = o(M)$. Let $C_2 = \frac{q\gamma}{Sg_c(M)}$. When adopting the caching policy in Theorem 1, the outage probability p_o is:

$$\begin{aligned} p_o &= 1 + (\gamma - 1) \\ &\cdot e^{-\gamma\left(\frac{1}{C_1}-1\right)} \left(\frac{C_1}{C_1+C_2}\right)^{\gamma} \left(\frac{C_2}{C_1+C_2}\right)^{\gamma\frac{C_2}{C_1}} \left(\frac{C_2}{C_1}\right)^{\gamma-1} \\ &- \left(\left(\frac{C_1}{C_2}\right)^{\gamma-1} - \left(\frac{C_1}{C_1+C_2}\right)^{\gamma-1}\right) \cdot \left(\frac{C_2}{C_1}\right)^{\gamma-1} \end{aligned} \tag{11}$$

Proof. See the proof of Proposition 3 in [1]. □

Corollary 1: Let $M \to \infty$, $N \to \infty$, and $q \to \infty$. Suppose $\gamma > 1$ and $g_c(M) \to \infty$. Consider $g_c(M) = o(M)$, $q = o(M)$, and $g_c(M) = \frac{\alpha_1 q}{S}$. When adopting the caching policy in Theorem 1 and considering $\alpha_1 = \Theta(1)$, we can obtain $p_o = \epsilon_2(\alpha_1)$, where $\epsilon_2(\alpha_1) > 0$ can be arbitrarily small. Furthermore, when $\alpha_1 = \omega(1)$, i.e., $q = o(g_c(M))$, we obtain $p_o = \Theta\left(\frac{1}{(\alpha_1)^{\gamma-1}}\right) = o(1)$.

Proof. This can be obtained by using Proposition 1 and $g_c(M) = \frac{\alpha_1 q}{S}$. See detailed proof in Appendix H in [1]. □

Theorem 2: Let $M \to \infty$, $N \to \infty$, and $q \to \infty$. Suppose $\gamma > 1$ and $g_c(M) \to \infty$. Consider $g_c(M) = o(M)$, $q = o(M)$, and $g_c(M) = \frac{\alpha_1 q}{S}$, where $\alpha_1 = \Omega(1)$. When adopting the caching policy in Theorem 1, the following throughput-outage performance is achievable:

$$T(P_o) = \Omega\left(\frac{(1-P_o)}{K}\sqrt{\frac{S}{\alpha_1 q}}\right), P_o = (11). \tag{12}$$

Proof. See proof in Appendix A. □

Corollary 2: Let $M \to \infty$, $N \to \infty$, and $q \to \infty$. Suppose $\gamma > 1$ and $g_c(M) \to \infty$. Consider $g_c(M) = o(M)$, $q = o(M)$, and $g_c(M) = \frac{\alpha_1 q}{S}$. When adopting the caching policy in Theorem 1 and considering $\alpha_1 = \Theta(1)$ to be large enough, the following throughput-outage performance is achievable:

$$T(P_o) = \Omega\left(\sqrt{\frac{S}{\alpha_1 q}}\right), P_o = \epsilon_2(\alpha_1), \tag{13}$$

where $\epsilon_2(\alpha_1) > 0$ can be arbitrarily small. Furthermore, when considering $\alpha_1 = \omega(1) \to \infty$, we obtain the following throughput-outage performance:

$$T(P_o) = \Omega\left(\sqrt{\frac{S}{\alpha_1 q}}\right), P_o = \Theta\left(\frac{1}{(\alpha_1)^{\gamma-1}}\right) = o(1). \tag{14}$$

Proof. Obtained directly from Theorem 1 and Corollary 1. □

Remark 3: Theorem 2 and Corollary 2 characterize the achievable throughput-outage performance.[3] Especially, Corollary 2 indicates that we can achieve the throughput $\Omega\left(\sqrt{\frac{S}{q}}\right)$ with a negligibly small outage probability. It also shows that when the outage probability converges to zero with the rate $(\alpha_1)^{\gamma-1}$, the achievable throughput is $\Omega\left(\sqrt{\frac{S}{\alpha_1 q}}\right)$. Besides, by using Corollary 2, we understand that when the popularity distribution has a light tail, i.e., $\gamma > 1$ and $q = o(M)$, the performance is restricted by the order of q, instead of M

IV. Outer Bound of the Throughput-outage Performance

In this section, we derive the outer bound of the throughput-outage performance. In the following, we say a point $(T(P_o), P_o)$ is dominant (thus serving as an outer bound point) if, for any caching and delivery scheme, either $T(P_o) \geq \overline{T}_{\text{user}}$ or $P_o \leq p_o$ is satisfied. Note that although there are different dominant points, we will specifically characterize the dominant points where P_o is either negligibly small or converging to zero.

Theorem 3: Let $M \to \infty$, $N \to \infty$, and $q \to \infty$. Suppose $\gamma > 1$ and $q = o(M)$. When considering $\alpha_1' = \Theta(1)$, the throughput-outage performance of the network is dominated by:

$$T(P_o) = \Theta\left(\sqrt{\frac{S}{\alpha_1' q}}\right), P_o = \epsilon_2'(\alpha_1'), \tag{15}$$

where $\epsilon_2'(\alpha_1') > 0$ can be arbitrarily small. Furthermore, when considering $\alpha_1' = \mathcal{O}\left(q^{\frac{1}{\gamma-1}}\right) \to \infty$ but $\alpha_1' q = o(M)$, the throughput-outage performance of the network is dominated by:

$$T(P_o) = \Theta\left(\sqrt{\frac{S}{\alpha_1' q}}\right), P_o = \Theta\left(\frac{1}{(\alpha_1')^{\gamma-1}}\right) = o(1), \tag{16}$$

Proof. See proof in Appendix B. □

Remark 4: By comparing between Corollary 2 and Theorem 3, we see that there is no gap between the achievable throughput-outage performance and the outer bound when $\gamma > 1$. This shows that the provided achievable scheme is orderwise optimal when the outage probability is either negligibly small or converging to zero.

V. Conclusions

In this work, we conducted a scaling law analysis for the throughput-outage performance of the cache-aided multi-hop D2D networks under the PPP and MZipf distribution for user distribution and popularity distribution, respectively. By demonstrating that there is no gap between the proposed achievable performance and outer bound, optimality is obtained. Specifically, when $q = \omega(1)$ and $\gamma > 1$, we show that the optimal throughput per user scaling is $\Theta\left(\sqrt{\frac{S}{q}}\right)$ when the outage probability is negligible.

Appendix A
Proof of Theorem 2

We here only outline the proof due to page limitation. The complete proof can be found in Appendix I in [1]. We consider $g_c(M) = \frac{\alpha_1 q}{S}$, where $\alpha_1 = \Omega(1)$. Consequently by Proposition 1, we can obtain the outage probability p_o. To compute the throughput of a cluster, we leverage the results in [15]. Recall that when using the achievable scheme in Sec. III.A, the multi-hop approach proposed in [15] is used for delivering both real and virtual files. We denote the throughput generated via transmitting real file as effective throughput; the throughput generated via transmitting virtual file as virtual throughput; and the sum of the real and virtual throughput as mixing throughput. Since only the effective throughput can be taken into account for $\overline{T}_{\text{user}}$, we want to compute its value.

Our approach is to first compute the mixing throughput, and then exclude the virtual throughput from it. From the definition, we know:

$$\overline{T}_{\text{user}} = \mathbb{E}_{\mathsf{n},\mathsf{P}}\left[\min_{u \in \mathcal{U}} \mathbb{E}\left[C_u \cdot 1_{\mathsf{H}_u} \mid \mathsf{n}, \mathsf{P}\right]\right], \tag{17}$$

where C_u is the mixing throughput of user u; 1_{H_u} is the indicating function of the event H_u defined as $\mathsf{H}_u = \{$the user u can find the desired file in the cluster$\}$. Thus, $1_{\mathsf{H}_u} = 1$ if user u can find the desired file in the cluster; otherwise $1_{\mathsf{H}_u} = 0$. Then according to the result in [15] and due to the frequency reuse scheme among different clusters, we have the following Theorem:

Theorem A.1 [15]: When using the proposed achievable scheme, with high probability (w.h.p.), users in a cluster with side length $\sqrt{\frac{g_c(M)}{N}}$ can achieve $C_u = \Omega\left(\frac{1}{K}\sqrt{\frac{1}{g_c(M)}}\right)$ of the mixing throughput simultaneously.

From Theorem A.1, we know that, w.h.p, there exists a $\epsilon = \Theta(1) > 0$ such that $C_u \geq \frac{\epsilon}{K}\sqrt{\frac{1}{g_c(M)}}$ for all users. We note that both Theorem A.1 and event 1_{H_u} have the symmetry property for all users. It is then sufficient that we consider an arbitrary user in the network. We then let $C_{\text{user}} = \frac{\epsilon}{K}\sqrt{\frac{1}{g_c(M)}}$ and recall that $P_h = 1 - p_o$ is the file hit-rate. By using (17), we obtain:

$$\overline{T}_{\text{user}} = \Omega\left(\frac{1-p_o}{K}\sqrt{\frac{S}{\alpha_1 q}}\right). \tag{18}$$

This complete the proof.

[3] Some simulations that show the results of Theorem 2 and Corollary 2 numerically can be found in Sec. III.D in [1].

Appendix B
Proof of Theorem 3

We again only outline the proof. The complete proof can be found in Appendix K in [1]. We first consider the network having $\mathsf{n} = n = \omega(q)$ uniformly distributed users and derive the outer bound of $T_{\text{user}}(n)$ and $p_o(n)$, where $T_{\text{user}}(n) = \mathbb{E}_{\mathsf{P}|n}\left[T_{\text{user}}(n,r)\right]$. Then, we obtain $\overline{T}_{\text{user}}$ and p_o via accommodating different n with high probability.

Suppose the network has $\mathsf{n} = n = \omega(q)$ users, where the location placement P of users follows the BPP. We denote $\lambda(n,\mathsf{P}) = \frac{\sum_{u\in\mathcal{U}}\overline{T}_u}{n}$ as the average throughput per user in the network and $\overline{L}(n,\mathsf{P})$ as the average distance between the source and destination in the network. Using Theorem 4.2 in [14], which describes the upper bound of the transport capacity of the network for any arbitrary placement of users and choice of transmission powers, we obtain

$$\lambda(n,\mathsf{P}) \leq \Theta\left(\frac{1}{\overline{L}(n,\mathsf{P})\sqrt{n}}\right). \tag{19}$$

To compute the upper bound of $\lambda(n,\mathsf{P})$, we need to find $\overline{L}(n,\mathsf{P})$. To do this, we first provide Lemmas 2 and 3 that will be used later (see Lemmas 7 and 8 in [1].):

Lemma 2: When $n = \omega(q)$ users are uniformly distributed within a network with unit size, the minimum size of an area to have $\Theta\left(\frac{q}{S}\right)$ users with high probability is $\Theta\left(\frac{q}{Sn}\right)$.

Lemma 3: Suppose $\gamma > 1$ and $n = \omega(q)$. Considering $q = o(M)$, we have the following results: (i) when a user searches through $n_{\text{s}} = o\left(\frac{q}{S}\right)$ different users in the network, we obtain $p_{\text{miss}}(n) \geq 1 - o(1)$; (ii) when a user searches through $n_{\text{s}} = \frac{\alpha_1' q}{S}$ different users, where $\alpha_1' = \Theta(1) > 0$, we obtain $p_{\text{miss}}(n) \geq \epsilon_{\text{miss}}(\alpha_1')$, where $\epsilon_{\text{miss}}(\alpha_1') = \Theta(1) > 0$ can be arbitrarily small; (iii) when a user searches through $n_{\text{s}} = \frac{\alpha_1' q}{S} < \frac{M}{S}$ different users, where $\alpha_1' = \mathcal{O}\left(q^{\frac{1}{\gamma-1}}\right) \to \infty$, we obtain: $p_{\text{miss}}(n) = \Omega\left(\frac{1}{(\alpha_1')^{\gamma-1}}\right)$.

From Lemmas 2 and 3, we conclude that to have a non-vanishing probability for a user to obtain the desired file, w.h.p., the the distance between the source and destination is at least $\Theta\left(\sqrt{\frac{q}{Sn}}\right)$. Furthermore, if we consider $\overline{L}(n,\mathsf{P}) = \Theta\left(\sqrt{\frac{\alpha_1' q}{Sn}}\right)$, we know that (w.h.p.) the distance between a source-destination pair is $\mathcal{O}\left(\sqrt{\frac{\alpha_1' q}{Sn}}\right)$; otherwise we should have $\overline{L}(n,\mathsf{P}) = \omega\left(\sqrt{\frac{\alpha_1' q}{Sn}}\right)$. As a result, w.h.p., the number of users searched by a user is $n_s = \mathcal{O}\left(\frac{\alpha_1' q}{S}\right)$. Note that above arguments are valid for any $n = \omega(q)$ and P. Consequently, by combining this with Lemma 3 and using (19) and the fact that $T_{\text{user}}(n,\mathsf{P}) \leq \lambda(n,\mathsf{P})$, we conclude that for all $n = \omega(q)$, we must have

$$T_{\text{user}}(n) = \mathcal{O}\left(\sqrt{\frac{S}{\alpha_1' q}}\right) \tag{20}$$

with $p_o(n) \geq \epsilon_{\text{miss}}(\alpha_1')$ when $\alpha_1' = \Theta(1)$; and $p_o(n) = \Omega\left(\frac{1}{(\alpha_1')^{\gamma-1}}\right)$ when $\alpha_1' = \omega(1)$ and $\alpha_1' = o\left(q^{\frac{1}{\gamma-1}}\right)$. Finally, recall that we consider $q = o(N)$ when $\gamma > 1$. Consequently, according to (20) and that $n = \omega(q)$ w.h.p. (see Lemma 6 in [1]), we obtain the theorem.

References

[1] M.-C. Lee, M. Ji, and A. F. Molisch, "Optimal throughput–outage analysis of cache-aided wireless multi-hop D2D networks – Derivations of scaling laws," May 2020, online available at https://arxiv.org/abs/2005.05149.

[2] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commmun. Mag.*, vol. 51, no. 4, pp. 142–149, April 2013.

[3] L. Li, G. Zhao, and R. S. Blum, "A survey of caching techniques in cellular networks: Research issues and challenges in content placement and delivery strategies," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 1710–1732, 2018.

[4] M. Ji, G. Gaire, and A. F. Molisch, "The throughput-outage tradeoff of wireless one-hop caching networks," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6833–6859, December 2015.

[5] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *IEEE J. Sel. Area Commun.*, vol. 34, no. 1, pp. 176–189, January 2016.

[6] S.-W. Jeon, S.-N. Hong, M. Ji, G. Caire, and A. F. Molisch, "Wireless multihop device-to-device caching networks," *IEEE Trans. Inf. Theory*, vol. 63, no. 3, pp. 1662–1676, 2017.

[7] M.-C. Lee, M. Ji, A. F. Molisch, and N. Sastry, "Throughput-outage analysis and evaluation of cache-aided d2d networks with measured popularity distributions," *IEEE Trans. on Wireless Commun.*, vol. 18, no. 11, pp. 5316–5332, November 2019.

[8] M. Mehrabi, D. You, V. Latzko, H. Salah, M. Reisslein, and F. H. P. Fitzek, "Device-enhanced mec: Multi-access edge computing (mec) aided by end device computation and caching: A survey," *IEEE Access*, vol. 7, pp. 166 079–166 108, 2019.

[9] S. Gitzenis, G. S. Paschos, and L. Tassiulas, "Asymptotic laws for joint content replication and delivery in wireless networks," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2760–2776, 2012.

[10] M. Ji, R.-R. Chen, G. Caire, and A. F. Molisch, "Fundamental limits of distributed caching in multihop d2d wireless networks," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 2950–2954.

[11] M.-C. Lee, M. Ji, A. F. Molisch, and N. Sastry, "Performance of caching-based d2d video distribution with measured popularity distributions," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[12] L. Qiu and G. Cao, "Popularity-aware caching increases the capacity of wireless networks," *IEEE Trans. Mobile Comput.*, vol. 19, no. 1, pp. 173–187, 2019.

[13] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. on inf. theory*, vol. 46, no. 2, pp. 388–404, 2000.

[14] A. Agarwal and P. R. Kumar, "Capacity bounds for ad hoc and hybrid wireless networks," *ACM SIGCOMM Computer Communication Review*, vol. 34, no. 3, pp. 71–81, 2004.

[15] M. Franceschetti, O. Dousse, D. N. Tse, and P. Thiran, "Closing the gap in the capacity of wireless networks via percolation theory," *IEEE Trans. Inf. Theory*, vol. 53, no. 3, pp. 1009–1018, 2007.

[16] E. Cohen and S. Shenker, "Replication strategies in unstructured peer-to-peer networks," *ACM SIGCOMM Computer Communication Review*, vol. 32, no. 4, pp. 177–190, 2002.

[17] L. Yin and G. Cao, "Supporting cooperative caching in ad hoc networks," *IEEE Trans. mobile computing*, vol. 5, no. 1, pp. 77–89, 2005.

[18] N. Golrezaei, A. D. Dimakis, and A. F. Molisch, "Scaling behavior for device-to-device communications with distributed caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4286–4298, July 2014.

[19] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," June 20015.

[20] M. Hefeeda and O. Saleh, "Traffic modeling and proportional partial caching for peer-to-peer systems," *IEEE/ACM Trans. Netw.*, vol. 16, no. 6, pp. 1447–1460, December 2008.

[21] A. F. Molisch, *Wireless Communications*, 2nd ed. IEEE Press-Wiley, 2012.