Throughput–Outage Analysis and Evaluation of Cache-Aided D2D Networks With Measured Popularity Distributions

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

***Abstract*— Caching of video files on user devices, combined with file exchange through device-to-device (D2D) communications is a promising method for increasing the throughput of wireless networks. Previous theoretical investigations showed that throughput can be increased by orders of magnitude, but assumed a Zipf distribution for modeling the popularity distribution, which was based on observations in *wired* networks. Thus the question whether cache-aided D2D video distribution can provide in practice the benefits promised by existing theoretical literature remains open. To answer this question, we provide new results specifically for popularity distributions of video requests of mobile users. Based on an extensive real-world dataset, we adopt a generalized distribution, known as Mandelbrot-Zipf (MZipf) distribution. We first show that this popularity distribution can fit the practical data well. Using this distribution, we analyze the throughput–outage tradeoff of the cache-aided D2D network and show that the scaling law is identical to the case of Zipf popularity distribution when the MZipf distribution is sufficiently skewed, implying that the benefits previously promised in the literature could indeed be realized in practice. To support the theory, practical evaluations using numerical experiments are provided, and show that the cache-aided D2D can outperform the conventional unicasting from base stations.**

***Index Terms*— Wireless caching network, device-to-device (D2D) communications, throughput-outage tradeoff, scaling laws.**

I. Introduction

WIRELESS data traffic is anticipated to increase at a rate of $50-100\%$ per year for the foreseeable future. The main driver for this development is video traffic, which accounts for about $2/3$ of all wireless data [2], and has emerged as one of the most important applications of 4G cellular services, as well as motivating the "enhanced mobile broadband" thrust of 5G. It is thus of paramount importance to find cost-effective ways for increasing the throughput of cellular networks for video distribution.

Traditional methods for throughput enhancement have treated video traffic just like any other traffic, meaning that each file transmission or on-demand streaming transmission is treated as a unicast. Consequently, it relies on the general throughput enhancement methods of cellular networks such as network densification, HetNets [3], massive MIMO, and use of additional spectrum (in particular mm-wave bands [4]). However, these approaches tend to be either very expensive, and/or not scalable.

On-demand video has, however, unique properties compared to other data, namely [5] (i) high concentration of the popularity distribution (i.e., a small number of videos accounts for the majority of the video traffic), and (ii) *asynchronous content reuse*, i.e., those files are watched by different people at different times.[1] This offers the possibility of employing caching as a part of the video distribution process, thereby facilitating improvement of spectral efficiency through use of dedicated memory [6]; the theoretical scaling laws show that if we double the memory size of each user device, then the per user throughput can also be doubled [6]. Such an approach is appealing because bandwidth is limited and expensive, while memory is relatively cheap and a rapidly growing hardware resource. Caching approaches include selfish on-device caching [7], femtocaching [5], coded caching [8]–[10], and caching combined with device-to-device (D2D) based file exchange [5]–[7]. On-device caching naturally uses the storage of users' own devices to cache files possibly watched by users in the future. Since the on-device caching requires the prediction of user behavior to gain large benefits, caching designs incorporating recommendation systems have been investigated [11]. Femtocaching or caching in base station (BS) exploits the storage in BSs to relax the requirement of the backhaul [5], [12], [13]. Coded caching, combining coding with multicasting, leverages the redundancy of cache memories and the broadcasting nature of the wire-

Manuscript received January 29, 2019; revised May 28, 2019; accepted July 30, 2019. Date of publication August 21, 2019; date of current version November 11, 2019. This work was supported in part by the National Science Foundation (NSF) under Grant CNS-1816699, Grant CCF-1423140, Grant CCF-1817154, and Grant SpecEES-1824558, and in part by the U.K. EPSRC under Grant EP/K024914/1 and Grant EP/P022723/1. This article was presented in part at the 2019 IEEE Global Communications Conference [1]. The associate editor coordinating the review of this article and approving it for publication was S. Dey. *(Corresponding author: Ming-Chun Lee.)*

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

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

N. Sastry is with the Department of Informatics, King's College London, London WC2B 4BG, U.K. (e-mail: nishanth.sastry@kcl.ac.uk).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2019.2935452

[1] The latter property distinguishes video streaming services such as Netflix, Amazon Prime, Hulu, and Youtube, from the traditional broadcast TV, which achieved high spectral efficiency by forcing viewers to watch particular videos at prescribed times.

less medium, and effectively facilitates throughput increase by using caching [8]–[10]. Cache-aided D2D exploits recent high-throughput D2D communications [14] and storage in user devices to gain benefits. This method has been shown to provide not only appealing throughput scaling laws (network throughput increasing linearly with the number of devices) [6], but also robustness in realistic propagation conditions [15]. It will therefore be the focus of this paper.

Cache-aided D2D considered in this paper has the following principle: each user device caches, at random, a subset of the video files (the particular caching distribution is a function of the video popularity distribution and other system parameters). When a user requests a file, it might be either already in this user's cache, or is obtained from a nearby device through short-distance D2D communications. This approach was first suggested by one of the authors in [16], and since then has been widely explored in the literature. Among the related papers, different goals, including optimizing outage probability [6], [15], [17], [18], throughput [6], [15], [19]–[21], [26], energy efficiency (EE) [21]–[23], and delay [24], [25], are pursued using various approaches, and different theoretical and practical aspects have been considered. For example, the information-theoretic throughput-outage scaling laws were explored in [6], [15], [26]. The tradeoff between throughput and EE was investigated in [21]. In [22], battery life was taken into account for optimal EE design. In [27], a joint scheduling, power control, and caching policy design was proposed. In [28] mobility was leveraged to maximize offloading data to D2D networks. Stochastic geometry was used to analyze cache-aided D2D in [29], [30]. In [31], cache-aided D2D with millimeter wave communications was investigated, considering detailed physical channel effects. Using the devices as relay nodes, caching-aided D2D with energy harvesting was proposed in [32]. To deal with time-varying popularity distributions, dynamic caching content replacement was discussed in [33]. Since the field of video caching has been of great interest in the past several years and several hundred related papers have been published, the above literature review cites only a sample of papers and topics.

Most existing papers assume the popularity distribution as the Zipf distribution (essentially a power law distribution). However, this assumption was based on observations in *wired* networks [34] with Youtube videos and with little empirical support for *wireless* network. A recent investigation [35] into wireless popularity distributions of general content showed little content reuse. However, as the authors of the paper pointed out, since video connections were run via a secure https connection so that the content of the videos could not be determined, this investigation could not uniquely identify video content reuse, and therefore the popularity distribution of video content reuse in wireless networks is still not clear. Consequently, the question remains open whether cache-aided D2D video distribution can achieve the significant gains promised in the literature. This paper aims to answer this question.

In particular, we use the *measured* video popularity distribution of the BBC (British Broadcasting Corporation) iPlayer, a popular video streaming application in the UK. Through appropriate postprocessing, we are able to extract the popularity distribution for the videos watched via cellular connections (these might be different from the files watched through wired connections). We find that this distribution is not well described by a Zipf distribution, but rather a Mandelbrot-Zipf (MZipf) distribution [36], which is somewhat less skewed. Such distribution, in contrast to the simple Zipf distribution, is characterized by two parameters: the Zipf factor γ and plateau factor q, and it degenerates to the Zipf distribution when $q = 0$. Thus the MZipf distribution generalizes the Zipf distribution. Considering this more general model, we investigate the benefits of the cache-aided D2D video distribution.

To understand the performance of the cache-aided D2D video distribution, we conduct a thorough throughput–outage tradeoff analysis following the framework in [6] but using a different analytical approach and aim to see the scaling law of the throughput-outage tradeoff when the more general MZipf distribution is considered. We derive the analytical formulation of the caching policy maximizing the probability of users to access the desired files via D2D communications. Based on this policy, we obtain the achievable throughput–outage tradeoff. Since the MZipf distribution has the additional factor q, the derived caching policy and achievable throughput–outage tradeoff can characterize the influence of q. This distinguishes our results from [6]. However, this does not imply the resulting scaling behavior is worse than the case with the Zipf distribution. In contrast, the results indicate that in a particular range of q, the same scaling law as considering the Zipf distribution can be obtained again when the MZipf distribution is considered; implying that the benefits promised by existing literature should be retained in practice. We emphasize that, after investigating the real-world data, we find that this range of q is valid in practice.

To support the theoretical analysis and verify the benefits of considering the MZipf model from the perspective of the network, numerical experiments are conducted in D2D networks considering MZipf distributions parameterized based on the real-world data and the realistic setup adopted from [15]. Results show that the cache-aided D2D scheme can provide orders of magnitude improvement of throughput for a negligible outage probability compared to conventional unicasting, and that the MZipf model can provide more accurate performance evaluations when compared with the Zipf model. Our main contributions are summarized below:

- Based on an extensive BBC iPlayer dataset, we extract the popularity distribution for the videos watched by mobile users. Such distribution is then modeled and parameterized by the MZipf distribution, which is a generalized version of the widely used Zipf distribution. To our best knowledge, this is the only work that reports the measured popularity distributions for mobile users and provides modeling results.
- To investigate the throughput–outage tradeoff of the cache-aided D2D networks considering a MZipf distribution, we generalize the theoretical treatment of [6] with a different but simpler proof technique. Such generalization

is non-trivial and several new techniques are used such that the influence of q can be explicitly expressed.

- We show that the scaling law of cache-aided D2D achieved in [6] is achievable in the case of the practical MZipf distribution; we also characterize the influences of the critical parameters γ and q of the MZipf distribution on the throughput–outage tradeoff. The question of whether the gains theoretically predicted in [6] hold with realistic (i.e., measured) popularity distributions has often been raised. The current paper answers that important question.
- To support the theoretical study, we conduct numerical experiments with practical details and show that the cache-aided D2D can significantly outperform the conventional unicasting. To verify the benefits of considering the MZipf model from the network perspective, we also show that simulations using the MZipf model can provide more accurate performance evaluations compared with the conventional Zipf model.

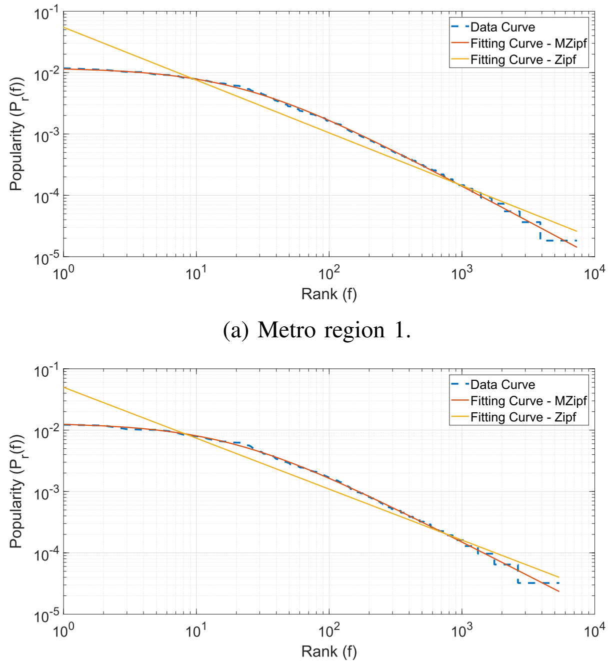
The remainder of the paper is organized as follows. In Sec. II, the dataset of video requests for mobile users is described and the corresponding modeling and parameterization are presented. In Sec. III, the theoretical analysis of throughput–outage tradeoff is provided and insights are discussed. We offer numerical experiments in Sec. IV to support the theory. Finally, we conclude the paper in Sec. V. Proofs of theorems and corollaries are relegated to appendices.

Scaling law order notation: given two functions f and g, we say that: (1) $f(n) = \mathcal{O}(g(n))$ if there exists a constant c and integer N such that $f(n) \leq cg(n)$ for $n > N$. (2) $f(n) = o(g(n))$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$. (3) $f(n) = \Omega(g(n))$ if $g(n) = \mathcal{O}(f(n))$. (4) $f(n) = \omega(g(n))$ if $g(n) = o(f(n))$. (5) $f(n) = \Theta(g(n))$ if $f(n) = \mathcal{O}(g(n))$ and $g(n) = \mathcal{O}(f(n))$.

II. Measured Data and Popularity Distribution Modeling

This work uses an extensive set of real-world data, namely the dataset of the BBC iPlayer [11], [37], [38], to obtain realistic video demand distributions. The BBC iPlayer is a video streaming service from BBC that provides video content for a number of BBC channels without charge. Content on the BBC iPlayer is available for up to 30 days depending on the policies. We consider two datasets covering June and July, 2014, which include 192,120,311 and 190,500,463 recorded access sessions, respectively. In each record, access information of the video content contains two important columns: *user id* and *content id*. *User id* is based on the long-term cookies that uniquely (in an anonymized way) identify users. *Content id* is the specific identity that uniquely identifies each video content separately. Although there are certain exceptions, *user id* and *content id* can generally help identify the user and the video content of each access. More detailed descriptions of the BBC iPlayer dataset can be found in [11], [37], [38].

To facilitate the investigation, preprocessing is conducted on the dataset. We notice that a user could access the same file multiple times, possibly due to temporary disconnnections from Internet and/or due to temporary pauses by users while viewing. Since a user is unlikely to access the same video after finishing watching the video within the period of a month [38] and a fetched file can be temporarily cached in the user device for later viewing, we consider multiple accesses made by the same user to the same file as a single unique access.



(a) Metro region 1.

(b) Metro region 2.

Fig. 1. Measured ordered popularity distribution of video files of the BBC iPlayer requested via the cellular operator in July of 2014. $\gamma = 0.86$ and $\gamma = 0.83$ for Zipf distributions in Metro regions 1 and 2, respectively. γ and q for the MZipf distributions are shown in Table I.

We then separate the data requested by cellular users from those requested via cabled connections or personal WiFi by observing the services of the Internet service providers (ISPs), resulting in $640,631$ different unique accesses (requests) among $267,424$ different users in June; $689,461$ different unique accesses among $327,721$ different users in July. We also separate the data between different regions by observing the Internet gateway through which the requests are routed. Specifically, we consider one of the four major cellular operators in the UK. We can geographically localize two of its gateways to two of the major metropolitan areas in the UK, and we believe are mainly intended to serve users from these metropolitan areas. Therefore, we use these two metropolitan region gateways to validate our results.

Based on these data, we plot the global popularity distribution and find that the Zipf distribution is not a good fit. Instead, a MZipf distribution [36] provides a good approximation as demonstrated in Fig. 1 (since data for other months and regions show similar results, we thus omit their depictions for brevity):

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

where $P_r(f)$ is the probability that users want to access file f, i.e., the request probability of users for file f, M is the number of files in the library, γ is the Zipf factor, and q

TABLE I

PARAMETERIZATION OF POPULARITY DISTRIBUTION USING THE MZIPF MODEL

Region	γ (June)	q (June)	M (June)
Whole UK	1.36	50	16823
Metro 1	1.23	33	6449
Metro 2	1.18	28	4859
Region	γ (July)	q (July)	M (July)
Whole UK	1.28	34	19379
Metro 1	1.16	22	7345
Metro 2	1.11	18	5405

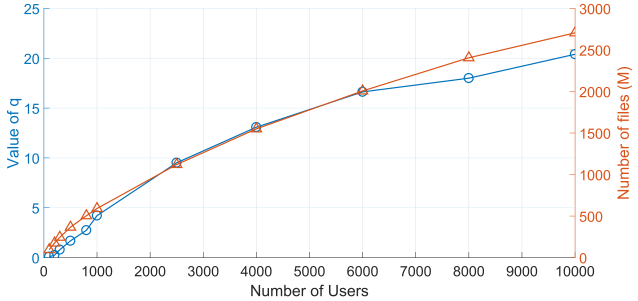


Fig. 2. Relation between q, M, and N using data in Metro region 1 of June, 2014.

is the plateau factor. We note that the MZipf distribution degenerates to a Zipf distribution when $q = 0$. We also note that a possible reason, as described in [39], for observing the MZipf distribution instead of the Zipf distribution is that a user only fetches the same file once.

A fitting that minimizes the Kullback-Leibler (KL) divergence between the data and model provides values of the parameters γ, q, and M as shown in Table I.[2] The results imply that up to a breakpoint, i.e., q, of approximately 20-50 files, the popularity distribution is relatively flat, and decays faster from there. Also importantly, we find that $\gamma > 1$ for all results, which has important implications for the throughput–outage scaling law due to caching. Moreover, we find that the values of q are much smaller (order-wise) than the values of M, which also has an important implication that the aggregate memory of the D2D network can easily surpass the number of files requested with similar probabilities and thus should be cached in the D2D network intuitively. Mathematically, in Sec. III, we will see that when the aggregate memory is smaller than the value of q (order-wise), the outage goes to 1 asymptotically as the library size M and q go to infinity, indicating poor performance. Finally, based on the data in Metro region 1 during June, 2014, Fig. 2 shows the relationship between the values of q, M, and the number of users N; we let N range here from 10 to $10,000$, covering the range of realistic values for the number of users in a cell. It can be observed that q is much smaller than M when N is realistic. Although not shown here for brevity, γ is (on average) between 0.2 and 1.1 for the range of N considered in Fig. 2, and γ generally increases when N increases.

Although our dataset cannot directly represent the global popularity distribution of a small area, e.g., a cell, due to the limitation discussed, our results are the best indication currently available, because to the best of our knowledge there are no publicly available data for *video reuse* of *mobile* data on a *per-cell* basis. We will thus make in the following the assumption that the popularity distribution at each location follows the global (over a particular region) popularity distribution and use the parameters of Metro regions 1 and 2 henceforth.

III. ACHIEVABLE THROUGHPUT–OUTAGE TRADEOFF

From the measured data, we understand that the MZipf distribution is more suitable for mobile data traffic. In this section, we thus generalize the theoretical treatment in [6] by considering the MZipf distribution and provide the achievable throughput–outage tradeoff analysis.

A. Network Setup

In this section, we describe the network model and define the throughput–outage tradeoff. Denote the number of users in the network as N. Our goal is to provide the asymptotic analysis when $N \to \infty$, $M \to \infty$, and $q \to \infty$.[3] We assume a network where user devices can communicate with each other through direct links. We consider the transmission policy using *clustering*, in which the devices are grouped geographically into clusters such that any device within one cluster can communicate with any other devices in the same cluster with a constant rate C bits/second, but not with devices in a different cluster. The network is split into equal-sized clusters. We adopt a grid network in which the users are placed on a regular grid [6]. As a result, $g_c(M) \leq N \in \mathbb{N}$, which is a function of M and denoted as the cluster size, is the number of users in a cluster and is a parameter to be chosen in order to analyze the throughput–outage tradeoff. Moreover, we say a potential link exists in the cluster if a user can find its desired file in the cluster through D2D communications and say that a cluster is *good* if it contains at least one potential link.

We assume only a single user in a cluster can use its potential link to obtain the requested file at a time, thus avoiding the interference between users in the same cluster. Besides, potential links of the same cluster are scheduled with equal probability (or, equivalently, in round robin). Therefore all users have the same average throughput. To avoid interference between clusters, we use a spatial reuse scheme with Time Division Multiple Access (TDMA). Denoting K as the reuse factor, 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 D2D communications. Note that the adopted reuse scheme is

[2] The KL divergence of a parameter set $\mathbf{x}$ is defined as $D_{\text{KL}}(\mathbf{x}) = \sum_m p_m^{\text{data}} \log \frac{p_m^{\text{data}}}{p_m^{\text{model}}(\mathbf{x})}$.

[3] We generally consider $q = \mathcal{O}(M)$ because, by definition, the MZipf distribution would converge to simple uniform distribution when $q = \omega(M)$. Besides, as a matter of practice, we can see from Table I that q is much smaller than M. Note that we view the case that $q = \Theta(1)$ is a constant simply as a degenerate case of our results. Also, based on the experimental results, γ changes within a (small) finite range, i.e., does not go to infinity, as M increases. We therefore approximate γ as a fixed constant for the sake of analysis.

analogous to the spatial reuse scheme in conventional cellular networks [40]. Besides, we use TDMA only as convenient example. Any scheme that allocates orthogonal resources to clusters with different colors is aligned with our model.

Although the assumptions above are made for the subsequent theoretical analysis, they are actually practical. Specifically, the adjustable size of the cluster can be implemented by adapting the transmit power - in other words, the transmit power is chosen such that communication between opposite corners of a cluster is possible. The link rate for the D2D communication is fixed when no adaptive modulation and coding, and of course this rate has to be smaller than the capacity for the longest-distance communication envisioned in this system. The signal-to-noise ratio (SNR) is determined by the pathloss; small-scale fading can be neglected since in highly frequency-selective channels, the effects of this fading can be eliminated by exploiting frequency diversity [40], and shadowing effects can be accommodated by adding a constant throughput loss from the system's point of view since the caching policy as well as the file delivery are implemented on a long-term scale. It must be emphasized that the above network is not optimum for D2D communications. Suitable power control, adaptive modulation and coding, etc., could all increase the spectral efficiency. However, our model provides both a useful lower bound on the performance as well as analytical tractability, which is important for comparability between different schemes. The information theoretical optimal throughput-outage tradeoff analysis is beyond the scope of this paper.

We denote S as the cache memory in a user device, i.e., a user can cache up to S files. Note that we do not consider S to grow to infinity as $N \to \infty$, $M \to \infty$, and $q \to \infty$, i.e., we consider $S = \Theta(1)$ as a fixed network parameter, in this paper. The aggregate memory in a cluster is then $Sg_c(M)$. An independent random caching policy is adopted for users to cache files. Denote $P_c(f)$ as the probability of caching file f, where $0 \leq P_c(f) \leq 1$ and $\sum_{f=1}^{M} P_c(f) = 1$. Using such caching policy, each user caches each file independently at random according to $P_c(f)$. We note that when using this policy, a user might cache the same file multiple times, and this policy is used for the sake of analysis.

Given the popularity distribution $P_r(\cdot)$, caching policy $P_c(\cdot)$, and transmission policy, we define the average throughput of a user u as $\overline{T}_u = \mathbb{E}[T_u]$, where T_u is a throughput realization of user u, and the expectation is taken over the realizations of the cached files and requests and scheduling results. The minimum average throughput is $\overline{T}_{\min} = \min\limits_u \overline{T}_u = \overline{T}_u$ due to the symmetry of the network (e.g., round robin scheduling). We define the number of users in outage N_o as the number of users that cannot find their requested files. Thus the average outage is:

$$p_o = \frac{1}{N}\mathbb{E}[N_o] = \frac{1}{N}\sum_u \mathbb{P}(\mathbb{E}[T_u \mid \mathsf{F}, \mathsf{G}] = 0) = 1 - P_u^c, \qquad (2)$$

where $\mathbb{E}[T_u \mid \mathsf{F}, \mathsf{G}]$ is the average throughput of user u conditioned on a set of requests F and a set of cached files G of users in the network; P_u^c is the probability that a user u can find its desired file in a cluster. Due to the symmetry of the network, P_u^c is the same for all users. P_u^c is also called "hit-rate" in some literature [18], [19]. We note that our network setup follows the framework in [6]. Thus please refer to [6] for more rigorous descriptions.

B. Prerequisite for the Analysis of Throughput-Outage Tradeoff

In this section, we analyze the achievable throughput–outage tradeoff defined by the following:

Definition [6]: For a given network and popularity distribution, a throughput–outage pair (T, P_o) is achievable if there exists a caching policy and a transmission policy with outage probability $p_o \leq P_o$ and minimum per-user average throughput $\overline{T}_{\text{min}} \geq T$.

Under the network setup considered in Sec. III.A, we determine the throughput–outage tradeoff by adopting the caching policy maximizing P_u^c and by adjusting the cluster size $g_c(M)$. We thus first provide the following theorem:

Theorem 1: We define $c_2 = qa'$, where $a' = \frac{\gamma}{S(g_c(M)-1)-1}$, and $c_1 \geq 1$ is the solution of the equality $c_1 = 1 + c_2 \log\left(1 + \frac{c_1}{c_2}\right)$. Let $M \to \infty$, $N \to \infty$, and $q \to \infty$. Suppose $g_c(M) \to \infty$ as $M \to \infty$, and denote m^* as the smallest index such that $P_c^*(m^*+1) = 0$. Under the network model in Sec. III.A, the caching distribution $P_c^*(\cdot)$ that maximizes P_u^c is:

$$P_c^*(f) = \left[1 - \frac{\nu}{z_f}\right]^+, \quad f = 1, \ldots, M, \qquad (3)$$

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

$$m^* = \Theta\left(\min\left(\frac{c_1 S g_c(M)}{\gamma}, M\right)\right). \qquad (4)$$

Proof: See Appendix A. □

Observe that $P_c^*(f)$ is monotonically decreasing and m^* determines the number of files whose $P_c^*(f) > 0$. Besides, we can observe that $c_1 \geq 1$ and $c_1 = 1$ only if $c_2 = o(1)$. Furthermore, we can see that $c_1 = \Theta(c_2)$ when $c_2 = \Omega(1)$. Thus, when considering $q = \Omega\left(\frac{Sg_c(M)}{\gamma}\right)$ and $\frac{c_1 S g_c(M)}{\gamma} < M$, we obtain $m^* = \Theta(\frac{c_1 S g_c(M)}{\gamma}) = \Theta(\frac{c_2 S g_c(M)}{\gamma}) = \Theta(q)$. Combining above results, Theorem 1 indicates that the caching policy should cover at least up to the file at rank q (order-wise) in the library. This is intuitive because the MZipf distribution has a relatively flat head and q characterizes the breaking point.

Using the result in Theorem 1, we then characterize P_u^c, i.e., the probability that a user can find the desired file in a cluster, in Corollaries 1 and 2:

Corollary 1: Let $M \to \infty$, $N \to \infty$, and $q \to \infty$. Suppose $g_c(M) \to \infty$ as $M \to \infty$. Consider $q = \mathcal{O}\left(\frac{Sg_c(M)}{\gamma}\right)$ and $g_c(M) < \frac{\gamma M}{c_1 S}$. Under the network model in Sec. III.A and the caching policy in Theorem 1, P_u^c is expressed as (5) on the top of next page.

Corollary 2: Let $M \to \infty$, $N \to \infty$, and $q \to \infty$. Suppose $g_c(M) \to \infty$ as $M \to \infty$. Consider $q = \mathcal{O}\left(\frac{Sg_c(M)}{\gamma}\right)$ and $g_c(M) = \frac{\rho M}{c_1 S}$, where $\rho \geq \gamma$. Define $D = \frac{q}{M}$. Under the network model in Sec. III.A and the caching policy in Theorem 1, P_u^c is lower bounded as (6) on the top of next page.

Proof: See Appendix B. □

$$P_u^c = \frac{\left(\frac{c_1 S g_c(M)}{\gamma} + q\right)^{1-\gamma}}{(M+q)^{1-\gamma} - (q+1)^{1-\gamma}} - \frac{(1-\gamma)\left(\frac{c_1 S g_c(M)}{\gamma} + q\right)^{-\gamma}\left(\frac{c_1 S g_c(M)}{\gamma}\right)}{(M+q)^{1-\gamma} - (q+1)^{1-\gamma}} - \frac{(q+1)^{1-\gamma}}{(M+q)^{1-\gamma} - (q+1)^{1-\gamma}}. \tag{5}$$

$$P_u^c \geq 1 - \frac{(1-\gamma)e^{-(\rho/c_1-\gamma)}}{(1+D)^{1-\gamma} - (D)^{1-\gamma}}\left[(1+D)^{\frac{\gamma}{S(g_c(M)-1)-1}+1} - (D)^{\frac{\gamma}{S(g_c(M)-1)-1}+1}\right]^{-(S(g_c(M)-1)-1)}. \tag{6}$$

C. Throughput-Outage Tradeoff for MZipf Distributions With $\gamma < 1$

Using the previous results, we characterize the throughput–outage tradeoff for $\gamma < 1$ in the following theorems.

Theorem 2: Let $M \to \infty$, $N \to \infty$, and $q \to \infty$. Suppose $g_c(M) \to \infty$ as $M \to \infty$. Consider $M = \mathcal{O}(N)$, $q = \mathcal{O}\left(\frac{Sg_c(M)}{\gamma}\right)$, and $\gamma < 1$. Denote $\alpha = \frac{1-\gamma}{2-\gamma}$ (i.e., $\gamma = \frac{2\alpha-1}{\alpha-1}$). Under the network model in Sec. III.A and the caching policy in Theorem 1, we characterize the throughput–outage tradeoff achievable by adopting the caching policy in Theorem 1 using three regimes:

(i) Define $c_4 = \frac{q}{M^\alpha}$. When $g_c(M) = c_3 M^\alpha$, where $c_3 = \Theta(1)$, the achievable throughput–outage tradeoff is described by (7) on the bottom of next page.

(ii) Define $c_5 = \frac{q}{g_c(M)}$. When $g_c(M) = \omega(M^\alpha) < \frac{\gamma M}{c_1 S}$, the achievable throughput–outage tradeoff is described by (8) on the bottom of next page.

(iii) Define $D = \frac{q}{M}$. When $g_c(M) = \frac{\rho M}{c_1 S}$, where $\rho \geq \gamma$, the achievable throughput-outage tradeoff is described by (9) on the bottom of next page.

Proof: See Appendix C. □

By comparing Theorem 2 with Theorem 5 in [6], we observe that, when $q = \mathcal{O}\left(\frac{Sg_c(M)}{\gamma}\right)$, the scaling order of the throughput-outage tradeoff in MZipf popularity distribution is identical to that in the Zipf popularity distribution.[4]

Theorem 2 indicates that the achievable throughput–outage tradeoff has the same scaling law as the Zipf distribution when the order of q is no larger than the the order of the aggragate memory, indicating that the performance improvement using the cache-aided D2D network with Zipf distribution can be retained when the popularity distribution follows the more practical MZipf distribution. In particular, since other regimes could have unacceptable high outage, the only regime we are interested in is the third regime of Theorem 2. We can then see from the results that the throughput scales with respect to $\Theta(\frac{S}{M})$, meaning that the throughput of cache-aided D2D scales much better than the conventional unicasting when N is much greater than M (small library), i.e., $T \propto \frac{S}{M} >> \frac{1}{N}$. Besides, the throughput scales linearly with respect to the memory size of each device. The results also imply that cache-aided D2D has the same scaling law as the coded multicasting scheme of [8] and is better than Harmonic Broadcasting [41]. For the detailed discussions regarding scaling laws of different schemes, please refer to [6].

Theorem 2 does not characterize the case that $q = \omega\left(\frac{Sg_c(M)}{\gamma}\right)$. We thus provide the relevant discussions. Specifically, we consider the regime that $q = \omega\left(\frac{Sg_c(M)}{\gamma}\right)$ while $q = \mathcal{O}(M)$. This is because when $q = \omega(M)$, the popularity distribution becomes a uniform distribution asymptotically, in which we are not interested. We then provide Theorem 3.

Theorem 3: Let $M \to \infty$, $N \to \infty$, and $q \to \infty$. Suppose $g_c(M) \to \infty$ as $M \to \infty$. Consider $\gamma < 1$, $q = \omega\left(\frac{Sg_c(M)}{\gamma}\right)$, and $q = \mathcal{O}(M)$ (i.e, $g_c(M) = o(M)$). Under the network model in Sec. III.A and the caching policy in Theorem 1, the achievable outage is lower bounded by 1 asymptotically, i.e., $P_o \geq 1 - o(1)$.

Proof: See Appendix D. □

Theorem 3 suggests that we should increase the cluster size such that the aggregate memory is at least the same order of q, i.e., $Sg_c(M) = \Omega(q)$; otherwise the outage will always go to 1. In practice, this implies the outage of the network will be excessive if the aggregate memory is not large enough to accommodate caching at least to the order of q files.

D. Throughput-Outage Tradeoff for MZipf Distributions With $\gamma > 1$

From Theorem 2, we understand that when $\gamma < 1$, the only meaningful regime is the third regime. In practice, this implies that it is necessary to have a high density D2D network (or on the other hand, a small library) for realizing the benefits of D2D caching. In this section, we want to see whether this condition can be relaxed when $\gamma > 1$, i.e., the popularity is more concentrated on the popular files located in the flat regime of the MZipf distribution. Since Theorem 3 suggests to have a sufficient aggregate memory, we thus focus on the first two regimes of Theorem 2. Specifically, we are interested in the scenario that $g_c(M) = o(M)$ and $q = \mathcal{O}\left(\frac{Sg_c(M)}{\gamma}\right)$:[5]

Theorem 4: Let $M \to \infty$, $N \to \infty$, and $q \to \infty$. Suppose $g_c(M) \to \infty$ as $M \to \infty$. Consider $\gamma > 1$, $g_c(M) = o(M) \leq N$, and $q = \mathcal{O}\left(\frac{Sg_c(M)}{\gamma}\right)$. Define $c_6 = \frac{q}{g_c(M)}$. Under the network model in Sec. III.A and the caching policy in Theorem 1, the achievable throughput–outage tradeoff is

$$T(P_o) = \frac{C}{K}\frac{1}{g_c(M)} + o\left(\frac{1}{g_c(M)}\right), \tag{10}$$

[4] By observing Theorem 2, it is then clear that we are not interested in cases that $g_c(M) = o(M^\alpha)$ and $g_c(M) = \omega(M)$ since the former one gives an even worse outage, i.e., $P_o \to 1$, and the latter one gives worse throughput when $P_o \to 0$.

[5] We actually can see from Theorem 4 that we need $q = \mathcal{O}\left(\frac{Sg_c(M)}{\gamma}\right)$ to bound P_o away from 1.

where $P_o = (c_6)^{\gamma-1}\frac{Sc_1+c_6}{(\frac{Sc_1}{\gamma}+c_6)^\gamma}$.

Proof: See appendix E. □

If $q = o\left(\frac{Sg_c(M)}{\gamma}\right)$, we obtain $c_1 = 1$ and $c_6 = o(1)$ by definition. We thus have Corollary 3:

Corollary 3: Let $M \to \infty$, $N \to \infty$, and $q \to \infty$. Suppose $g_c(M) \to \infty$ as $M \to \infty$. Consider $\gamma > 1$, $g_c(M) = o(M) \leq N$, and $q = o\left(\frac{Sg_c(M)}{\gamma}\right)$. Under the network model in Sec. III.A and the caching policy in Theorem 1, the achievable throughput–outage tradeoff is

$$T(P_o) = \frac{C}{K}\frac{1}{g_c(M)} + o\left(\frac{1}{g_c(M)}\right), \tag{11}$$

where $P_o = o(1)$.

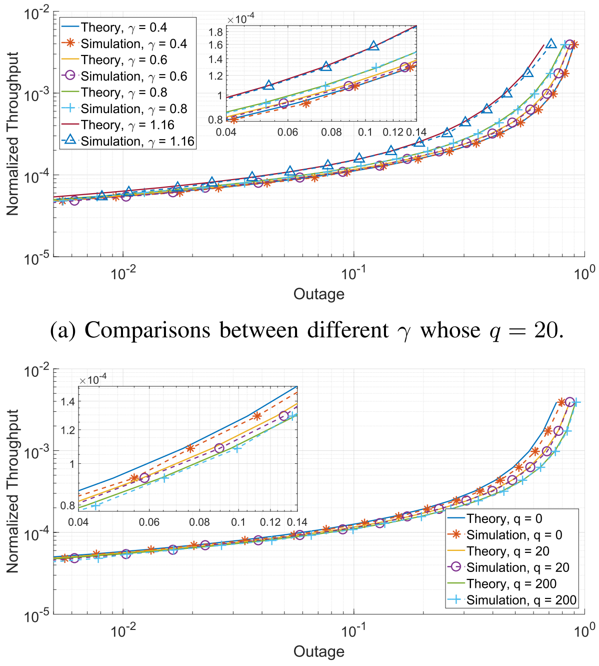
From Theorem 4 and Corollary 3, we observe that when $\gamma > 1$, we obtain the scaling law that is better than $\Theta(\frac{S}{M})$ but worse than $\Theta(\frac{S}{q})$. In practice, it implies that when $\gamma > 1$ and the aggregate memory is larger than the order of q, the improvement of the cache-aided D2D could still be significant even if we have a large library. This relaxes the condition that we need a small library to have significant benefits when $\gamma < 1$.

E. Finite-Dimensional Simulations

Finally, we provide results from finite-dimensional simulations in Fig. 3, which compares theoretical (solid lines) and simulated (dashed lines) curves. In Fig. 3, we adopt $K = 4$, $S = 1$, $M = 1000$, and $N = 10000$. We observe that our analysis can effectively characterize (with small gap) the throughput–outage tradeoff even with finite dimensional setups. This is not common, as indicated by [6], when analyzing the scaling behavior of wireless networks. We note that although not being shown here for brevity, simulations with other parameters, e.g., $N = 5000$ and/or $M = 1500$, show similar results.

IV. Evaluations of Cache-Aided D2D Networks

Our theoretical analysis shows that the cache-aided D2D scheme outperforms the conventional unicasting even if the popularity distribution follows the more practical MZipf distribution. To support the theory, we present simulations of the throughput-outage tradeoff using MZipf distributions parameterized according to the real-world data in the network considering practical setups as in [15]. For the simualtions, communications between users occur at 2.4 GHz. We assume a cell of dimensions 0.36km^2 ($600\text{m} \times 600\text{m}$) that contains buildings as well as streets/outdoor environments. We assume $N = 10000$ users in the cell, i.e., on average, there are $2 \sim 3$ nodes in each square of 10×10 meters. The cell contains a Manhattan grid of square buildings with side length of 50m, separated by streets of width 10m. Each building is made up of offices of size $6.2\text{m} \times 6.2\text{m}$. Within the cell, users (devices) are distributed at random according to a uniform distribution. Due to our geometrical setup, each node is assigned to be outdoors or indoors, and in the latter case placed in a particular office. Since 2.4 GHz communication can penetrate walls, we have to account for different scenarios, which are indoor communication (Winner model A1), outdoor-to-indoor communication (B4), indoor-to-outdoor communication (A2), and outdoor communication (B1) (see [15]).



(a) Comparisons between different γ whose $q = 20$.

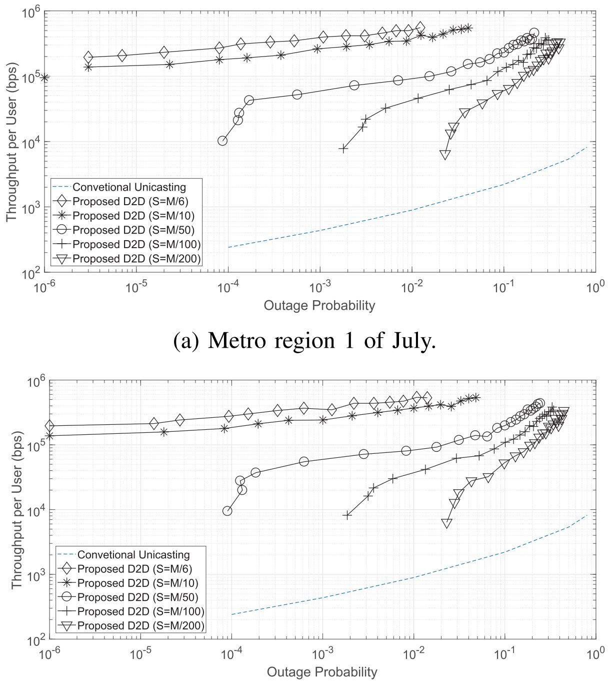
(b) Comparisons between different q whose $\gamma = 0.6$.

Fig. 3. Comparison between the normalized theoretical result (solid lines) and normalized simulated result (dashed lines) in networks adopting $K = 4$, $S = 1$, $M = 1000$, and $N = 10000$.

$$T(P_o) = \frac{C}{K}\frac{M^{-\alpha}}{c_3}\left(1 - \exp\left(\frac{-c_3}{2}\left[\left(\frac{Sc_1c_3}{\gamma} + c_4\right)^{-\gamma}(Sc_1c_3 + c_4) - (c_4)^{1-\gamma}\right]\right)\right) + o(M^{-\alpha}),$$
$$P_o = 1 - M^{-\alpha}\left[\left(\frac{Sc_1c_3}{\gamma} + c_4\right)^{-\gamma}(Sc_1c_3 + c_4) - (c_4)^{1-\gamma}\right] \tag{7}$$

$$T(P_o) = \frac{C}{K}\frac{1}{g_c(M)} + o\left(\frac{1}{g_c(M)}\right), P_o = 1 - \frac{(g_c(M))^{1-\gamma}}{(M + c_5g_c(M))^{1-\gamma} - (c_5g_c(M)+1)^{1-\gamma}}\left[\left(\frac{Sc_1}{\gamma} + c_5\right)^{-\gamma}(Sc_1 + c_5) - (c_5)^{1-\gamma}\right] \tag{8}$$

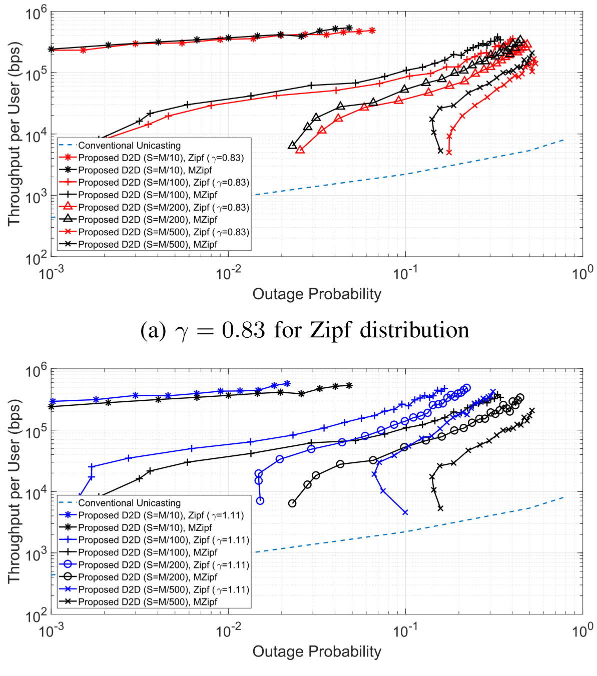
$$T(P_o) = \frac{C}{K}\frac{Sc_1}{\rho M} + o\left(\frac{1}{M}\right), P_o = \frac{(1-\gamma)e^{-(\rho/c_1-\gamma)}}{(1+D)^{1-\gamma} - (D)^{1-\gamma}}\left[(1+D)^{\frac{\gamma}{S(g_c(M)-1)-1}+1} - (D)^{\frac{\gamma}{S(g_c(M)-1)-1}+1}\right]^{-(S(g_c(M)-1)-1)} \tag{9}$$



(a) Metro region 1 of July.

(b) Metro region 2 of July.

Fig. 4. Throughput outage tradeoff in networks assuming mixed office scenario for propagation channel; varying local storage size.



(a) $\gamma = 0.83$ for Zipf distribution

(b) $\gamma = 1.11$ for Zipf distribution

Fig. 5. Throughput outage tradeoff comparisons between different models for Metro region 2 of July in networks assuming mixed office scenario for propagation channel; varying local storage size.

The number of clusters in a cell is varied from $2^2 = 4, 3^2 = 9, \ldots 27^2 = 729$; a frequency reuse factor of $K = 4$ is used to minimize the inter-cluster interference. The cache memory on each device S is kept as a parameter that will be varied in the simulations. To provide some real-world connections: storage of an hour-long video in medium video quality (750 kbps, suitable for a cellphone) takes about 300 MByte. Thus, storing 100 files with current cellphones is reasonably realistic, and given the continuous increase in memory size, even storing 500 files is not prohibitive (assuming some incentivization by network operators or other entities).

In terms of channel models, we mostly employ the Winner channel models with a minor modification, motivated by the fact that it is difficult for establish a D2D link at low SNR [15], that no D2D communication is possible for a distance larger than 100 m. In particular, we directly use Winner II channel models with antenna heights of 1.5m, as well as the probabilistic Line of Sight (LOS) and Non Line of Sight (NLOS) models. We add a probabilistic body shadowing loss (σ_{L_b}) with a lognormal distribution, where for LOS, $\sigma_{L_b} = 4.2$ and for NLOS, $\sigma_{L_b} = 3.6$ to account for the blockage of radiation by the person holding the device; see [42]. More details about the channel model can be found in [15].

Since Metro regions 1 and 2 of the dataset cover much smaller regions compared to the whole UK, and thus are expected to describe better the effects that might be encountered within a particular cell (though they are still much larger than a cell), we use their corresponding parameters for MZipf distributions in the simulations.

Fig. 4(a) shows the throughput-outage tradeoff for different cache sizes on each device in Metro region 1. An outage of 10% implies that 90% of traffic can be offloaded to the D2D communications. We can see that the throughput of 10^5 bps can be achieved if the cache size of each user is up to $1/10$ of the library size. Even for $S = M/50$, i.e., approximate 100 files (30 GB), the advantage compared to conventional unicasting described in [15] is two orders of magnitude. Even just caching of 30 files ($M/200$) also provides significant throughput gains, though only for outage probabilities > 0.01. The results for Metro region 2 (Fig. 4(b)) are very similar.

Finally, we verify whether using the more detailed model of the MZipf distribution has an impact on the *performance* of the caching system. Note that here it is not important whether the throughput is better or worse with a specific model, but whether it is *correct*. In other words, is there a difference in performance when using the MZipf fit (more complicated, but better fit, as discussed above), or the Zipf fit? The short answer is that indeed there is a difference, as explained in detail in the following.

In Fig. 5, we consider modeling results of Metro region 2 of July, and compare the proposed MZipf model with the Zipf model adopting the best-fitting parameter $\gamma = 0.83$ (fits the whole data curve) and the best-tail-fitting parameter $\gamma = 1.11$ (fits the power law of the tail). When comparing using the MZipf model to using the Zipf model with the best-fitting parameter ($\gamma = 0.83$), we can observe a performance gap between them, and such gap increases as the storage size decreases. This is because a Zipf model fails to model the flattened head of the popularity distribution, and this drawback is significant when the storage size of devices is not large enough to completely store all files in the flattened head. When comparing with the Zipf model using the best-tail-fitting parameter ($\gamma = 1.11$), the gap between the Zipf and MZipf models is even more significant, indicating simply fitting the power law of the tail could lead to a fairly inaccurate result. Above results indicate that, when using the inaccurate Zipf distribution to evaluate the system performance, it might generate an inaccurate result. Therefore it is necessary to use the MZipf distribution for modeling and analysis. As a remark, we observe in some curves that decreasing the throughput (by increasing the cluster size) does not improve the outage. This is because when the cluster size is large and the storage size of devices is small, users need to fetch the desired files from their neighbors with distances larger than 100 m. Since we assume that a D2D link with a distance longer than 100 m is prohibited, a large cluster size accompanied by a small storage size leads to high channel outage and increases the overall outage.

V. Conclusion

To answer the open question whether cache-aided D2D for video distribution can provide in practice the benefits promised in the literature, we analyze and evaluate the throughput–outage performance considering measured popularity distributions. Using an extensive dataset, we observe that the widely used Zipf distribution cannot effectively describe the popularity distribution of real wireless traffic data. We thus propose using a generalized version of Zipf distribution, i.e., the MZipf distribution, to model and parameterize the real data. Comparisons using measurements and numerical simulations verify the accuracy and necessity of this modeling. Considering such generalized modeling, we generalize the theoretical treatment in [6] and analyze the throughout–outage tradeoff. In particular, we show the impact of the plateau factor q of the MZipf distribution in the optimal caching distribution and the throughput-outage tradeoff. Theoretical results show that the scaling behavior of the cache-aided D2D is identical to case of Zipf distribution under some parameter regimes validated by real data, implying that the benefits in the case of the Zipf distribution could be retained. To support the theory, extensive numerical evaluations considering practical propagation scenarios and other details are provided, and show that the cache-aided D2D for video distribution significantly outperforms the conventional unicasting. Since the theory and numerical experiments both suggest positive results, we thus conclude that the cache-aided D2D for video distribution can in practice provide the benefits promised in the existing literature.

Appendix A
Proof of Theorem 1

In this section, our goal is to find the caching policy that maximizes P_u^c. Note that the probability that a user u can find its desired file f in the cluster through D2D communications is $1-(1-P_c(f))^{S(g_c(M)-1)}$. Then by using the law of total probability, we have

$$P_u^c = \sum_{f=1}^{M} P_r(f)\left(1-(1-P_c(f))^{S(g_c(M)-1)}\right). \quad (12)$$

To maximize P_u^c, we follow the similar approach based on convex minimization and KKT conditions in Appendix C of [6] and obtain

$$P_c^*(f) = \left[1-\left(\frac{\lambda}{P_r(f)S(g_c(M)-1)}\right)^{\frac{1}{S(g_c(M)-1)-1}}\right]^+. \quad (13)$$

Next, we need to find the λ such that $\sum_{f=1}^{M} P_c^*(f) = 1$. Let $\nu = \left(\frac{\lambda}{S(g_c(M)-1)}\right)^{\frac{1}{S(g_c(M)-1)-1}}$ and $z_f = (P_r(f))^{\frac{1}{S(g_c(M)-1)-1}}$. Note that z_f is non-increasing with respect to f since $P_r(f)$ is non-increasing. By following the similar argument in appendix C of [6], we obtain that $\nu = \frac{m^*-1}{\sum_{f=1}^{m^*}\frac{1}{z_f}}$, satisfying $\nu \geq z_{m^*+1}$ and $\nu \leq z_{m^*}$. Thus if m^* is a unique integer in $\{1, 2, \cdots, M-1\}$, it satisfies:

$$m^* \geq 1 + z_{m^*+1}\sum_{f=1}^{m^*}\frac{1}{z_f}$$

and

$$m^* \leq 1 + z_{m^*}\sum_{f=1}^{m^*}\frac{1}{z_f}.$$

Then in order to determine m^* as a function of $g_c(M)$ in the assumption that $g_c(M) \to \infty$ as $M \to \infty$, we need to evaluate

$$\begin{aligned} z_{m^*+1}&\sum_{f=1}^{m^*}\frac{1}{z_f} \\ &= (m^*+q+1)^{\frac{-\gamma}{S(g_c(M)-1)-1}}\sum_{f=1}^{m^*}(f+q)^{\frac{\gamma}{S(g_c(M)-1)-1}} \\ &= \left(\frac{1}{m^*+q+1}\right)^{a'}\sum_{f=1}^{m^*}(f+q)^{a'}, \end{aligned} \quad (14)$$

and

$$\begin{aligned} z_{m^*}\sum_{f=1}^{m^*}\frac{1}{z_f} &= (m^*+q)^{\frac{-\gamma}{S(g_c(M)-1)-1}}\sum_{f=1}^{m^*}(f+q)^{\frac{\gamma}{S(g_c(M)-1)-1}} \\ &= \left(\frac{1}{m^*+q}\right)^{a'}\sum_{f=1}^{m^*}(f+q)^{a'}, \end{aligned} \quad (15)$$

where $a' = \frac{\gamma}{S(g_c(M)-1)-1}$. We first characterize $\sum_{f=1}^{m^*}(f+q)^{a'}$. By using the fundamental concept of integration, we observe that

$$\begin{aligned} \sum_{f=1}^{m^*}(f+q)^{a'} &\leq \int_1^{m^*+1}(x+q)^{a'}dx \\ &= \frac{(m^*+q+1)^{a'+1}-(q+1)^{a'+1}}{a'+1}; \\ \sum_{f=1}^{m^*}(f+q)^{a'} &\geq (1+q)^{a'} + \int_1^{m^*}(x+q)^{a'}dx \\ &= (1+q)^{a'} + \frac{(m^*+q)^{a'+1}-(q+1)^{a'+1}}{a'+1}. \end{aligned} \quad (16)$$

By using (16), we can obtain the upper (UB 1) bound and lower bound (LB 1) of (14):

$$\begin{aligned}&\text{LB 1}\\&=\left(\frac{q+1}{m^*+q+1}\right)^{a'}+\frac{1}{a'+1}\\&\quad\cdot\left[(m^*+q)\left(\frac{m^*+q}{m^*+q+1}\right)^{a'}-(q+1)\left(\frac{q+1}{m^*+q+1}\right)^{a'}\right],\\&\text{UB 1}\\&=\frac{1}{a'+1}\left[(m^*+q+1)-(q+1)\left(\frac{q+1}{m^*+q+1}\right)^{a'}\right].\end{aligned}\tag{17}$$

Similarly, we can obtain the upper (UB 2) bound and lower bound (LB 2) of (15):

$$\begin{aligned}\text{LB 2}&=\left(\frac{q+1}{m^*+q}\right)^{a'}\\&\quad+\frac{1}{a'+1}\left[(m^*+q)-(q+1)\left(\frac{q+1}{m^*+q}\right)^{a'}\right],\\\text{UB 2}&=\frac{1}{a'+1}\\&\quad\cdot\left[(m^*+q+1)\left(\frac{m^*+q+1}{m^*+q}\right)^{a'}\right.\\&\quad\left.-(q+1)\left(\frac{q+1}{m^*+q}\right)^{a'}\right].\end{aligned}\tag{18}$$

We then define $c_1 = m^*a'$ and $c_2 = qa'$. Notice that $a' \downarrow 0$ as $g_c(M) \to \infty$. We therefore obtain (19) and (20) on the bottom of this page, where $\delta_i(a')$, $i = 1, \dots, 4$, tend to zeros as $a' \downarrow 0$.

Then we denote that

$$\begin{aligned}1-\delta_1(a')&=\left(\frac{c_2+a'}{c_1+c_2+a'}\right)^{a'}=(\nu_1)^{a'},\\1-\delta_2(a')&=\left(\frac{c_1+c_2}{c_1+c_2+a'}\right)^{a'}=(\nu_2)^{a'},\\1-\delta_3(a')&=\left(\frac{c_2+a'}{c_1+c_2}\right)^{a'}=(\nu_3)^{a'},\\1-\delta_4(a')&=\left(\frac{c_1+c_2+a'}{c_1+c_2}\right)^{a'}=(\nu_4)^{a'}.\end{aligned}\tag{21}$$

It follows that

$$\frac{c}{a'}\delta_i(a')=\frac{c\left[1-(\nu_i)^{a'}\right]}{a'}\overset{a'\to 0}{=}-c\log(\nu_i),\quad i=1,\dots,4,\tag{22}$$

where the second equality is by L'Hôspital's rule. Thus, suppose $c = \mathcal{O}(c_1 + c_2)$, we obtain

$$\begin{aligned}&\frac{c}{a'}\delta_1(a')\overset{a'\to 0}{=}c\log\left(1+\frac{c_1}{c_2}\right);\ \frac{c}{a'}\delta_2(a')\overset{a'\to 0}{=}0;\\&\frac{c}{a'}\delta_3(a')\overset{a'\to 0}{=}c\log\left(1+\frac{c_1}{c_2}\right);\ \frac{c}{a'}\delta_4(a')\overset{a'\to 0}{=}0.\end{aligned}\tag{23}$$

By using the above results and that $m^* = \frac{c_1}{a'}$, it follows that, when $a' \to 0$, we obtain

$$\frac{c_1/a'+\epsilon}{a'+1}+1\lessapprox\frac{c_1}{a'}\lessapprox\frac{c_1/a'+\epsilon}{a'+1}+1,\tag{24}$$

where $\epsilon = c_2 \log\left(1 + \frac{c_1}{c_2}\right)$. Thus, we obtain

$$\frac{c_1}{a'}\left(1-\frac{1}{a'+1}\right)=\frac{c_1}{a'+1}\cong 1+\frac{\epsilon}{a'+1},\tag{25}$$

leading to $c_1 \cong a' + 1 + \epsilon = 1 + \epsilon$. We then conclude that

$$m^*=\frac{c_1}{a'}=\frac{c_1(S(g_c(M)-1)-1)}{\gamma}+\mathcal{O}(1),\tag{26}$$

where c_1 satisfies the equality $c_1 = 1 + c_2 \log\left(1 + \frac{c_1}{c_2}\right)$ and $c_2 = qa'$. This indicates $m^* = \frac{c_1 S g_c(M)}{\gamma}$ to the leading order. Besides, it should be clear that if $\frac{c_1 S g_c(M)}{\gamma} \geq M$, we have $m^* = M$. We also note that when $q = 0$, our result degenerates to the results in [6] (Observe that when $q = 0$, we obtain $c_2 = 0$ and $c_1 = 1$).

$$\begin{aligned}\text{LB 1}&=\left(\frac{\frac{c_2}{a'}+1}{\frac{c_1+c_2}{a'}+1}\right)^{a'}+\frac{1}{a'+1}\left[\left(\frac{c_1+c_2}{a'}\right)\left(\frac{\frac{c_1+c2}{a'}}{\frac{c_1+c_2}{a'}+1}\right)^{a'}-\left(\frac{c_2}{a'}+1\right)\left(\frac{\frac{c_2}{a'}+1}{\frac{c_1+c_2}{a'}+1}\right)^{a'}\right]=1-\delta_1(a')+\frac{1}{1+a'}\\&\quad\cdot\left[\left(\frac{c_1+c_2}{a'}\right)(1-\delta_2(a'))-\left(\frac{c_2}{a'}+1\right)(1-\delta_1(a'))\right]=\frac{\left[\left(\frac{c_1+c_2}{a'}\right)(1-\delta_2(a'))-\left(\frac{c_2}{a'}\right)(1-\delta_1(a'))-a'(1-\delta_1(a'))\right]}{1+a'}\\\text{UB 1}&=\frac{1}{a'+1}\left[\left(\frac{c_1+c_2}{a'}+1\right)-\left(\frac{c_2}{a'}+1\right)\left(\frac{\frac{c_2}{a'}+1}{\frac{c_1+c_2}{a'}+1}\right)^{a'}\right]=\frac{\left[\left(\frac{c_1+c_2}{a'}+1\right)-\left(\frac{c_2}{a'}+1\right)(1-\delta_1(a'))\right]}{1+a'}.\end{aligned}\tag{19}$$

$$\begin{aligned}\text{LB 2}&=\left(\frac{\frac{c_2}{a'}+1}{\frac{c_1+c_2}{a'}}\right)^{a'}+\frac{1}{a'+1}\left[\left(\frac{c_1+c_2}{a'}\right)-\left(\frac{c_2}{a'}+1\right)\left(\frac{\frac{c_2}{a'}+1}{\frac{c_1+c_2}{a'}}\right)^{a'}\right]=\frac{\left[\left(\frac{c_1+c_2}{a'}\right)-\left(\frac{c_2}{a'}\right)(1-\delta_3(a'))-a'(1-\delta_3(a'))\right]}{1+a'}\\\text{UB 2}&=\frac{1}{a'+1}\left[\left(\frac{c_1+c_2}{a'}+1\right)\left(\frac{\frac{c_1+c_2}{a'}+1}{\frac{c_1+c_2}{a'}}\right)^{a'}-\left(\frac{c_2}{a'}+1\right)\left(\frac{\frac{c_2}{a'}+1}{\frac{c_1+c_2}{a'}}\right)^{a'}\right]\\&=\frac{\left[\left(\frac{c_1+c_2}{a'}+1\right)(1+\delta_4(a'))-\left(\frac{c_2}{a'}+1\right)(1-\delta_3(a'))\right]}{1+a'}.\end{aligned}\tag{20}$$

$$
\begin{aligned}
P_u^c &= \sum_{f=1}^{M} P_r(f)\left(1-(1-P_c(f))^{S(g_c(M)-1)}\right) = \sum_{f=1}^{m^*} P_r(f)\left(1-\left(\frac{\nu}{z_f}\right)^{S(g_c(M)-1)}\right)\\
&\overset{(a)}{\leq} \sum_{f=1}^{m^*} P_r(f) - \sum_{f=1}^{m^*} P_r(f)\left(\frac{P_r(m^*+1)}{P_r(f)}\right)\cdot\left(\frac{P_r(m^*+1)}{P_r(f)}\right)^{\frac{1}{S(g_c(M)-1)-1}}\\
&= \sum_{f=1}^{m^*} P_r(f) - P_r(m^*+1)\sum_{f=1}^{m^*}\left(\frac{f+q}{m^*+1+q}\right)^{\frac{\gamma}{S(g_c(M)-1)-1}}\\
&= \frac{H(\gamma,q,1,m^*)}{H(\gamma,q,1,m)} - \frac{(m^*+q+1)^{-\gamma}}{H(\gamma,q,1,m)}\sum_{f=1}^{m^*}\left(\frac{f+q}{m^*+1+q}\right)^{\frac{\gamma}{S(g_c(M)-1)-1}}\\
&\overset{(b)}{\leq} \frac{\frac{1}{1-\gamma}(m^*+q)^{1-\gamma} - \frac{1}{1-\gamma}(q+1)^{1-\gamma} + (1+q)^{-\gamma} - (m^*+q+1)^{-\gamma}\sum_{f=1}^{m^*}\left(\frac{f+q}{m^*+1+q}\right)^{\frac{\gamma}{S(g_c(M)-1)-1}}}{\frac{1}{1-\gamma}(M+q+1)^{1-\gamma} - \frac{1}{1-\gamma}(q+1)^{1-\gamma}}\\
&= \frac{(m^*+q)^{1-\gamma} - (q+1)^{1-\gamma} + (1-\gamma)(1+q)^{-\gamma}}{(M+q+1)^{1-\gamma} - (q+1)^{1-\gamma}} - \frac{(1-\gamma)(m^*+q+1)^{-\gamma}\left(\frac{1}{m^*+1+q}\right)^{\frac{\gamma}{S(g_c(M)-1)-1}}\sum_{f=1}^{m^*}(f+q)^{\frac{\gamma}{S(g_c(M)-1)-1}}}{(M+q+1)^{1-\gamma} - (q+1)^{1-\gamma}}\\
&\overset{(c)}{\leq} \frac{(m^*+q)^{1-\gamma} - (q+1)^{1-\gamma} + (1-\gamma)(1+q)^{-\gamma}}{(M+q+1)^{1-\gamma} - (q+1)^{1-\gamma}}\\
&\quad - \frac{(1-\gamma)(m^*+q+1)^{-\gamma}\left(\frac{1}{m^*+1+q}\right)^{\frac{\gamma}{S(g_c(M)-1)-1}}\left[(1+q)^{\frac{\gamma}{S(g_c(M)-1)-1}} + \int_1^{m^*}(x+q)^{\frac{\gamma}{S(g_c(M)-1)-1}}dx\right]}{(M+q+1)^{1-\gamma} - (q+1)^{1-\gamma}}\\
&= \frac{(m^*+q)^{1-\gamma} - (q+1)^{1-\gamma} + (1-\gamma)(1+q)^{-\gamma}}{(M+q+1)^{1-\gamma} - (q+1)^{1-\gamma}} - \frac{(1-\gamma)(m^*+q+1)^{-\gamma}\left(\frac{1}{m^*+1+q}\right)^{\frac{\gamma}{S(g_c(M)-1)-1}}}{(M+q+1)^{1-\gamma} - (q+1)^{1-\gamma}}\\
&\quad \cdot\left[(1+q)^{\frac{\gamma}{S(g_c(M)-1)-1}} + \frac{1}{\frac{\gamma}{S(g_c(M)-1)-1}+1}\left((m^*+q)^{\frac{\gamma}{S(g_c(M)-1)-1}+1} - (q+1)^{\frac{\gamma}{S(g_c(M)-1)-1}+1}\right)\right]\\
&\overset{(d)}{=} \frac{\left(\frac{c_1 S g_c(M)}{\gamma}+q\right)^{1-\gamma} - (q+1)^{1-\gamma}}{(M+q)^{1-\gamma} - (q+1)^{1-\gamma}} - \frac{(1-\gamma)\left(\frac{c_1 S g_c(M)}{\gamma}+q\right)^{-\gamma}\left(\frac{c_1 S g_c(M)}{\gamma}\right)}{(M+q)^{1-\gamma} - (q+1)^{1-\gamma}} + o\left(\frac{\left(\frac{c_1 S g_c(M)}{\gamma}+q\right)^{1-\gamma}}{(M+q)^{1-\gamma} - (q+1)^{1-\gamma}}\right). \qquad (27)
\end{aligned}
$$

Appendix B
Proof of Corollaries 1 and 2

Before starting the main proof, we first provide a useful Lemma:

Lemma 1: Denote $\sum_{m=a}^{b}(m+q)^{-\gamma} = H(\gamma,q,a,b)$. When $\gamma \neq 1$, we have

$$
\begin{aligned}
&\frac{1}{1-\gamma}\left[(b+q+1)^{1-\gamma} - (a+q)^{1-\gamma}\right] \leq H(\gamma,q,a,b)\\
&\qquad \leq \frac{1}{1-\gamma}\left[(b+q)^{1-\gamma} - (a+q)^{1-\gamma}\right] + (a+q)^{-\gamma}.
\end{aligned}
$$

Proof. Consider $\gamma \neq 1$. By the fundamental calculus, we have

$$
\begin{aligned}
H(\gamma,q,a,b) &= \sum_{m=a}^{b}(m+q)^{-\gamma} \geq \int_a^{b+1}\frac{dx}{(x+q)^\gamma}\\
&= \frac{1}{1-\gamma}\left[(b+q+1)^{1-\gamma} - (a+q)^{1-\gamma}\right],\\
H(\gamma,q,a,b) &= \sum_{m=a}^{b}(m+q)^{-\gamma} \leq (a+q)^{-\gamma} + \int_a^{b}\frac{dx}{(x+q)^\gamma}\\
&= \frac{1}{1-\gamma}\left[(b+q)^{1-\gamma} - (a+q)^{1-\gamma}\right] + (a+q)^{-\gamma}.
\end{aligned}
$$

A. Proof of Corollary 1

We consider $g_c(M) < \frac{\gamma M}{c_1 S}$ and $q = \mathcal{O}\left(\frac{S g_c(M)}{\gamma}\right)$. We thus obtain $c_1 = \mathcal{O}(1)$ and $m^* < M$. The probability that a user u finds the desired file in the cluster is then expressed as in (27) on the bottom of next page, where (a) is because $\nu \geq z_{m^*+1}$; (b) uses results in Lemma 1; (c) exploits Riemann sum and $m^* \geq 1$; (d) uses Theorem 1 that $m^* = \frac{c_1 S g_c(M)}{\gamma}$ and $g_c(M) \to \infty$.

Similarly, we can obtain (28), where (a) is because $\nu \leq z_{m^*}$ on the top of next next page. By combining the results in (27) and (28), Corollary 1 is proved.

B. Proof of Corollary 2

When $g_c(M) = \frac{\rho M}{c_1 S}$, where $\rho \geq \gamma$, we obtain $m^* = M$. Thus, results in Corollary 1 is no longer appropriate. Now since $m^* = M$, we thus have $\nu = \frac{M-1}{\sum_{f=1}^{M}\frac{1}{z_f}}$. We define $D = \frac{q}{M}$. Then

$$
P_u^c = \sum_{f=1}^{M} P_r(f)\left(1-\left(\frac{\nu}{z_f}\right)^{S(g_c(M)-1)}\right)
$$

$$
\begin{aligned}
P_u^c &\overset{(a)}{\geq} \sum_{f=1}^{m^*} P_r(f) - \sum_{f=1}^{m^*} P_r(f)\left(\frac{P_r(m^*)}{P_r(f)}\right)\cdot\left(\frac{P_r(m^*)}{P_r(f)}\right)^{\frac{1}{S(g_c(M)-1)-1}} \\
&= \frac{H(\gamma,q,1,m^*)}{H(\gamma,q,1,m)} - \frac{(m^*+q)^{-\gamma}}{H(\gamma,q,1,m)}\sum_{f=1}^{m^*}\left(\frac{f+q}{m^*+q}\right)^{\frac{\gamma}{S(g_c(M)-1)-1}} \\
&\geq \frac{\frac{1}{1-\gamma}(m^*+q+1)^{1-\gamma} - \frac{1}{1-\gamma}(q+1)^{1-\gamma} - (m^*+q)^{-\gamma}\sum_{f=1}^{m^*}\left(\frac{f+q}{m^*+q}\right)^{\frac{\gamma}{S(g_c(M)-1)-1}}}{\frac{1}{1-\gamma}(M+q)^{1-\gamma} - \frac{1}{1-\gamma}(q+1)^{1-\gamma} + (1+q)^{-\gamma}} \\
&= \frac{(m^*+q+1)^{1-\gamma} - (q+1)^{1-\gamma}}{(M+q)^{1-\gamma} - (q+1)^{1-\gamma} + (1-\gamma)(1+q)^{-\gamma}} - \frac{(1-\gamma)(m^*+q)^{-\gamma}\left(\frac{1}{m^*+q}\right)^{\frac{\gamma}{S(g_c(M)-1)-1}}\sum_{f=1}^{m^*}(f+q)^{\frac{\gamma}{S(g_c(M)-1)-1}}}{(M+q)^{1-\gamma} - (q+1)^{1-\gamma} + (1-\gamma)(1+q)^{-\gamma}} \\
&\geq \frac{(m^*+q+1)^{1-\gamma} - (q+1)^{1-\gamma}}{(M+q)^{1-\gamma} - (q+1)^{1-\gamma} + (1-\gamma)(1+q)^{-\gamma}} - \frac{(1-\gamma)(m^*+q)^{-\gamma}\left(\frac{1}{m^*+q}\right)^{\frac{\gamma}{S(g_c(M)-1)-1}}\left[\int_1^{m^*+1}(x+q)^{\frac{\gamma}{S(g_c(M)-1)-1}}dx\right]}{(M+q)^{1-\gamma} - (q+1)^{1-\gamma} + (1-\gamma)(1+q)^{-\gamma}} \\
&= \frac{\left(\frac{c_1Sg_c(M)}{\gamma}+q\right)^{1-\gamma} - (q+1)^{1-\gamma} +}{(M+q)^{1-\gamma} - (q+1)^{1-\gamma}} - \frac{(1-\gamma)\left(\frac{c_1Sg_c(M)}{\gamma}+q\right)^{-\gamma}\left(\frac{Sg_c(M)}{\gamma}\right)}{(M+q)^{1-\gamma} - (q+1)^{1-\gamma}} + o\left(\frac{\left(\frac{c_1Sg_c(M)}{\gamma}+q\right)^{1-\gamma}}{(M+q)^{1-\gamma} - (q+1)^{1-\gamma}}\right), \qquad (28)
\end{aligned}
$$

$$
\begin{aligned}
&= 1 - \nu^{S(g_c(M)-1)}\sum_{f=1}^{M}\frac{P_r(f)}{(z_f)^{S(g_c(M)-1)}} \\
&= 1 - \left(\frac{M-1}{\sum_{f=1}^{M}P_r(f)^{\frac{-1}{S(g_c(M)-1)-1}}}\right)^{S(g_c(M)-1)} \\
&\quad \cdot \sum_{f=1}^{M} P_r(f)^{\frac{-1}{S(g_c(M)-1)-1}} \\
&= 1 - (M-1)^{S(g_c(M)-1)} \\
&\quad \cdot \left(\sum_{f=1}^{M}\left(\frac{(f+q)^{-\gamma}}{H(\gamma,q,1,M)}\right)^{\frac{-1}{S(g_c(M)-1)-1}}\right)^{-(S(g_c(M)-1)-1)} \\
&= 1 - \frac{(M-1)^{S(g_c(M)-1)}}{H(\gamma,q,1,M)} \\
&\quad \cdot \frac{1}{\left(\sum_{f=1}^{M}(f+q)^{\frac{\gamma}{S(g_c(M)-1)-1}}\right)^{(S(g_c(M)-1)-1)}}. \qquad (29)
\end{aligned}
$$

Denoting $S(g_c(M)-1)-1$ as φ, it follows from (29) that (30) on the top of next page can be obtained. This complete the Proof of Corollary 2.

Appendix C
Proof of Theorem 2

In this section, we provide the proof for Theorem 2, which lets $M \to \infty$, $N \to \infty$, and $q \to \infty$ and consider $g_c(M) \to \infty$ as $M \to \infty$. We first outline the proof. From Corollaries 1 and 2, we obtain the lower bound of P_u^c, which is determined by the cluster size $g_c(M)$ and the condition of q. Since the outage probability $P_o = 1 - P_u^c$, therefore we can obtain the upper bound of the outage. Subsequently, for each outage regime, we obtain the lower bound of $\overline{T}_{\text{min}}$ by computing the lower bound of the sum throughput $\overline{T}_{\text{sum}}$ and using the result that $\overline{T}_{\text{min}} = \frac{1}{N}\overline{T}_{\text{sum}}$, following the fact that each user is symmetric and has the same average throughput. Since the achievable upper bound of the outage probability and the corresponding lower bound of the throughput can be obtained, we characterize the achievable throughput-outage tradeoff. In Theorem 2, we consider $\gamma < 1$ and the regime covering $q = \mathcal{O}\left(\frac{Sg_c(M)}{\gamma}\right)$. The cases that $\gamma < 1$ and $q = \omega\left(\frac{Sg_c(M)}{\gamma}\right)$ will be considered later in Theorem 3.

The main flow for computing $\overline{T}_{\text{sum}}$ is the following (see also Appendix D in [6]). Denote L as the number of active links, we have

$$\overline{T}_{\text{sum}} = C \cdot \mathbb{E}[L] = C \cdot E[\text{number of active cluster}], \qquad (31)$$

where C is the constant link rate and the second equality is because only one transmission is allowed in a cluster in a time-frequency slot. Then noticing that

$$
\begin{aligned}
&\mathbb{E}[\text{number of active cluster}] \\
&\quad \geq \frac{1}{K}\mathbb{E}[\text{number of good cluster}] \\
&\quad = \frac{1}{K}\left(\text{number of total clusters} \cdot \mathbb{P}(W > 0)\right), \qquad (32)
\end{aligned}
$$

where the K is reuse factor. Recall that a good cluster is where there exists at least one potential link in the cluster. Thus $W = \sum_{u=1}^{g_c(M)} \mathbf{1}_u$ is the number of potential links, where $\mathbf{1}_u$ is the indicator that equals to one if user u can access the desired file in the cluster; otherwise $\mathbf{1}_u = 0$.

A. Proof of Regime 1

In this section, we consider $g_c(M) = c_3M^\alpha$, where $c_3 = \Theta(1)$, and $q = \mathcal{O}\left(\frac{Sg_c(M)}{\gamma}\right)$. We define $c_4 = \frac{q}{M^\alpha}$. According to Corollary 1, we can obtain P_u^c as in (33) on the top of next page, where (a) is because $q = o(M)$ and $M \to \infty$, and (b) is because

$$
\begin{aligned}
(\alpha-1)(1-\gamma) &\overset{(d)}{=} (\alpha-1)\left(1 - \frac{2\alpha-1}{\alpha-1}\right) \\
&= (\alpha-1)\left(\frac{-\alpha}{\alpha-1}\right) = -\alpha, \qquad (34)
\end{aligned}
$$

$$
\begin{aligned}
P_u^c &\geq 1 - \frac{(M-1)^{S(g_c(M)-1)}}{\frac{1}{1-\gamma}(M+q+1)^{1-\gamma} - \frac{1}{1-\gamma}(q+1)^{1-\gamma}} \cdot \frac{1}{\left((1+q)^{\frac{\gamma}{\varphi}} + \int_1^M (x+q)^{\frac{\gamma}{\varphi}} dx\right)^{\varphi}} \\
&= 1 - \frac{(1-\gamma)\,(M-1)^{S(g_c(M)-1)}}{(M+q+1)^{1-\gamma} - (q+1)^{1-\gamma}} \frac{1}{\left[(1+q)^{\frac{\gamma}{\varphi}} + \left(\frac{1}{\frac{\gamma}{\varphi}+1}\right)\left((M+q)^{\frac{\gamma}{\varphi}+1} - (q+1)^{\frac{\gamma}{\varphi}+1}\right)\right]^{\varphi}} \\
&= 1 - \frac{(1-\gamma)\,(M-1)^{S(g_c(M)-1)}}{(M+q+1)^{1-\gamma} - (q+1)^{1-\gamma}} \cdot \underbrace{\left(1 - \frac{\gamma}{\varphi+\gamma}\right)^{(\varphi+\gamma)\frac{\varphi}{\varphi+\gamma}(-1)}}_{=e^{\gamma}} \\
&\quad \cdot \left[\left(\frac{\gamma}{\varphi}+1\right)(1+q)^{\frac{\gamma}{\varphi}} + (M+q)^{\frac{\gamma}{\varphi}+1} - (q+1)^{\frac{\gamma}{\varphi}+1}\right]^{-\varphi} \\
&= 1 - (1-\gamma)\frac{(M-1)^{S(g_c(M)-1)}}{(M)^{S(g_c(M)-1)}} \frac{M^{1-\gamma}}{(M+DM+1)^{1-\gamma} - (DM+1)^{1-\gamma}} \cdot e^{\gamma} \\
&\quad \cdot \left[\left(\frac{\gamma}{\varphi}+1\right)\frac{1}{M} \cdot \left(\frac{1+DM}{M}\right)^{\frac{\gamma}{\varphi}} + (1+D)^{\frac{\gamma}{\varphi}+1} - \left(D + \frac{1}{M}\right)^{\frac{\gamma}{\varphi}+1}\right]^{-\varphi} \\
&= 1 - (1-\gamma)\underbrace{\left(1 - \frac{1}{M}\right)^{S(\frac{\rho M}{c_1 S}-1)}}_{=e^{-\rho/c_1}} \frac{1}{(1+D+\frac{1}{M})^{1-\gamma} - (D+\frac{1}{M})^{1-\gamma}} \cdot e^{\gamma} \\
&\quad \cdot \left[\left(\frac{\gamma}{\varphi}+1\right)\frac{1}{M}\left(D + \frac{1}{M}\right)^{\frac{\gamma}{\varphi}} + (1+D)^{\frac{\gamma}{\varphi}+1} - \left(D + \frac{1}{M}\right)^{\frac{\gamma}{\varphi}+1}\right]^{-\varphi} \\
&= 1 - \frac{(1-\gamma)e^{-(\rho/c_1-\gamma)}}{(1+D)^{1-\gamma} - (D)^{1-\gamma}} \left[(1+D)^{\frac{\gamma}{\varphi}+1} - (D)^{\frac{\gamma}{\varphi}+1}\right]^{-\varphi} + o(1).
\end{aligned}
\tag{30}
$$

$$
\begin{aligned}
P_u^c &= \frac{\left(\frac{c_1 S g_c(M)}{\gamma} + q\right)^{1-\gamma}}{(M+q)^{1-\gamma} - (q+1)^{1-\gamma}} - \frac{(1-\gamma)\frac{c_1 S g_c(M)}{\gamma}\left(\frac{c_1 S g_c(M)}{\gamma} + q\right)^{-\gamma}}{(M+q)^{1-\gamma} - (q+1)^{1-\gamma}} - \frac{(q+1)^{1-\gamma}}{(M+q)^{1-\gamma} - (q+1)^{1-\gamma}} \\
&= \frac{\left(\frac{c_1 c_3 S M^{\alpha}}{\gamma} + c_4 M^{\alpha}\right)^{1-\gamma}}{(M+q)^{1-\gamma} - (q+1)^{1-\gamma}} - \frac{(1-\gamma)\frac{c_1 c_3 S M^{\alpha}}{\gamma}\left(\frac{c_1 c_3 S M^{\alpha}}{\gamma} + c_4 M^{\alpha}\right)^{-\gamma}}{(M+q)^{1-\gamma} - (q+1)^{1-\gamma}} - \frac{(c_4 M^{\alpha}+1)^{1-\gamma}}{(M+q)^{1-\gamma} - (q+1)^{1-\gamma}} \\
&\overset{(a)}{=} \frac{M^{\alpha(1-\gamma)}}{M^{1-\gamma}}\left[\left(\frac{S c_1 c_3}{\gamma} + c_4\right)^{1-\gamma} - (1-\gamma)\frac{S c_1 c_3}{\gamma}\left(\frac{S c_1 c_3}{\gamma} + c_4\right)^{-\gamma} - (c_4)^{1-\gamma}\right] + o\left(\frac{M^{\alpha(1-\gamma)}}{M^{1-\gamma}}\right) \\
&= M^{(\alpha-1)(1-\gamma)}\left[\left(\frac{S c_1 c_3}{\gamma} + c_4\right)^{-\gamma}(S c_1 c_3 + c_4) - (c_4)^{1-\gamma}\right] + o\left(M^{-\alpha}\right) \\
&\overset{(b)}{=} M^{-\alpha}\left[\left(\frac{S c_1 c_3}{\gamma} + c_4\right)^{-\gamma}(S c_1 c_3 + c_4) - (c_4)^{1-\gamma}\right] + o\left(M^{-\alpha}\right).
\end{aligned}
\tag{33}
$$

in which (d) is because

$$\alpha = \frac{1-\gamma}{2-\gamma} <=> (\alpha-1)\gamma = 2\alpha - 1 <=> \gamma = \frac{2\alpha-1}{\alpha-1}. \tag{35}$$

Now we lower bound $\overline{T}_{\min}$ by using (32) and computing $\mathbb{P}(W > 0)$.[6] We first introduce the definition of self-bounding property and a corresponding Lemma:

Definition [10]: Let $\mathcal{X} \subseteq \mathbb{R}$ and consider a non-negative ν-variate function $g : \mathcal{X} \to [0, \infty)$. We say that g has the self-bounding property if there exists a function $g_i : \mathcal{X}^{\nu-1} \to \mathbb{R}$ such that, for all $x_1, \ldots, x_\nu \subseteq \mathcal{X}^\nu$ and all $i = 1, \ldots, \nu$,

$$
\begin{aligned}
&0 \leq g(x_1, \cdots, x_\nu) - g_i(x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_\nu) \leq 1; \\
&\quad \sum_{i=1}^{\nu}(g(x_1, \cdots, x_\nu) - g_i(x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_\nu)) \\
&\leq g(x_1, \cdots, x_\nu).
\end{aligned}
\tag{36}
$$

Lemma 2 (p. 182, Th. 6.12 in [43]): Consider $\mathcal{X} \subseteq \mathbb{R}$ and the random vector $X = (X_1, \ldots, X_\nu) \in \mathcal{X}^\nu$, where $X_1, \ldots, X_\nu$ are mutually statistically independent. Denote $Y = g(X)$, where $g(.)$ has the self-bounding property. Then, for any $0 < \mu \leq \mathbb{E}[Y]$, we have

$$\mathbb{P}(Y - \mathbb{E}[Y] \leq -\mu) \leq \exp\left(-\frac{\mu^2}{2\mathbb{E}[Y]}\right). \tag{37}$$

[6] Note that the proof technique used for this part is based on the concentration of functions with the self-bounding property and is different from the one in [6].

We observe that the sum function $g(x_1, \dots, x_\nu) = \sum_{i=1}^{\nu} x_i$ has self-bounding property when $x_i, \forall i$, are binary, i.e., $x_i \in \{0, 1\}$. Thus, $W = \sum_{u=1}^{g_c(M)} \mathbf{1}_u$ satisfies the conditions of Lemma 2. By using Lemma 2 and considering $\mu = \mathbb{E}[W]$, we obtain $\mathbb{P}(W \leq 0) \leq \exp\left(-\frac{\mathbb{E}[W]}{2}\right)$. It follows that

$$\mathbb{P}(W > 0) > 1 - \exp\left(-\frac{\mathbb{E}[W]}{2}\right). \tag{38}$$

Using (32) and (38), we thus obtain

$$\begin{aligned}
&\mathbb{E}[\text{number of active cluster}] \\
&\quad \geq \frac{1}{K}\left(\text{number of total clusters} \cdot \mathbb{P}(W > 0)\right) \\
&\quad \geq \frac{N}{K g_c(M)} \exp\left(1 - \exp\left(-\frac{\mathbb{E}[W]}{2}\right)\right) \\
&\quad = \frac{N}{K c_3 M^\alpha}\left(1 - \exp\left(-\frac{\mathbb{E}[W]}{2}\right)\right).
\end{aligned} \tag{39}$$

To compute $\mathbb{E}[W]$, we note that $\mathbb{E}[W] = E\left[\sum_{u=1}^{g_c(M)} \mathbf{1}_u\right] = g_c(M) P_u^c$. Thus,

$$\begin{aligned}
&\mathbb{E}[W] \\
&\quad = c_3 M^\alpha \cdot M^{-\alpha} \cdot \\
&\qquad \left(\left[\left(\frac{S c_1 c_3}{\gamma} + c_4\right)^{-\gamma} (S c_1 c_3 + c_4) - (c_4)^{1-\gamma}\right] + o(1)\right) \\
&\quad = c_3\left[\left(\frac{S c_1 c_3}{\gamma} + c_4\right)^{-\gamma} (S c_1 c_3 + c_4) - (c_4)^{1-\gamma}\right] + o(1).
\end{aligned} \tag{40}$$

By using (31), (39), (40), and that $\overline{T}_{\text{min}} = \frac{1}{N}\overline{T}_{\text{sum}}$, we obtain

$$\begin{aligned}
\overline{T}_{\text{min}} \geq \frac{C}{K}\frac{M^{-\alpha}}{c_3}\Bigg(1 - {}& \\
&\exp\left(\frac{-c_3\left[\left(\frac{S c_1 c_3}{\gamma} + c_4\right)^{-\gamma} (S c_1 c_3 + c_4) - (c_4)^{1-\gamma}\right]}{2}\right)\Bigg) \\
&+ o(M^{-\alpha}).
\end{aligned}$$

Finally, by exploiting the perturbation argument similar to appendix J in [6], we obtain the achievable throughput-outage tradeoff for regime 1 in the theorem as in (41) on the top of next page.

B. Proof of Regime 2

In this section, we consider $g_c(M) = \omega(M^\alpha) < \frac{\gamma M}{c_1 S}$ and $q = \mathcal{O}\left(\frac{S g_c(M)}{\gamma}\right)$. We define $c_5 = \frac{q}{g_c(M)}$. Again by using Corollary 1, we obtain P_u^c as (42) on the top of next page. Then we again use the same approach as used in regime 1 to obtain the lower bound of $\overline{T}_{\text{min}}$. We first compute

$$\begin{aligned}
&E[W] \\
&\quad = g_c(M) P_u^c \\
&\quad = \frac{g_c(M)(g_c(M))^{1-\gamma}\left[\left(\frac{S c_1}{\gamma} + c_5\right)^{-\gamma} (S c_1 + c_5) - (c_5)^{1-\gamma}\right]}{(M + c_5 g_c(M))^{1-\gamma} - (c_5 g_c(M) + 1)^{1-\gamma}}
\end{aligned}$$

$$+ o\left(\frac{(g_c(M))^{2-\gamma}}{M^{1-\gamma}}\right) \overset{(a)}{=} \infty, \tag{43}$$

where (a) is because $g_c(M) < \frac{\gamma M}{c_1 S}$, $q = \mathcal{O}\left(\frac{S g_c(M)}{\gamma}\right)$, and

$$\begin{aligned}
g_c(M)\left(\frac{g_c(M)}{M}\right)^{1-\gamma} &\overset{(b)}{=} g_c(M)\omega\left(\frac{M^{\alpha(1-\gamma)}}{M^{1-\gamma}}\right) \\
&\overset{(c)}{=} g_c(M)\omega(M^{-\alpha}) \overset{(d)}{=} \omega(1) = \infty
\end{aligned} \tag{44}$$

where (b) is because $g_c(M) = \omega(M^\alpha)$; (c) follows the same derivations as in (35); (d) is again because $g_c(M) = \omega(M^\alpha)$. Consequently, we obtain

$$\overline{T}_{\text{min}} \geq \frac{C}{K}\frac{1}{g_c(M)} + o\left(\frac{1}{g_c(M)}\right), \tag{45}$$

since $\exp(-E[W]/2) \to 0$. Again by using a perturbation argument, it follows that

$$T(P_o) = \frac{C}{K}\frac{1}{g_c(M)} + o\left(\frac{1}{g_c(M)}\right), \tag{46}$$

where

$$\begin{aligned}
P_o = 1 - {}& \frac{(g_c(M))^{1-\gamma}}{(M + c_5 g_c(M))^{1-\gamma} - (c_5 g_c(M) + 1)^{1-\gamma}} \\
&\cdot \left[\left(\frac{S c_1}{\gamma} + c_5\right)^{-\gamma} (S c_1 + c_5) - (c_5)^{1-\gamma}\right].
\end{aligned} \tag{47}$$

C. Proof of Regime 3

Finally, we consider $g_c(M) = \frac{\rho M}{c_1 S}$, where $\rho \geq \gamma$ and $q = \mathcal{O}\left(\frac{S g_c(M)}{\gamma}\right)$. Thus, instead of using Corollary 1, Corollary 2 is adopted. By Corollary 2, we obtain

$$\begin{aligned}
P_o \leq {}& \frac{(1-\gamma)e^{-(\rho/c_1 - \gamma)}}{(1+D)^{1-\gamma} - (D)^{1-\gamma}} \cdot \Big[(1+D)^{\frac{\gamma}{S(g_c(M)-1)-1}+1} \\
&- (D)^{\frac{\gamma}{S(g_c(M)-1)-1}+1}\Big]^{-(S(g_c(M)-1)-1)} + o(1).
\end{aligned} \tag{48}$$

To compute the lower bound of $\overline{T}_{\text{min}}$, it is clear that $\mathbb{E}[W] \to \infty$ because both P_u^c and $g_c(M)$ in regime 3 are larger than their counterparts in regime 2. Consequently, we obtain

$$\overline{T}_{\text{min}} \geq \frac{C}{K}\frac{1}{g_c(M)} + o\left(\frac{1}{g_c(M)}\right) = \frac{C}{K}\frac{S c_1}{\rho M} + o\left(\frac{1}{g_c(M)}\right). \tag{49}$$

Again by using a perturbation argument, we obtain the achievable throughput-outage tradeoff in (50) on the top of the next page.

APPENDIX D
PROOF OF THEOREM 3

Observe that P_o goes to 1 when we consider $q = \mathcal{O}\left(\frac{S g_c(M)}{\gamma}\right)$ and $g_c(M) = o(M) < \frac{\gamma M}{c_1 S}$ according to Theorem 2 (regimes 1 and 2). By intuition, it follows that P_o also goes to 1 when we consider $q = \omega\left(\frac{S g_c(M)}{\gamma}\right)$ while $q = \mathcal{O}(M)$ since increasing the value of q degrades the

$$\begin{aligned} T(P_o) &= \frac{C}{K}\frac{M^{-\alpha}}{c_3}\left(1-\exp\left(\frac{-c_3}{2}\left[\left(\frac{Sc_1c_3}{\gamma}+c_4\right)^{-\gamma}(Sc_1c_3+c_4)-(c_4)^{1-\gamma}\right]\right)\right)+o(M^{-\alpha}),\\ P_o &= 1-M^{-\alpha}\left[\left(\frac{Sc_1c_3}{\gamma}+c_4\right)^{-\gamma}(Sc_1c_3+c_4)-(c_4)^{1-\gamma}\right]. \end{aligned} \tag{41}$$

$$\begin{aligned} P_u^c &= \frac{\left(\frac{c_1Sg_c(M)}{\gamma}+q\right)^{1-\gamma}}{(M+q)^{1-\gamma}-(q+1)^{1-\gamma}} - \frac{(1-\gamma)\frac{c_1Sg_c(M)}{\gamma}\left(\frac{c_1Sg_c(M)}{\gamma}+q\right)^{-\gamma}}{(M+q)^{1-\gamma}-(q+1)^{1-\gamma}} - \frac{(q+1)^{1-\gamma}}{(M+q)^{1-\gamma}-(q+1)^{1-\gamma}}\\ &= \frac{\left(\frac{c_1Sg_c(M)}{\gamma}+c_5g_c(M)\right)^{1-\gamma}-(1-\gamma)\frac{c_1Sg_c(M)}{\gamma}\left(\frac{c_1Sg_c(M)}{\gamma}+c_5g_c(M)\right)^{-\gamma}-(c_5g_c(M)+1)^{1-\gamma}}{(M+c_5g_c(M))^{1-\gamma}-(c_5g_c(M)+1)^{1-\gamma}}\\ &= \frac{(g_c(M))^{1-\gamma}\left[\left(\frac{Sc_1}{\gamma}+c_5\right)^{1-\gamma}-(1-\gamma)\frac{Sc_1}{\gamma}\left(\frac{Sc_1}{\gamma}+c_5\right)^{-\gamma}-(c_5)^{1-\gamma}\right]}{(M+c_5g_c(M))^{1-\gamma}-(c_5g_c(M)+1)^{1-\gamma}}\\ &= \frac{(g_c(M))^{1-\gamma}\left[\left(\frac{Sc_1}{\gamma}+c_5\right)^{-\gamma}(Sc_1+c_5)-(c_5)^{1-\gamma}\right]+o\left(\left(\frac{g_c(M)}{M}\right)^{1-\gamma}\right)}{(M+c_5g_c(M))^{1-\gamma}-(c_5g_c(M)+1)^{1-\gamma}}. \end{aligned} \tag{42}$$

$$T(P_o) = \frac{C}{K}\frac{Sc_1}{\rho M}+o\left(\frac{1}{M}\right), P_o = \frac{(1-\gamma)e^{-(\rho/c_1-\gamma)}}{(1+D)^{1-\gamma}-(D)^{1-\gamma}}\left[(1+D)^{\frac{\gamma}{S(g_c(M)-1)-1}+1}-(D)^{\frac{\gamma}{S(g_c(M)-1)-1}+1}\right]^{-(S(g_c(M)-1)-1)} \tag{50}$$

concentration of the popularity distribution which increases the outage. This leads to Theorem 3. Rigorously, observe that

$$P_u^c = \sum_{f=1}^{M} P_r(f)G(f), \tag{51}$$

where $G(f) = \left(1-(1-P_c(f))^{S(g_c(M)-1)}\right)$. Then denote the optimal caching policy for $P_r(f;\gamma,q_1)$ as $P_c^{q_1}(f)$ and the optimal caching policy for $P_r(f;\gamma,q_2)$ as $P_c^{q_2}(f)$, where both $P_c^{q_1}(f)$ and $P_c^{q_2}(f)$ are monotonically decreasing with respect to f (see Appendix A). Considering $q_1 < q_2$, we want to show the following

$$\begin{aligned} &\sum_{f=1}^{M} P_r(f;\gamma,q_1)\left(1-(1-P_c^{q_1}(f))^{S(g_c(M)-1)}\right)\\ &\quad\overset{(a)}{\geq} \sum_{f=1}^{M} P_r(f;\gamma,q_1)\left(1-(1-P_c^{q_2}(f))^{S(g_c(M)-1)}\right)\\ &\quad\overset{(b)}{>} \sum_{f=1}^{M} P_r(f;\gamma,q_2)\left(1-(1-P_c^{q_2}(f))^{S(g_c(M)-1)}\right), \end{aligned} \tag{52}$$

is true. Since (a) is true simply because $P_c^{q_1}(f)$ is the optimal policy for $P_r(f;\gamma,q_1)$, it thus suffices when showing (b) is true.

To show the (b) of (52) is true, we note that when $g < h$ and $\epsilon > 0$,

$$\sum_{f=1}^{M} P_r(f)G(f)+\epsilon G(g)-\epsilon G(h) > \sum_{f=1}^{M} P_r(f)G(f) \tag{53}$$

because $G(f)$ is monotonically decreasing when $P_c(f)$ is monotonically decreasing with respect to f. Eq. (53) indicates that, given the caching policy is monotonically decreasing, when we add ϵ to the popularity with lower index (better rank) by subtracting ϵ from the one with higher index, we can improve P_u^c. Then notice that when $q_1 < q_2$, we obtain:

$$\begin{aligned} P_r(1;\gamma,q_1) &= \frac{(1+q_1)^{-\gamma}}{\sum_{f=1}^{M}(f+q_1)^{-\gamma}} \geq \frac{(1+q_2)^{-\gamma}}{\sum_{f=1}^{M}(f+q_2)^{-\gamma}}\\ &= P_r(1;\gamma,q_2) \end{aligned} \tag{54}$$

and

$$\frac{(f+q_1)^{-\gamma}}{(f+1+q_1)^{-\gamma}} > \frac{(f+q_2)^{-\gamma}}{(f+1+q_2)^{-\gamma}}, \quad f=1,2,\ldots,M, \tag{55}$$

i.e., starting with a larger value, $P_r(f;\gamma,q_1)$ decreases faster than $P_r(f;\gamma,q_2)$ with respect to f. By using (53), (54), and (55), we can then obtain

$$\begin{aligned} &\sum_{f=1}^{M} P_r(f;\gamma,q_1)\left(1-(1-P_c^{q_2}(f))^{S(g_c(M)-1)}\right)\\ &\quad-\sum_{f=1}^{M} P_r(f;\gamma,q_2)\left(1-(1-P_c^{q_2}(f))^{S(g_c(M)-1)}\right) > 0, \end{aligned}$$

which proves the (b) of (52) is true.

Appendix E
Proof of Theorem 4

We consider $g_c(M) = o(M) \leq N$ and $q = \mathcal{O}\left(\frac{Sg_c(M)}{\gamma}\right)$. Since these regimes imply $g_c(M) < \frac{\gamma M}{c_1 S}$, we should apply

$$
\begin{aligned}
P_u^c &= \frac{\left(\frac{c_1 S g_c(M)}{\gamma}+q\right)^{1-\gamma}}{(M+q)^{1-\gamma}-(q+1)^{1-\gamma}} - \frac{(1-\gamma)\frac{c_1 S g_c(M)}{\gamma}\left(\frac{c_1 S g_c(M)}{\gamma}+q\right)^{-\gamma}}{(M+q)^{1-\gamma}-(q+1)^{1-\gamma}} - \frac{(q+1)^{1-\gamma}}{(M+q)^{1-\gamma}-(q+1)^{1-\gamma}} \\
&= \frac{(c_6 g_c(M)+1)^{1-\gamma}}{(c_6 g_c(M)+1)^{1-\gamma}-(M+c_6 g_c(M))^{1-\gamma}} - \frac{\left(\frac{c_1 S g_c(M)}{\gamma}+c_6 g_c(M)\right)^{1-\gamma}}{(c_6 g_c(M)+1)^{1-\gamma}-(M+c_6 g_c(M))^{1-\gamma}} \\
&\quad - \frac{(\gamma-1)\frac{c_1 S g_c(M)}{\gamma}\left(\frac{c_1 S g_c(M)}{\gamma}+c_6 g_c(M)\right)^{-\gamma}}{(c_6 g_c(M)+1)^{1-\gamma}-(M+c_6 g_c(M))^{1-\gamma}} \\
&\overset{(a)}{\geq} \frac{(c_6 g_c(M)+1)^{1-\gamma}}{(c_6 g_c(M)+1)^{1-\gamma}} - \frac{\left(\frac{c_1 S g_c(M)}{\gamma}+c_6 g_c(M)\right)^{1-\gamma}}{(c_6 g_c(M)+1)^{1-\gamma}} - \frac{(\gamma-1)\frac{c_1 S g_c(M)}{\gamma}\left(\frac{c_1 S g_c(M)}{\gamma}+c_6 g_c(M)\right)^{-\gamma}}{(c_6 g_c(M)+1)^{1-\gamma}} \\
&= 1-\left(\frac{c_6 g_c(M)+1}{\frac{c_1 S g_c(M)}{\gamma}+c_6 g_c(M)}\right)^{\gamma-1} - \frac{(\gamma-1)(c_6 g_c(M)+1)^{\gamma-1}}{\left(\frac{c_1 S g_c(M)}{\gamma}+c_6 g_c(M)\right)^{\gamma}\left(\frac{c_1 S g_c(M)}{\gamma}\right)^{-1}} \\
&= 1-\left(\frac{c_6 g_c(M)+1}{\frac{c_1 S g_c(M)}{\gamma}+c_6 g_c(M)}\right)^{\gamma}\left(\left(\frac{c_6 g_c(M)+1}{\frac{c_1 S g_c(M)}{\gamma}+c_6 g_c(M)}\right)^{-1}+(\gamma-1)\left(\frac{c_6 g_c(M)+1}{\frac{c_1 S g_c(M)}{\gamma}}\right)^{-1}\right) \\
&= 1-\left(\frac{c_6 g_c(M)+1}{\frac{c_1 S g_c(M)}{\gamma}+c_6 g_c(M)}\right)^{\gamma}\frac{c_6 g_c(M)+c_1 S g_c(M)}{c_6 g_c(M)+1} \\
&= 1-\left(\frac{c_6}{\frac{S c_1}{\gamma}+c_6}\right)^{\gamma}\frac{c_6+S c_1}{c_6}-o(1) = 1-(c_6)^{\gamma-1}\frac{S c_1+c_6}{\left(\frac{S c_1}{\gamma}+c_6\right)^{\gamma}}-o(1).
\end{aligned} \tag{56}
$$

Corollary 1. We define $c_6 = \frac{q}{g_c(M)}$. When $\gamma > 1$, we obtain P_u^c in (56) on the top of this page, where (a) is because

$$(1+c_6 g_c(M))^{1-\gamma} > (1+c_6 g_c(M))^{1-\gamma} - (M+c_6 g_c(M))^{1-\gamma} > 0.$$

Then notice that $\mathbb{E}[W] = g_c(M) P_u^c \to \infty$ since $g_c(M) \to \infty$ and $c_6 = \mathcal{O}(1)$. Consequently, $P(W > 0) \to 1$ by Lemma 2 (see Appendix C.A). It follows that

$$\overline{T}_{\min} \geq \frac{C}{K}\frac{1}{g_c(M)} + o\left(\frac{1}{g_c(M)}\right).$$

Finally, again by the fact that $P_o = 1 - P_u^c$ and the perturbation argument, we obtain Theorem 4.

REFERENCES

[1] M.-C. Lee, M. Ji, A. F. Molisch, and N. Sastry, "Performance of caching-based D2D video distribution with measured popularity distributions," Dec. 2019, *arXiv:1806.05380*. [Online]. Available: https://arxiv.org/abs/1806.05380

[2] *Cisco Virtual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021*, Cisco Syst., San Jose, CA, USA, 2018.

[3] J. G. Andrews, "Seven ways that HetNets are a cellular paradigm shift," *IEEE Commun. Mag.*, vol. 51, no. 3, pp. 136–144, Mar. 2013.

[4] T. S. Rappaport *et al.*, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, May 2013.

[5] 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 Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.

[6] M. Ji, G. Caire, 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, Oct. 2015.

[7] N. Golrezaei, P. Mansourifard, A. F. Molisch, and A. G. Dimakis, "Base-station assisted device-to-device communications for high-throughput wireless video networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3665–3676, Jul. 2014.

[8] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[9] U. Niesen and M. A. Maddah-Ali, "Coded caching with nonuniform demands," *IEEE Trans. Inf. Theory*, vol. 63, no. 2, pp. 1146–1158, Feb. 2017.

[10] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, "Order-optimal rate of caching and coded multicasting with random demands," *IEEE Trans. Inf. Theory*, vol. 63, no. 6, pp. 3923–3949, Jun. 2017.

[11] D. Karamshuk, N. Sastry, M. Al-Bassam, A. Secker, and J. Chandaria, "Take-away TV: Recharging work commutes with predictive preloading of catch-up TV content," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2091–2101, Aug. 2016.

[12] B. N. Bharath, K. G. Nagananda, and H. V. Poor, "A learning-based approach to caching in heterogenous small cell networks," *IEEE Trans. Commun.*, vol. 64, no. 4, pp. 1674–1686, Apr. 2016.

[13] S. H. Chae and W. Choi, "Caching placement in stochastic wireless caching helper networks: Channel selection diversity via caching," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6626–6637, Oct. 2016.

[14] K. Doppler, M. Rinne, C. Wijting, C. B. Ribeiro, and K. Hugl, "Device-to-device communication as an underlay to LTE-advanced networks," *IEEE Commun. Mag.*, vol. 47, no. 12, pp. 42–49, Dec. 2009.

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

[16] N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Wireless video content delivery through distributed caching and peer-to-peer gossiping," in *Proc. IEEE ASILOMAR*, Nov. 2011, pp. 1177–1180.

[17] X. Song, Y. Geng, X. Meng, J. Liu, W. Lei, and Y. Wen, "Cache-enabled device to device networks with contention-based multimedia delivery," *IEEE Access*, vol. 5, pp. 3228–3239, 2017.

[18] D. Malak, M. Al-Shalash, and J. G. Andrews, "Spatially correlated content caching for device-to-device communications," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 56–70, Jan. 2018.

[19] Z. Chen, N. Pappas, and M. Kountouris, "Probabilistic caching in wireless D2D networks: Cache hit optimal versus throughput optimal," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 584–587, Mar. 2017.

[20] B. Chen, C. Yang, and G. Wang, "High-throughput opportunistic cooperative device-to-device communications with caching," *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 7527–7539, Aug. 2017.

[21] M.-C. Lee and A. F. Molisch, "Caching policy and cooperation distance design for base station-assisted wireless D2D caching networks: Throughput and energy efficiency optimization and tradeoff," *IEEE Trans. Wireless Commun.*, vol. 17, no. 11, pp. 7500–7514, Nov. 2018.

[22] B. Chen, C. Yang, and A. F. Molisch, "Cache-enabled device-to-device communications: Offloading gain and energy cost," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4519–4536, Jul. 2017.

[23] T. Deng, G. Ahani, P. Fan, and D. Yuan, "Cost-optimal caching for D2D networks with user mobility: Modeling, analysis, and computational approaches," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3082–3094, May 2018.

[24] Y. Wang, X. Tao, X. Zhang, and Y. Gu, "Cooperative caching placement in cache-enabled D2D underlaid cellular network," *IEEE Commun. Lett.*, vol. 21, no. 5, pp. 1151–1154, May 2017.

[25] L. Pei, Z. Yang, C. Pan, W. Huang, and M. Chen, "Joint bandwidth, caching and association optimization for D2D assisted wireless networks," in *Proc. IEEE INFOCOM*, Apr. 2018, pp. 505–510.

[26] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 849–869, Feb. 2016.

[27] L. Zhang, M. Xiao, G. Wu, and S. Li, "Efficient scheduling and power allocation for d2d-assisted wireless caching networks," *IEEE J. Sel. Areas Commun.*, vol. 64, no. 6, pp. 2438–2452, Jun. 2016.

[28] R. Wang, J. Zhang, S. H. Song, and K. B. Letaief, "Mobility-aware caching in D2D networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5001–5015, Aug. 2017.

[29] D. Malak, M. Al-Shalash, and J. G. Andrews, "Optimizing content caching to maximize the density of successful receptions in device-to-device networking," *IEEE Trans. Commun.*, vol. 64, no. 10, pp. 4365–4380, Oct. 2016.

[30] M. Naslcheraghi, M. Afshang, and H. S. Dhillon, "Modeling and performance analysis of full-duplex communications in cache-enabled D2D networks," in *Proc. IEEE ICC*, May 2018, pp. 1–6.

[31] S. Vuppala, T. X. Vu, S. Gautam, S. Chatzinotas, and B. Ottersten, "Cache-aided millimeter wave ad hoc networks with contention-based content delivery," *IEEE Trans. Commun.*, vol. 66, no. 8, pp. 3540–3554, Aug. 2018.

[32] S. Gautam, X. V. Thang, S. Chatzinotas, and B. Ottersten, "Cache-aided simultaneous wireless information and power transfer (SWIPT) with relay selection," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 1, pp. 187–201, Jan. 2019.

[33] M.-C. Lee, H. Feng, and A. F. Molisch, "Dynamic caching content replacement in base station assisted wireless D2D caching networks," *IEEE Trans. Commun.*, to be published.

[34] M. Zink, K. Suh, Y. Gu, and J. Kurose, "Watch global, cache local: YouTube network traffic at a campus network: Measurements and implications," in *Proc. SPIE/ACM MMCN*, 2008, Art. no. 681805. [Online]. Available: https://scholarworks.umass.edu/cs_faculty_pubs/177

[35] X. Hu and A. Striegel, "Redundancy elimination might be overrated: A quantitative study on wireless traffic," in *Proc. IEEE INFOCOM WKSHPS*, May 2017, pp. 754–759.

[36] 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, Dec. 2008.

[37] G. Nencioni *et al.*, "SCORE: Exploiting global broadcasts to create offline personal channels for on-demand access," *IEEE/ACM Trans. Netw.*, vol. 24, no. 4, pp. 2429–2442, Aug. 2016.

[38] M.-C. Lee, A. F. Molisch, N. Sastry, and A. Raman, "Individual preference probability modeling and parameterization for video content in wireless caching networks," *IEEE/ACM Trans. Netw.*, vol. 27, no. 2, pp. 676–690, Apr. 2019.

[39] K. P. Gummadi, R. J. Dunn, S. Saroiu, S. D. Gribble, H. M. Levy, and J. Zahorjan, "Measurement, modeling, and analysis of a peer-to-peer file-sharing workload," in *Proc. ACM SOSP*, Oct. 2003, pp. 314–329.

[40] A. F. Molisch, *Wireless Communication*, 2nd ed. Hoboken, NJ, USA: Wiley, 2011.

[41] L.-S. Juhn and L.-M. Tseng, "Harmonic broadcasting for video-on-demand service," *IEEE Trans. Broadcast.*, vol. 43, no. 3, pp. 268–271, Sep. 1997.

[42] J. Karedal, A. J. Johansson, F. Tufvesson, and A. F. Molisch, "A measurement-based fading model for wireless personal area networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 11, pp. 4575–4585, Nov. 2008.

[43] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. London, U.K.: Oxford Univ. Press, 2013.