The Throughput-Outage Tradeoff of Wireless One-Hop Caching Networks

Mingyue Ji, *Student Member, IEEE*, Giuseppe Caire, *Fellow, IEEE*,
and Andreas F. Molisch, *Fellow, IEEE*

The authors are with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089, USA. (e-mail: {mingyuej, caire, molisch}@usc.edu). A short version of this work was presented at the 2013 IEEE International Symposium on Information Theory, ISIT 2013, Istanbul, July 7-11, 2013.

This work was supported by the Intel VAWN (Video Aware Wireless Networks) program and the National Science Foundation under grants CCF-1423140 and CIF-1161801.

Abstract

We consider a wireless device-to-device (D2D) network where the nodes have pre-cached information from a library of available files. Nodes request files at random. If the requested file is not in the on-board cache, then it is downloaded from some neighboring node via one-hop "local" communication. An outage event occurs when a requested file is not found in the neighborhood of the requesting node, or if the network admission control policy decides not to serve the request. We characterize the optimal throughput-outage tradeoff in terms of tight scaling laws for various regimes of the system parameters, when both the number of nodes and the number of files in the library grow to infinity. Our analysis is based on Gupta and Kumar *protocol model* for the underlying D2D wireless network, widely used in the literature on capacity scaling laws of wireless networks without caching. Our results show that the combination of D2D spectrum reuse and caching at the user nodes yields a per-user throughput independent of the number of users, for any fixed outage probability in $(0, 1)$. This implies that the D2D caching network is "scalable": even though the number of users increases, each user achieves constant throughput. This behavior is very different from the classical Gupta and Kumar result on ad-hoc wireless networks, for which the per-user throughput vanishes as the number of users increases. Furthermore, we show that the user throughput is directly proportional to the fraction of cached information over the whole file library size. Therefore, we can conclude that D2D caching networks can turn "memory" into "bandwidth" (i.e., doubling the on-board cache memory on the user devices yields a 100% increase of the user throughout).

Index Terms

Throughput-outage tradeoff, scaling laws, caching wireless networks, device-to-device communications.

I. Introduction

Data traffic generated by wireless and mobile devices is predicted to increase by something between one and two orders of magnitude [1] in the next five years, mainly due to wireless video streaming. Traditional methods for increasing the area spectral efficiency, such as using more spectrum and/or deploying more base stations, are either insufficient to provide the necessary wireless throughput increase, or are too expensive. Thus, exploring alternative strategies that leverage different and cheaper network resources is of great practical and theoretical interest.

The bulk of wireless video traffic is due to asynchronous *video on demand*, where users request video files from some library (e.g., iTunes, Netflix, Hulu or Amazon Prime) at arbitrary times. This type of traffic differs significantly from *live streaming*. The latter is essentially a lossy multicasting problem, for which the broadcast nature of the wireless channel can be naturally exploited (see for example [2]–[6]). The theoretical foundation of schemes for live streaming relies on well-known information theoretic settings for one-to-many transmission of a common message with possible refinement information, such as successive refinement [7]–[9] or multiple description coding [10]–[12].

In contrast, the *asynchronous* nature of video on demand prevents from taking advantage of multicasting, despite the significant overlap of the requests (people wish to watch a few very popular files). Hence, even though users keep requesting the same few popular files, the asynchronism of their requests is large with respect to the duration of the video itself, such that the probability that a single transmission from the base station is useful for more than one user is *essentially zero*. Due to this reason, current practical implementation of video on demand over wireless networks is handled at the application layer, requiring a dedicated data connection (typically TCP/IP) between each client (user) and the server (base station), for each streaming user, as if users were requesting independent information.

One of the most promising approaches to take advantage of the inherent *asynchronous content reuse* is *caching*, widely used in content distribution networks over the (wired) Internet [13]. In [14], [15], the idea of deploying dedicated "helper nodes" with large caches, that can be refreshed via wireless at the cellular network off-peak time, was proposed as a cost-effective alternative to providing large capacity wired backhaul to a network of densely deployed small cells. An even more radical view considers caching directly at the wireless users, exploiting the fact that modern devices have tens and even hundreds of GBytes of largely under-utilized storage space, which represents an enormous, cheap and yet untapped network resource.

Recently, a *coded multicasting* scheme exploiting caching at the user nodes was proposed in [16]. In this

scheme, the files in the library are divided in blocks (packets) and users cache carefully designed subsets of such packets. Then, for given set of user demands, the base station sends to all users (multicasting) a common codeword formed by a sequence of packets obtained as linear combinations of the original file packets. As noticed in [16], coded multicasting can handle any form of asynchronism by suitable sub-packetization. Hence, the scheme is able to create multicasting opportunities through coding, exploiting the overlap of demands while eliminating the asynchronism problem. For the case of arbitrary (adversarial) demands, the coded multicasting scheme of [16] is shown to perform within a small gap, independent of the number of users, of the cache size and of the library size, from the cut-set bound of the underlying compound channel.[1] However, the scheme has some significant drawbacks that makes it not easy to be implemented in practice: 1) the construction of the caches is combinatorial and the sub-packetization explodes exponentially with the library size and number of users; 2) changing even a single file in the library requires a significant reconfiguration of the user caches, making the cache update difficult. In [17], similar near-optimal performance of coded caching is shown to be achieved also through a *random caching scheme*, where each user caches a random selection of bits from each file in the library. In this case, though, the combinatorial complexity of the coded caching scheme is transferred from the caching phase to the (coded) delivery phase, where the construction of the multicast codeword requires solving multiple clique cover problems with fixed clique size (known to be NP-complete), for which a greedy algorithm is shown to be efficient.

Our contributions: In this paper, we focus on an alternative approach that involves random independent caching at the user nodes and device-to-device (D2D) communication. Instead of creating multicasting opportunities by coding, we exploit the spatial reuse provided by concurrent multiple short-range D2D transmissions. Inspired by the current standardization of a D2D mode for LTE (the 4-th generation of cellular systems) [18], we restrict to one-hop communication. Under such assumption, requiring that all users must be served for any request configuration is too constraining. Therefore, we introduce the possibility of outages, i.e., that some requests are not served, because of some network admission control policy (to be discussed in details later on). For the system described in Section II, we define the throughput-outage region and obtain achievability and converses that are sufficiently tight to characterize the throughput-outage scaling laws within a small gap of the *constants of the leading term*. Furthermore, our analysis shows very good agreement with with finite-dimensional simulation results.

In the relevant regime of small outage probability, the throughput of the D2D one-hop caching network

[1]The compound nature of this model is due to the fact that the scheme handles adversarial demands.

behaves in the same near-optimal way as the throughput of coded multicasting [16], [17], while the system architecture is significantly more straightforward for a practical implementation. In particular, for fixed cache size M, as the number of users n and the number of files m become large with $nM \gg m$, the throughput of the D2D one-hop caching network grows linearly with M, and it is inversely proportional to m, but it is independent of n. Hence, D2D one-hop caching networks are very attractive to handle situations where a relatively small library of popular files (e.g., the 500 most popular movies and TV shows of the week) is requested by a large number of users (e.g., 10,000 users per km^2 in a typical urban environment). In this regime, the proposed system is able to efficiently turn memory into bandwidth, in the sense that the per-user throughput increases proportionally to the cache capacity of the user devices. We believe that this conclusion is important for the design of future wireless systems, since bandwidth is a much more scarce and expensive resource than storage capacity.

Related literature: The analysis of the capacity scaling laws[2] for large D2D (or "ad-hoc") wireless networks has been the subject of a vast body of literature. Gupta and Kumar [19] proposed a network model where n are randomly placed on some planar region and communicate through multiple hops. They characterized the transport capacity scaling as $n \to \infty$, under the same *protocol model* considered in our paper (see Section II). For random assignment of source-destination pairs, [19] showed that the per-link capacity must vanish as $O(\frac{1}{\sqrt{n}})$. In addition, a multi-hop relaying scheme was shown to achieve throughput scaling $\Omega(\frac{1}{\sqrt{n \log n}})$. The same results were confirmed, using a somehow simpler and more general analysis technique, in [20]. The multicast capacity of large wireless networks has been investigated in [21], [22]. Finally, Franceschetti, Dousse, Tse and Thiran [23] closed the $\sqrt{\log(n)}$ gap between upper bound and achievability in [19], [20] by creating an almost deterministically placed grid sub-network with vertical and horizontal "highways" that relay messages with very short hops. The existence of such grid subnetwork is guaranteed with high probability by percolation theory.

Given the fact that randomly placed nodes yield the same scaling laws of nodes placed on a deterministic squared grid, in this work we assume a grid network from the start. This allows to focus on the essential aspect of caching at the nodes, and avoid the analytical complication of randomly placed nodes. The same approach is taken in [24], where *multi-hop* D2D communication is considered under the protocol model for a network of nodes placed deterministically on a squared grid. If the aggregate distributed

[2]Scaling law order notation: given two functions f and g, we say that: 1) $f(n) = O(g(n))$ 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 $g(n) = O(f(n))$. 4) $f(n) = \omega(g(n))$ if $g(n) = o(f(n))$. 5) $f(n) = \Theta(g(n))$ if $f(n) = O(g(n))$ and $g(n) = O(f(n))$.

storage space in the network is larger than the total size of the library, multi-hop guarantees that all user requests can be served by the network. Under the same assumption made here of the user requests distribution, [24] finds a deterministic replication caching scheme and a multi-hop routing scheme that achieves order-optimal average throughput. Besides the consideration of multihop and single hop, there are several other key differences between our work and [24]. First, [24] considers a deterministic caching placement approach, which depends on the files popularity distribution. This approach is not robust when users move between cells. In contrast, mobility is easily handled by our scheme which is based on independent random caching. Next, in [24], the transmission range is fixed, where each node can only reach its four neighbors. Besides the deterministic caching placement, the key aspect of the problem is the design of the routing protocol and analyze the traffic through the *bottleneck link* of the network. Our work focuses on determining the transmission range within which nodes can access their neighbors caches in one hop. This, in turn, determines the point of the throughput-outage tradeoff at which the system operates. Finally, [24] only gives the order of the throughput as the number of users n goes to infinity, but does not characterizes the multiplicative constant of the throughput leading term in the scaling law. Therefore, it is difficult to understand in which regime of (large but finite n) the scaling laws become relevant. In passing, we notice that this is a common problem in several studies focused on scaling laws of large wireless networks. In our case, we provide outer bounds and inner (achievable) bounds to the throughput-outage tradeoff, which are tight enough to characterize also the constants of the leading terms, within a bounded gap. In particular, the analysis of our achievability scheme matches well with finite-dimensional simulations.

Preliminary work of the present paper is given in [25], where only the sum throughput was considered irrespectively of user outage probability. The analysis in [25] considers a heuristic random caching policy, while here we find the optimal random caching distribution. More importantly, the total sum throughput is not a sufficient characterization of the performance of D2D one-hop caching networks: in certain regimes of the number of users and file library size, it can be shown that in order to achieve a large sum throughput only a small portion of the users should be served, while leaving the majority of the users in outage. In contrast, the throughput-outage tradeoff region considered here is able to capture the notion of *fairness*, since it focuses on the minimum per-user average throughput and on the fraction of users which are denied service.

The paper is organized as follows. Section II introduces the network model and the precise problem formulation of the throughput-outage tradeoff in D2D one-hop caching networks. The main results on the outer bound and achievability of the throughput-outage tradeoff are presented in Sections III and

IV, respectively. In Section V we presents some concluding remarks. All proofs are relegated in the Appendices, in order to maintain the flow of exposition.

II. Network Model and Problem Formulation

We consider a network deployed over a unit-area squared region and formed by n nodes $\mathcal{U} = \{1, \ldots, n\}$ placed on a regular grid with minimum node distance $1/\sqrt{n}$ (see Fig. 2). Each user $u \in \mathcal{U}$ makes a request to a file $\mathsf{f}_u \in \mathcal{F} = \{1, \ldots, m\}$ in an i.i.d. manner, according to a given request probability mass function $\{P_r(f) : f \in \mathcal{F}\}$. In order to model the asynchronous content reuse and forbid any form of "for-free" multicasting by "overhearing", we consider the following theoretical model. We assume that each file is formed by a sequence of L packets. Each user demand corresponds to a file index $f \in \mathcal{F}$ and a segment of $L' < L$ consecutive packets, starting at some initial index ℓ, uniformly and independently distributed over $\{1, \ldots, L - L' + 1\}$. The packets of the requested segment are downloaded sequentially. We measure the cache size in files, and in order to compute the system performance we consider first the limit for large file size ($L \to \infty$) with L' finite, and then study the system scaling laws for $n, m \to \infty$. Hence, the probability that users request overlapping segments vanishes for $L \to \infty$ for any finite n, m, L', thus preventing the trivial use of naive multicasting (i.e., overhearing common messages). In contrast, the probability that two users request segments of the *same file* depends on the library size m and on the request distribution P_r. We hasten to say that this model is just a way to express in precise mathematical terms the notion of *asynchronous content reuse*, such that the overlap of the demands and the overlap of concurrent transmissions are decoupled.[3] Fig. 1 shows qualitatively our model assumptions.

In our system, D2D communication obeys the following *protocol model* [19]: a node u can receive successfully a packet from node v if $d(u, v) \leq R$ and if no other node v' at distance $d(u, v') < (1 + \Delta)R$ is transmitting. The transmission range R is a design parameter that can be set as a function of m and n. We consider the following definitions:

Definition 1: **(Network)** A network if formed by a set of user nodes $\mathcal{U}$ and a set of files $\mathcal{F} = \{1, \ldots, m\}$. Nodes in $\mathcal{U}$ are placed in the two-dimensional unit-square region, and their transmissions obeys the protocol model. In general, all $n(n-1)$ directed links between user nodes, subject to the protocol model, define an interference (conflict) graph. Only the links in an independent set in the interference graph can be active simultaneously. $\Diamond$

[3] As a side note, we observe also that the segmentation of large files into smaller packets (or "chunks") to be sequentially downloaded is consistent with current video streaming protocols such as DASH [26]–[29].

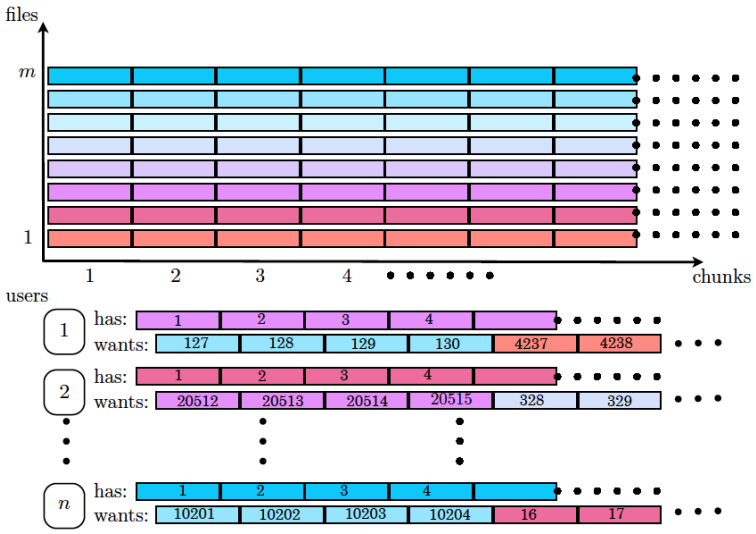


Fig. 1. Qualitative representation of our system assumptions: each user caches an entire file, formed by a very large number of packets L. Then, users place random requests of segments of L' packets from files of the library, starting at random initial points. In the figure, we have $L' = 4$.

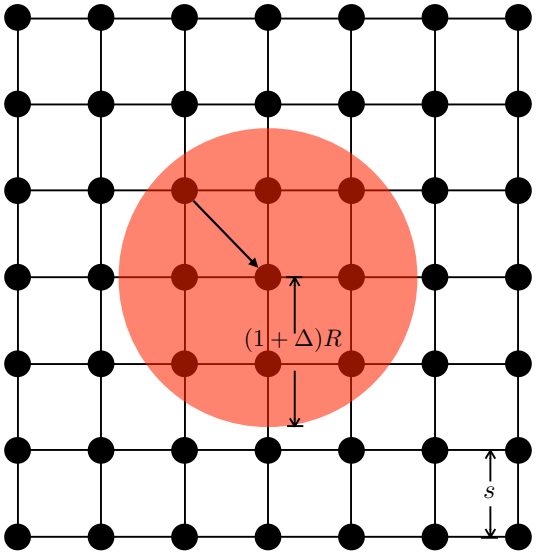


Fig. 2. Grid network with $n = 49$ nodes (black circles) with minimum separation $s = \frac{1}{\sqrt{n}}$. The red area is the disk where the protocol model imposes no other concurrent transmission. R is the case transmission range and Δ is the interference parameter, such that the forbidden disk around the receiver has radius $(1 + \Delta)R$.

Definition 2: **(Cache placement)** The cache placement Π_c is a rule to assign files from the library $\mathcal{F}$ to the user nodes $\mathcal{U}$ with "replacement" (i.e., with possible replication). Let $\mathsf{G} = \{\mathcal{U}, \mathcal{F}, \mathcal{E}\}$ be a bipartite graph with "left" nodes $\mathcal{U}$, "right" nodes $\mathcal{F}$ and edges $\mathcal{E}$ such that $(u, f) \in \mathcal{E}$ indicates that file f is assigned to the cache of user u. A bi-partite cache placement graph G is feasible if the degree of each user node is not larger than the maximum user cache size equal to M files. Let $\mathcal{G}$ denote the set of all feasible bi-partite graphs G. Then, Π_c is a probability mass function over $\mathcal{G}$, i.e., a particular cache placement $\mathsf{G} \in \mathcal{G}$ is assigned with probability $\Pi_c(\mathsf{G})$. $\lozenge$

Notice that deterministic cache placements are special cases, corresponding to a single probability mass equal to 1 on the desired assignment G. In contrast, we will be interested in "decentralized" random caching placements constructed as follows: each user node $u \in \mathcal{U}$ selects its cache content in an i.i.d. manner, by independently generating at random M indices $f_{u,1}, \ldots, f_{u,M}$ according to the same caching probability mass function $\{P_c(f) : f \in \mathcal{F}\}$.

Definition 3: **(Random requests)** At each request time (integer multiples of L'), each user $u \in \mathcal{U}$ makes a request to a segment of length L' of chunks from file $\mathsf{f}_u \in \mathcal{F}$, selected independently with probability P_r. The vector of current requests f is a random vector taking on values in $\mathcal{F}^n$, with product joint probability mass function $\mathbb{P}(\mathsf{f} = (f_1, \ldots, f_n)) = \prod_{i=1}^n P_r(f_i)$. $\lozenge$

In this paper, we assume that P_r is a Zipf distribution with parameter $0 < \gamma < 1$ [30], i.e., $P_r(f) = \frac{f^{-\gamma}}{H(\gamma,1,m)}$ for $f = 1, \ldots, m$, and $H(\gamma, x, y) \triangleq \sum_{i=x}^{y} \frac{1}{i^\gamma}$.

Definition 4: **(Transmission policy)** The transmission policy Π_t is a rule to activate the D2D links in the network. Let $\mathcal{L}$ denote the set of all directed links. Let $\mathcal{A} \subseteq 2^{\mathcal{L}}$ denote the set of all feasible subsets of links (this is a subset of the power set of $\mathcal{L}$, formed by all independent sets in the network interference graph). Let $\mathsf{A} \subset \mathcal{A}$ denote a feasible set of simultaneously active links according to the protocol model. Then, Π_t is a conditional probability mass function over $\mathcal{A}$ given f (requests) and G (cache placement), assigning probability $\Pi_t(\mathsf{A}|\mathsf{f}, \mathsf{G})$ to $\mathsf{A} \in \mathcal{A}$. $\lozenge$

We may think of Π_t as a way of scheduling simultaneously compatible sets of links (subject to the protocol model). Modeling the scheduling policy in a probabilistic manner allows the analytical convenience of defining the average per-user throughput (see below) as an ensemble average. As a matter of fact, deterministic link activation rules can be included by defining the average throughput as a time-average. For example, a bounded deterministic delay per user can be guaranteed by activating groups of compatible links (forming a maximal independent set in the network interference graph) in a deterministic round-robin sequence, such that each user is served with a deterministic delay.

Definition 5: **(Useful received bits per slot)** For given P_r, Π_c and Π_t, and user $u \in \mathcal{U}$, we define the

random variable T_u as the number of useful received information bits per slot unit time by user u at a given scheduling time (irrelevant because of stationarity). This is given by

$$T_u = \sum_{v:(u,v)\in \mathsf{A}} c_{u,v} 1\{\mathsf{f}_u \in \mathsf{G}(v)\} \tag{1}$$

where f_u denotes the file requested by user node u, $c_{u,v}$ denotes the rate of the link (u,v), and $\mathsf{G}(v)$ denotes the content of the cache of node v, i.e., the neighborhood of node v in the cache placement graph G. $\lozenge$

Consistently with the protocol model, $c_{u,v}$ depends only on the active link $(u,v) \in \mathsf{A}$ and not on the whole set of active links A. Furthermore, we shall obtain our results under the simplifying assumption (usually made under the protocol model [19]) that $c_{u,v} = C$ for all $(u,v) \in \mathsf{A}$. The indicator function $1\{\mathsf{f}_u \in \mathsf{G}(v)\}$ expresses the fact that only the bits relative to the file f_u requested by user u are "useful" and count towards the throughput. Obviously, scheduling links (u,v) for which $\mathsf{f}_u \notin \mathsf{G}(v)$ is useless for the sake of the throughput defined above. Hence, we can restrict our transmission policies to those activating only links (u,v) for which $\mathsf{f}_u \in \mathsf{G}(v)$. These links are referred to as "potential links", i.e., links potentially carrying useful data. Potential links included in A are "active links", at the given scheduling slot.

The average throughput for user $u \in \mathcal{U}$ is given by $\overline{T}_u = \mathbb{E}[T_u]$, where expectation is with respect to the random triple $(\mathsf{f},\mathsf{G},\mathsf{A}) \sim \prod_{i=1}^{n} P_r(f_i)\Pi_c(\mathsf{G})\Pi_t(\mathsf{A}|\mathsf{f},\mathsf{G})$. We say that user u is in outage if $\mathbb{E}[T_u|\mathsf{f},\mathsf{G}] = 0$. This condition captures the event that no link (u,v) with $\mathsf{f}_u \in \mathsf{G}(v)$ is scheduled with positive probability, for given requests vector f and cache placement G. In other words, a user u for which $\mathbb{E}[T_u|\mathsf{f},\mathsf{G}] = 0$ experiences a "long" lack of service (zero rate), as far as the cache placement is G and the request vector is f.

Definition 6: **(Number of nodes in outage)** The number of nodes in outage is given by

$$\mathsf{N}_o = \sum_{u\in\mathcal{U}} 1\{\mathbb{E}[T_u|\mathsf{f},\mathsf{G}] = 0\}. \tag{2}$$

Notice that N_o is a random variable, function of f and G. $\lozenge$

Definition 7: **(Average outage probability)** The average (across the users) outage probability is given by

$$p_o = \frac{1}{n}\mathbb{E}[\mathsf{N}_o] = \frac{1}{n}\sum_{u\in\mathcal{U}} \mathbb{P}\left(\mathbb{E}[T_u|\mathsf{f},\mathsf{G}] = 0\right). \tag{3}$$

$\lozenge$

In this work we focus on max-min fairness, i.e., we express the throughput-outage tradeoff in terms of the minimum average user throughput, defined as

$$\overline{T}_{\min} = \min_{u \in \mathcal{U}} \left\{ \overline{T}_u \right\}. \tag{4}$$

Notice that that the max-min fairness criterion in our setting is *essential* to make the outage probability p_o defined in (3) meaningful. In fact, for $0 \leq p_o' < p_o \leq 1$, consider a system that achieves outage probability p_o by serving only a fraction $1-\lambda$ of users with outage probability $p_o' = \frac{p_o - \lambda}{1-\lambda}$, while leaving the remaining fraction λ of users permanently idle. In this case, we have $\overline{T}_{\min} = 0$ since there are $\lambda n > 0$ users with identically zero throughput. Hence, a system that permanently excludes some users in favor of others is certainly not optimal in terms of the throughput-outage tradeoff as defined below:

Definition 8: **(Throughput-Outage Tradeoff)** For a given network and request probability distribution P_r, an throughput-outage pair (T, p) is *achievable* if there exists a cache placement Π_c and a transmission policy Π_t with outage probability $p_o \leq p$ and minimum per-user average throughput $\overline{T}_{\min} \geq T$. The throughput-outage achievable region $\mathcal{T}$ is the closure of all achievable throughput-outage pairs (T, p). In particular, we let $T^*(p) = \sup\{T : (T, p) \in \mathcal{T}\}$. $\lozenge$

Notice that $T^*(p)$ is the result of the following optimization problem:

$$\begin{aligned} &\text{maximize} && \overline{T}_{\min} \\ &\text{subject to} && p_o \leq p, \end{aligned} \tag{5}$$

where the maximization is with respect to the cache placement and transmission policies Π_c, Π_t. Hence, it is immediate to see that $T^*(p)$ is non-decreasing in p. The range of feasible outage probability, in general, is an interval $[p_{o,\min}, 1]$ for some $p_{o,\min} \geq 0$. We say that an achievable point (T, p) dominates an achievable point (T', p') if $p \leq p'$ and $T \geq T'$. The Pareto boundary of $\mathcal{T}$ consists of all achievable points that are not dominated by other achievable points, i.e., it is given by $\{(T^*(p), p) : p \in [p_{o,\min}, 1]\}$.

It is also immediate to see that the throughput-outage tradeoff region is convex. In fact, consider two achievable points $(\overline{T}^{(1)}_{\min}, p_o^{(1)})$ and $(\overline{T}^{(2)}_{\min}, p_o^{(2)})$ corresponding to caching placements G_1 and G_2, with probability assignments $\Pi_c^{(1)}$ and $\Pi_c^{(2)}$. For $\lambda \in [0, 1]$, the caching placement G with mixture probability assignment $\Pi_c = \lambda \Pi_c^{(1)} + (1-\lambda)\Pi_c^{(2)}$ achieves $p_o = \lambda p_o^{(1)} + (1-\lambda) p_o^{(2)}$. For this value of outage probability, the best possible strategy achieves

$$T^*(p_o) \geq \min_u \left\{ \lambda \overline{T}_u^{(1)} + (1-\lambda)\overline{T}_u^{(2)} \right\} \geq \lambda \min_u \overline{T}_u^{(1)} + (1-\lambda) \min_u \overline{T}_u^{(2)},$$

where, by definition, $\overline{T}^{(1)}_{\min} = \min_u \overline{T}_u^{(1)}$ and $\overline{T}^{(2)}_{\min} = \min_u \overline{T}_u^{(2)}$. Hence, the segment joining any

two achievable throughput-outage points $(\overline{T}^{(1)}_{\min}, p_o^{(1)})$ and $(\overline{T}^{(2)}_{\min}, p_o^{(2)})$ is contained into the achievable throughput-outage region.

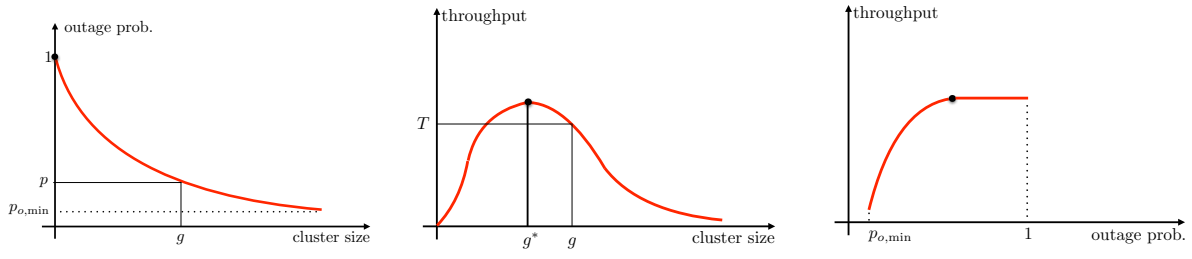
We conclude this section by providing the intuition behind the tension between outage and throughput, and explaining through an intuitive argument why $T^*(p)$ is non-decreasing for $p \in [p_{o,\min}, 1]$. The key tradeoff quantity here is the cooperation cluster size g, that is, the size of the set of nearest neighbor nodes among which any node u can look for its desired file f_u. On one hand, we would like to have g large, in order to take advantage of the content reuse, i.e., the larger g, the larger the probability that any user can find and retrieve its desired file. On the other hand, we would like to have g small, in order to take advantage of the spatial reuse, i.e., the smaller g, the larger the number of simultaneously active links that the network can support. Therefore, g describes the tradeoff between content reuse and spatial reuse.

As g increases the probability that user u does not find its desired file decreases. Hence, p_o is a decreasing function of g (see Fig.3(a)).

Since nodes can retrieve their desired files within a cluster of size g, then the communication range of the D2D links must be enough to communicate across such clusters. The average number of active links that can be activated without violating the protocol model is of the order of the number of disjoint clusters in the network, i.e., $\frac{n}{g}$. With a probability $1 - p_o$ that any user cannot find its requested file, the average per-user throughput is roughly given by $T \propto \frac{1-p_o}{n} \times \frac{n}{g} = \frac{1-p_o}{g}$. Since for small g we have $p_o \uparrow 1$, and for large g we have $p_o \downarrow p_{o,\min}$, where the latter is the probability that a node u does not find its requested file in the whole network, we clearly see that T must be increasing for small g and decreasing for large g (see Fig.3(b)).

Now, consider the constraint on the outage probability $p_o \leq p$. If p is small, then the constraint must be satisfied with equality, yielding the corresponding value of g (Fig.3(a)) which in turns yields a corresponding value of T (Fig.3(b)). As p increases, we obtain the concave increasing part of the throughput vs outage Pareto boundary curve qualitatively depicted in Fig. 3(c). However, when p becomes larger than some threshold, the optimal throughput is obtained by letting $g = g^*$, which is the size that achieves the maximum unconstrained throughput (see Fig. 3(b)). This means that for values of p beyond this threshold value, the throughput curve reaches a bound (horizontal line), equal to the unconstrained maximum throughput, as shown in Fig. 3(c).

It follows from the upper bounds developed in Section III that this intuitive argument, even though it is developed for a cluster-based achievability strategy, holds true also for the upper bounds, despite the fact that the latter do not assume any a priori transmission strategy other than the one-hop constraint and



(a) (b) (c)

Fig. 3. Qualitative behavior of the tradeoff between throughput and outage probability, by ways of the tradeoff parameter g, which represents the size of the cluster of nodes over which any node can look for its desired requested file and download it by D2D on-hop communication.

the protocol model.

III. Outer Bounds

Under the one-hop restriction, network topology and protocol model given in Section II, we can provide an outer bound $T^{\mathrm{ub}}(p)$ on the throughput-outage tradeoff, such that the ensemble of points $\{(T^{\mathrm{ub}}(p), p) : p \in [0,1]\}$ dominates the tradeoff region $\mathcal{T}$. In what follows, the quantity

$$\alpha \stackrel{\Delta}{=} \frac{1-\gamma}{2-\gamma} \tag{6}$$

plays an important role. Notice that $0 \leq \alpha < 1/2$ under the assumption made here that $0 < \gamma < 1$. The following results are proved in Appendices A and B:

Theorem 1: When $\lim_{n\to\infty} \frac{m^\alpha}{n} = 0$, the throughput-outage region is dominated by the set of points $(T^{\mathrm{ub}}(p), p)$ given by :

$$T^{\mathrm{ub}}(p) = \begin{cases} \frac{16CM}{\Delta^2 m(1-p)^{\frac{1}{1-\gamma}}} + o\left(\frac{1}{m(1-p)^{\frac{1}{1-\gamma}}}\right), & p = 1 - \left(\frac{Mg_R(m)}{m}\right)^{1-\gamma} \\ \min\left\{\frac{16CM}{\Delta^2 m(1-p)^{\frac{1}{1-\gamma}}}, f_{\mathrm{ub}}(\rho')m^{-\alpha}\right\} + o\left(m^{-\alpha}\right), & 1-(M\rho')^{1-\gamma}m^{-\alpha} \leq p < 1-(M\rho^*)^{1-\gamma}m^{-\alpha} \\ f_{\mathrm{ub}}(\rho^*)m^{-\alpha} + o\left(m^{-\alpha}\right), & 1-(M\rho^*)^{1-\gamma}m^{-\alpha} \leq p \leq 1, \end{cases} \tag{7}$$

where $\rho' \geq \rho^*$ and ρ^* is the solution (with respect to ρ) of the equation

$$\zeta(\rho) = \log(1 + (2-\gamma)\zeta(\rho)) \tag{8}$$

with

$$\zeta(\rho) \triangleq \left(\left(1 + \frac{3\Delta}{2}\right)^{\frac{2}{2-\gamma}} \rho \right)^{2-\gamma} M^{1-\gamma}, \tag{9}$$

$g_R(m)$ is any function such that $g_R(m) = \omega\left(m^{\alpha}\right)$ and $g_R(m) \leq \min\{\frac{m}{M}, n\}$, and where

$$f_{\mathrm{ub}}(\rho) \triangleq \frac{16C}{\Delta^2 \rho} \left(1 - e^{-\zeta(\rho)}\right). \tag{10}$$

□

Theorem 2: When there exists a positive constant ξ such that $\xi \leq \lim_{n\to\infty} \frac{m^{\alpha}}{n} \leq \frac{16}{\Delta^2 \rho^*}$, the throughput-outage region is dominated by the set of points $(T^{\mathrm{ub}}(p), p)$ given by:

$$T^{\mathrm{ub}}(p) =$$
$$\begin{cases} \min\left\{\frac{16CM}{\Delta^2 m (1-p)^{\frac{1}{1-\gamma}}}, f_{\mathrm{ub}}(\rho') m^{-\alpha}\right\} + o(m^{-\alpha}), & 1 - (M\rho')^{1-\gamma} m^{-\alpha} \leq p < 1 - (M\rho^*)^{1-\gamma} m^{-\alpha} \\ f_{\mathrm{ub}}(\rho^*) m^{-\alpha} + o(m^{-\alpha}), & 1 - (M\rho^*)^{1-\gamma} m^{-\alpha} \leq p \leq 1, \end{cases} \tag{11}$$

where ρ^* is the solution of (8), and $\rho' \in [\rho^*, \frac{16}{\Delta^2} \frac{n}{m^{\alpha}}]$. □

Theorem 3: When $\lim_{n\to\infty} \frac{m^{\alpha}}{n} > \frac{16}{\Delta^2 \rho^*}$ (ρ^* being the solution of (8)), the throughput-outage region is dominated by the set of points $(T^{\mathrm{ub}}(p), p)$ given by :

$$T^{\mathrm{ub}}(p) = C \left(\frac{Mn}{m}\right)^{1-\gamma} + o\left(\left(\frac{n}{m}\right)^{1-\gamma}\right), \quad 1 - \left(\frac{Mn}{m}\right)^{1-\gamma} \leq p \leq 1. \tag{12}$$

□

Notice that the range of p in Theorems 2 and 3 is limited to $[p_{o,\min}, 1]$ with $p_{o,\min} = 1 - (M\rho')^{1-\gamma} m^{-\alpha}$ (for Theorem 2) and $p_{o,\min} = 1 - \left(\frac{Mn}{m}\right)^{1-\gamma}$ (for Theorem 3), showing that in these regimes the outage probability cannot be small. As a matter of fact, of all the regimes identified by Theorems 1, 2 and 3, the only *practically interesting* one is the first regime of Theorem 1. In particular, Theorem 2 and 3 show that, when $\lim_{n\to\infty} \frac{m^{\alpha}}{n}$ is bounded away from zero, any scheme for the one-hop D2D caching network yields outage probability that goes to 1, which is clearly not an acceptable. In contrast, $\lim_{n\to\infty} \frac{m}{n} = \kappa < \infty$, there might exist schemes achieving some fixed target outage probability value $p \in [0, 1)$, as $n, m \to \infty$. Intuitively, the function $g_R(m)$ plays the role of the size of the cooperation cluster of neighboring nodes within which each user can find its requested file. For example, choosing $g_R(m) = \beta m$ for some constant $\beta \leq \min\left\{\frac{1}{M}, \frac{1}{\kappa}\right\}$, both conditions $g_R(m) = \omega(m^{\alpha})$ and $g_R(m) \leq \min\left\{\frac{m}{M}, n\right\}$ are satisfied, for all sufficiently large n. Notice that for $\kappa \leq M$, the choice $\beta = 1/M$ yields that the outer bound contains points of the type $(O(M/m), 0)$ (zero outage probability). We shall see in the next section that throughput-outage points with throughput $\Omega(M/m)$ and fixed p bounded away from 1 are achievable.

For a conventional unicast system where users are served by a single omniscient node (e.g., a base station) that can store the whole library, the throughput scaling is $O(1/n)$.[4] Hence, in the case of $nM \gg m$, the combination of caching and D2D spatial reuse yields a very large throughput relative gain with respect to a conventional system. It is also interesting to notice that despite the first regime of Theorem 1 requires that m grows more slowly than $n^{1/\alpha}$, the only practically interesting sub-regime is $m = o(n)$, otherwise conventional unicast from the base station yields better throughput scaling and zero outage probability (all users are served).

All other regimes in Theorems 1 – 3 are included for completeness, in order to prove mathematically a somehow intuitively expected "negative" result: unless the library size m is small with respect to the aggregate caching memory nM, caching cannot achieve significant throughput gains with respect to conventional unicast from a single base station. This result is expected since, in this case, the *asynchronous content reuse* that the D2D caching network tries to exploit is essentially non-existent.

It is also important to notice that here we are considering the (realistic) case of a "heavy tail" Zipf request distribution with $\gamma \in (0, 1)$. If $\gamma > 1$, then a finite number of files collects essentially all the request probability mass, and this case is similar to the case of $m = O(1)$, which is a special case of $m = o(n)$. As a matter of fact, Zipf-distributed requests with $\gamma \in (0, 1)$ have been observed experimentally [31]–[33].

In the next section, we show that the upper bounds obtained here are tight in the scaling laws, and that the constants of the leading terms can be determined within constant gaps. This is obtained by exhibiting and analyzing a specific achievability strategy.

IV. Achievable Throughput-Outage Tradeoff

Consistently with the outer bounds in Theorems 1 – 3 and the concluding remarks in Section III, we consider achievability only in the "small library" regime $\lim_{n\to\infty} \frac{m^\alpha}{n} = 0$, for which there is hope to achieve some target outage probability strictly less than 1. We obtain an achievable inner bound on the achievable throughput-outage tradeoff region by considering a transmission policy based on *clustering* and a caching policy based on *independent random caching*.

Clustering: the network is divided into clusters of equal size, denoted by $g_c(m)$, and independent of the users' demands and cache placement realizations. A user can only look for the requested file inside

[4]This is obviously achieved by TDMA, serving users on different time slots in a round-robin fashion. Notice that even if a more refined physical layer including a Gaussian broadcast channel is considered, the throughput scaling remains the same.

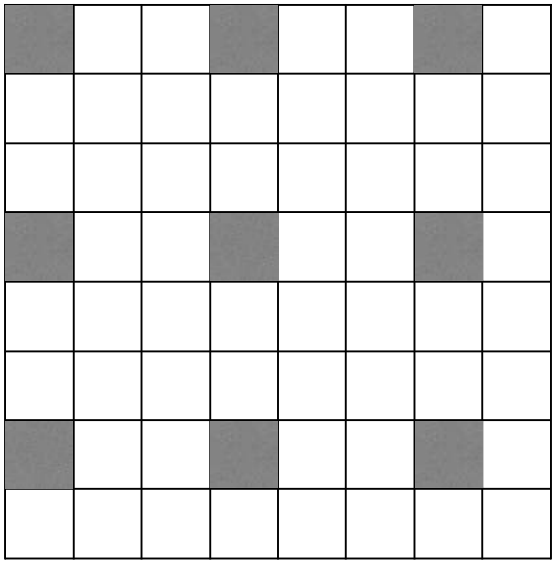


Fig. 4. Example of TDMA reuse scheme: each square represents a cooperation cluster. Gray squares represent the concurrently active clusters in a given time slot. In other times slots, other patterns of concurrently active clusters obtained by shifting the pattern in the figure are activated. In this particular example, the TDMA reuse parameter is $K = 9$.

its own cluster. For each user whose demand is found inside its cluster, we say that a *potential link* exists in the cluster. If a cluster contains at least one potential link, we say that this cluster is *good*. We use an *interference avoidance* transmission policy Π_t^* for which at most one concurrent transmission is allowed in each cluster, over any time-frequency slot (transmission resource). Furthermore, potential links inside the same cluster are scheduled with equal probability (or, equivalently, in round robin), such that all users have the same throughput $\overline{T}_u = \overline{T}_{\min}$. To avoid interference between clusters, we use a conventional TDMA with spatial reuse scheme [34, Ch. 17], very similar to the spatial reuse scheme of a cellular network, where each cluster acts as a cell. In short, a "coloring" scheme with K colors is applied to the clusters such that clusters with the same color can be concurrently active on the same time slot, without violating the protocol model. The resulting groups of clusters are assigned to K orthogonal time slots, and are activated in a round-robin fashion. In particular, we use $K = \left(\left\lceil \sqrt{2}(1+\Delta)\right\rceil + 1\right)^2$ in order to guarantee that concurrently active clusters do not interference with each other. Fig. 4 shows an example for $K = 9$.

Random Caching: we consider a caching policy Π_c^* where each node independently caches M files according to a common probability distribution P_c^*, given by the following result (proved in Appendix C):

Theorem 4: Under the system model assumptions and the clustering scheme described above, the caching distribution P_c^* that maximizes the probability that any user $u \in \mathcal{U}$ finds its requested file inside its corresponding cluster is given by

$$P_c^*(f) = \left[1 - \frac{\nu}{z_f}\right]^+, \quad f = 1, \ldots, m, \tag{13}$$

where $\nu = \frac{m^*-1}{\sum_{j=1}^{m^*} \frac{1}{z_j}}$, $z_j = P_r(j)^{\frac{1}{M(g_c(m)-1)-1}}$, and $m^* = \Theta\left(\min\{\frac{M}{\gamma} g_c(m), m\}\right)$. □

The following theorem (proved in Appendices D) yields an inner bound on the achievable outage-throughput tradeoff region:

Theorem 5: Assume $\lim_{n\to\infty} \frac{m^\alpha}{n} = 0$. Then, the throughput-outage tradeoff achievable by random caching and clustering behaves as:

$$T(p) = \begin{cases} \frac{C}{K}\frac{M}{\rho_1 m} + o(1/m), & p = (1-\gamma)e^{\gamma-\rho_1} \\ \frac{CA}{K}\frac{M}{m(1-p)^{\frac{1}{1-\gamma}}} + o\left(\frac{1}{m(1-p)^{\frac{1}{1-\gamma}}}\right), & p = 1 - a\left(\frac{g_c(m)}{m}\right)^{1-\gamma} \\ \frac{CB}{K} m^{-\alpha} + o\left(m^{-\alpha}\right), & 1 - a\rho_2^{1-\gamma} m^{-\alpha} \leq p \leq 1 - ab^{1-\gamma} m^{-\alpha} \\ \frac{CD}{K} m^{-\alpha} + o\left(m^{-\alpha}\right), & 1 - ab^{1-\gamma} m^{-\alpha} \leq p \leq 1, \end{cases} \tag{14}$$

where we define $a = \gamma^\gamma M^{1-\gamma}$, $b = \left(\frac{1-\gamma}{a}\right)^{\frac{1}{2-\gamma}}$, $A \triangleq \gamma^{\frac{\gamma}{1-\gamma}}$, $B \triangleq \frac{a\rho_2^{1-\gamma}}{1+a\rho_2^{2-\gamma}}$, $D \triangleq \frac{ab^{1-\gamma}}{1+ab^{2-\gamma}}$ and where ρ_1 and ρ_2 are positive parameters satisfying $\rho_1 \geq \gamma$ and $\rho_2 \geq b$. The cluster size $g_c(m)$ is any function of m satisfying $g_c(m) = \omega\left(m^\alpha\right)$ and $g_c(m) \leq \gamma m/M$. □

In all cases, the achievable throughput scaling law both for p bounded away from 1 and $p \to 1$ coincide with the outer bounds of Theorem 1. Therefore, these throughput scaling laws are tight up to some gap in the constants of the leading terms.

In the rest of this section we compare the achievable throughput scaling law of Theorem 5 with the outer bounds of Section III and with the performance achievable by other schemes. In particular, we focus on the interesting regime of small library (Theorems 1 and 5). Since $\alpha < 1/2$, even in this regime the library size m can grow faster than n^2. However, we restrict to the practically relevant regime of $m = O(n)$ (linear or sub-linear in n). Choosing $g_c(m) = \beta m$ for some $\beta > 0$, it is apparent from the first and second line of (14) that p strictly bounded away from 1. By fixing a small but positive target outage probability, the per-user average throughput of the D2D one-hop caching network with random (decentralized) caching scales as $T^*(p) = \Theta\left(\max\left\{\frac{1}{n}, \frac{M}{m}\right\}\right)$, where the scaling $\Theta\left(\frac{1}{n}\right)$ can be trivially achieved by letting the whole network to be a single cluster (e.g., transmission radius $R = \sqrt{2}$) and serving one demand per unit time. This scaling is equivalent to conventional unicast from a single

omniscient node which can be regarded as the state of the art of today's (single cell) systems, with a base station or access point serving individual requests without exploiting the asynchronous content reuse. We notice that, when $nM \gg m$, the throughput of the D2D caching network achieves per-user throughput that increases linearly with M. In this regime, caching in the user nodes and exploiting the dense spatial reuse of the D2D network is a very attractive approach, since storage space is much "cheaper" than scarce resources such as bandwidth or dense base station deployment (the reader will forgive this vague statement in this context).

It is interesting to notice that our analysis is able to characterize also the constant of the leading term within a bounded gap. This is a fortunate fact that does not happen often for the scaling analysis of wireless network capacity (e.g., see [19]–[23]). For example, upper and lower bounds (Theorem 1 and 5, respectively) and finite-dimensional simulation results are compared in Fig. 5, which shows both theoretical (solid lines) and simulated (dashed lines) curves of the throughput (y-axis) vs. outage (x-axis) tradeoff for different values of γ. In this simulation, the throughput is normalized by C, so that it is independent of the link rate. In particular, the theoretical curves show the dominant term in (14) divided by C. In these examples we used $m = 1000$, $n = 10000$, $M = 1$ and the spatial reuse factor $K = 4$. The Zipf parameter γ varies from 0.1 to 0.6, corresponding to the curves from the left (blue) to the right (cyan).

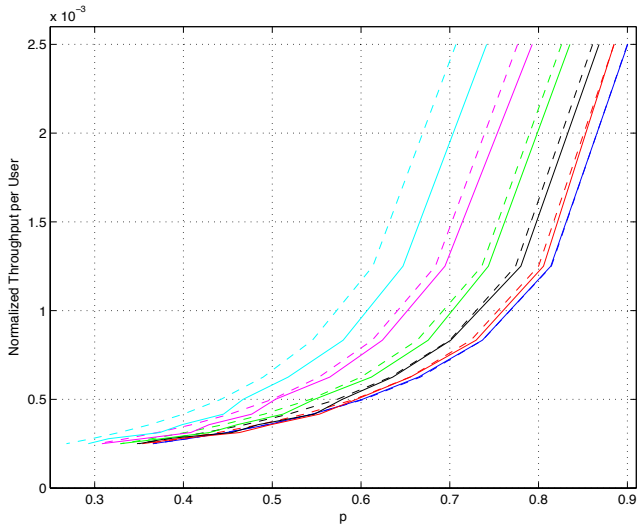


Fig. 5. Comparison between the normalized theoretical result (solid lines) and normalized simulated result (dashed lines) in terms of the minimum throughput per user vs. outage probability, where $m = 1000$, $n = 10000$, $M = 1$ and spatial reuse factor $K = 4$. The Zipf parameter γ varies from 0.1 to 0.6 (from the left (blue) to the right (cyan)).

As anticipated in Section I, in order to understand the potential of the combination of D2D one-hop communication and caching in the user devices, it is instructive to compare the scaling laws achieved by the D2D caching network with those achievable by other possible approaches. We have already discussed conventional unicast, achieving $\Theta(\frac{1}{n})$ throughput.

When the number of files m is less than the number of users n, an alternative consists of broadcasting all files on orthogonal channels. In order to guarantee that any requesting user can start its playback within a delay much shorter than the whole file duration, also in the presence of asynchronous requests, a well-known approach consists of Harmonic Broadcasting [35]–[38]. This scheme broadcasts continuously to all users a common message formed by the $m' \leq m$ most probable files, each of which is encoded in the way proposed in [35], refined in [36], [37] and whose optimality in an information theoretic sense was established in [38]. Without entering the details, each file of length L packets is divided into blocks of length L/N packets and encoded with bandwidth expansion factor $H(1,1,N) = \sum_{i=1}^{N} \frac{1}{i}$, such that the storage space at each user node is $0 < M < 1$ (less than one entire file is stored at each given time), and the time that any user must wait between the instant at which the streaming session is requested and the instant at which the playback starts is not larger than L/N packets. In order to allow the users to start playback within a finite delay while $L \to \infty$, the ratio L/N must be finite, i.e., it must be $N = \Theta(L)$. Furthermore, for $m \to \infty$ and $0 < \gamma < 1$, in order to have outage probability bounded away from 1, we need $m' = \Theta(m)$. Therefore, the bandwidth expansion factor of harmonic broadcasting in this regime is $m' \log N = \Theta(m \log L)$. It follows that the throughput of Harmonic Broadcasting scales as $\Theta(\frac{1}{m \log L})$. From a strictly technical viewpoint, since in our system assumptions we study the system performance by first letting $L \to \infty$, and then considering n, m that simultaneously grow in some relation, the throughput of harmonic broadcasting under our assumption is identically zero. In practice, for large but finite L, the gain of D2D caching over Harmonic Broadcasting can be appreciated by by comparing the term $\frac{M}{m}$ with the term $\frac{1}{m \log L}$ in the per-user throughput. It is clear that Harmonic Broadcasting does not take advantage of the user nodes storage memory, and in addition suffers an arbitrarily large multiplicative penalty as the length of the files L increases.

Finally, we examine the coded multicasting scheme of [16], already briefly described in Section I, which represent another example of one-hop network with caching in the user nodes, able to make efficient (and in fact, near-optimal in an information theoretic sense) use of caching. The rate analysis provided in [16] shows that the number of equivalent file multicast transmissions from the base station

in order to satisfy any set of users' requests is given by

$$N(n,m,M) = n\left(1-\frac{M}{m}\right)\frac{1}{1+\frac{Mn}{m}},$$

such that the minimum average throughput per user is given by

$$T = \frac{C_0\left(1+\frac{Mn}{m}\right)}{n\left(1-\frac{M}{m}\right)},$$

where C_0 is the common downlink rate at which the base station can send the multicast coded message to all users. For large m, n and finite M, the scaling of the per-user throughput given again by $\Theta\left(\max\left\{\frac{1}{n}, \frac{M}{m}\right\}\right)$, where the two terms inside the max are realized depending on whether $nM \gg m$, or $nM \ll m$. Interestingly and somehow surprisingly, this is the same scaling behavior of the D2D caching network studied in this paper. [5]

V. Conclusions

In this paper we have considered a wireless device-to-device (D2D) network where the nodes have pre-cached information from a fixed library of possible files, users request files at random and, if the requested file is not in the on-board cache, then it is downloaded from some neighboring node via one-hop "local" communication. To model the wireless network, we have followed the simple *protocol model*, widely considered in the analysis of the transport capacity scaling laws of wireless ad-hoc networks.

We have proposed a model that captures mathematically the *asynchronous content reuse* typical of on-demand video streaming, where the users' requests have strong overlap and concentrate on a small set of popular movies, but the demands are completely asynchronous, such that "naive multicasting" is not effective.

In our model, a user is in outage when its requested file is not found within the allowed transmission range. We have defined the optimal tradeoff between minimum per-user average throughput and the average fraction of users in outage, that we refer to as outage probability. Then, we have characterized such optimal tradeoff in terms of tight scaling laws in all the scaling regimes of the system parameters, when both the number of nodes and the number of files in the library grow to infinity.

The main result of this work is that, in the relevant regime "small library", i.e., when $m = O(n)$ and the aggregate cache capacity of the network, nM is much larger that the library size m, the throughput of

[5]For a performance comparison between the D2D caching network of the present work, coded multicasting in [16], Harmonic Broadcasting in [35] and conventional unicasting under realistic assumptions on the underlying D2D and cellular physical channels, please see [39].

the D2D one-hop caching network is proportional to M/m and independent of n. Hence, D2D one-hop caching networks are very attractive to handle situations where a relatively small library of popular files (e.g., the 500 most popular movies and TV shows of the week) is requested by a large number of users (e.g., 10,000 users per km^2 in a typical urban environment). In this regime, the proposed system is able to efficiently turn memory into bandwidth, in the sense that the per-user throughput increases proportionally to the cache capacity M of the user devices. Since the latter follows the doubling rate of Moore's law, caching in the user devices can achieve orders of magnitude throughput gains without requiring more bandwidth.

Interestingly, the same throughput scaling law is achieved by coded multicasting [16], [17] for a different one-hop network topology with caching at the user nodes, where a single central node (e.g., a base station) multicast network-coded codewords formed by EXORing data packets. It is worthwhile to point out that, although these schemes yield the same throughput scaling law, they achieve their (order-optimal) "caching gain" according to two completely different principles. The D2D caching network exploits the dense spatial reuse provided by caching, i.e., by replicating the same file many times in the network, any user with high probability can find its requested file at short distance, such that many simultaneously active links can be supported on the same time slot. In contrast, coded multicasting achieves its gain by using network coding in order to create a single message which is simultaneously useful for many users. While in the D2D network transmissions should be "as local as possible" in order to exploit spatial reuse, in the coded multicast network transmissions should be as global as possible, in order to benefit the largest number of users. In a recent follow-up paper [40], we have investigated a decentralized version of network-coded scheme of [16] for the same D2D network of the present work, without any omniscient node that has the whole file library. It turns out that coded multicasting gain and spatial reuse gain do not cumulate. Thus, it seems that the throughput scaling law obtained here is somehow an inherent limitation of one-hop networks with caching in the user nodes.

APPENDIX A
PROOF OF THEOREMS 1 AND 2

We first provide an outline of the proof and then dig into the details.

1) We define $T_{\rm sum} = \sum_{u=1}^{n} \overline{T}_u$ and let $(T^*_{\rm sum}(p), p)$ be the solution of

$$\begin{aligned} &\text{maximize} && T_{\rm sum} \\ &\text{subject to} && p_o \leq p, \end{aligned} \tag{15}$$

where the maximization is with respect to the cache placement and transmission policies Π_c, Π_t. As for $T^*(p)$, also $T^*_{\rm sum}(p)$ is non-decreasing in p. Furthermore, the inequality $T^*(p) \leq \frac{1}{n} T^*_{\rm sum}(p)$ follows immediately from the definition of $T^*_{\rm sum}(p)$ and $T^*(p)$.

2) We parameterize problem (15) with respect to the number of nodes in a disk of radius R, referred to (for brevity) as "disk size" and indicated by $g_R(m)$, where R denotes the one-hop transmission range of the protocol model. For any value $g_R(m) = g$, let $\mathcal{T}^*_{\rm sum}(g)$ denote the largest achievable sum throughput with disk size g, and let $p^*_o(g)$ denote the corresponding outage probability. While obtaining exact expressions for $\mathcal{T}^*_{\rm sum}(g)$ and for $p^*_o(g)$ is difficult, we shall obtain an upper bound $\mathcal{T}^{\rm ub}_{\rm sum}(g) \geq \mathcal{T}^*_{\rm sum}(g)$ and a lower bound $p^{\rm lb}(g) \leq p^*_o(g)$. By the monotonicity property said above, it follows that $(\mathcal{T}^{\rm ub}_{\rm sum}(g), p^{\rm lb}(g))$ dominates $(\mathcal{T}^*_{\rm sum}(g), p^*_o(g))$ and, as a consequence, $(\frac{1}{n}\mathcal{T}^{\rm ub}_{\rm sum}(g), p^{\rm lb}(g))$ dominates $(T^*(p), p)$ for $p = p^*_o(g)$. Also, we have that the set of outage probability values $p^*_o(g)$ obtained by varying g includes the feasibility domain $[p_{o,\min}, 1]$ of the original problem (5). This implies that the set of points $(\frac{1}{n}\mathcal{T}^{\rm ub}_{\rm sum}(g), p^{\rm lb}(g))$, obtained by varying g, dominates the Pareto boundary of the throughput-outage region $\mathcal{T}$.

3) Finally, we shall consider separately the different regimes of the outer bound, by "eliminating" the parameter $g_R(m)$. Conceptually, this can be obtained by letting $p = p^{\rm lb}(g)$, solving for g as a function of p and replacing the result into $\mathcal{T}^{\rm ub}_{\rm sum}(g)$. The resulting outer bound shall be denoted simply by $(T^{\rm ub}(p), p)$, given by Theorems 1 and 2.

We focus first on Theorem 1, where $\lim_{n\to\infty} \frac{m^\alpha}{n} = 0$, and consider in details step 2) of the above outline. In the following, we shall implicitly ignore the non-integer effects when they are irrelevant for the scaling laws. For example, recalling that the network has node density n (we have n nodes in the unit square), the disk size is given (up to integer rounding) by

$$g_R(m) = \pi R^2 n \stackrel{\Delta}{=} g. \tag{16}$$

For given disk size g, a lower bound on p_o can be obtained by observing that $1 - p_o$ is upper bounded by the maximum over the users $u = 1, \cdots, n$, of the probability that user u can be served by the D2D network. A necessary condition for this to happen is that the message f_u is found in the cache of some node inside a disk of size g centered at node u. We denote such event by $\mathcal{F}^u_g$.[6] If $g \geq m/M$, then the outage probability lower bound is zero, since we can arrange the files in the caches such that at least

[6]Notice: events are defined in the probability space of the triple $(\mathsf{f}, \mathsf{G}, \mathsf{A}) \sim \prod_{i=1}^n P_r(f_i)\Pi_c(\mathsf{G})\Pi_t(\mathsf{A}|\mathsf{f}, \mathsf{G})$, of requests, cache placements and transmission scheduling decisions.

one node u finds all files in the library within a radius R. Hence, assuming $g < m/M$, we have

$$\begin{aligned} 1-p_o &\leq \max_u \mathbb{P}\left(\mathcal{F}_g^u\right) \\ &\overset{(a)}{\leq} \sum_{f=1}^{Mg} P_r(f) = \sum_{f=1}^{Mg} \frac{f^{-\gamma}}{H(\gamma,1,m)} \\ &= \frac{H(\gamma,1,Mg)}{H(\gamma,1,m)}, \end{aligned} \tag{17}$$

where (a) follows by caching all most popular Mg files within a disk of radius R form a given user. In order to estimate the value of $H(\cdot,\cdot,\cdot)$, we have the following lemma:

Lemma 1: For $\gamma \neq 1$, then

$$\frac{1}{1-\gamma}(y+1)^{1-\gamma} - \frac{1}{1-\gamma}x^{1-\gamma} \leq H(\gamma,x,y) \leq \frac{1}{1-\gamma}y^{1-\gamma} - \frac{1}{1-\gamma}x^{1-\gamma} + \frac{1}{x^\gamma}. \tag{18}$$

For $\gamma = 1$, then

$$\log(y+1) - \log(x) \leq H(\gamma,x,y) \leq \log(y) - \log(x) + \frac{1}{x}. \tag{19}$$

Proof: See Appendix E. ■

From (17) and Lemma 1, we have the lower bound

$$p_o^*(g) \geq p^{\mathrm{lb}}(g) \triangleq \begin{cases} 0 & \text{for } g \geq \frac{m}{M} \\ 1 - \frac{\frac{1}{1-\gamma}(Mg)^{1-\gamma} - \frac{1}{1-\gamma} + 1}{\frac{1}{1-\gamma}m^{1-\gamma} - \frac{1}{1-\gamma}} & \text{for } g < \frac{m}{M} \end{cases} \tag{20}$$

Next, we seek an upper bound on $\mathcal{T}_{\mathrm{sum}}^*(g)$ as a function of the disk size g. According to the protocol model (see Section II), the throughput T_{sum} is given by

$$T_{\mathrm{sum}} = C \cdot \mathbb{E}\left[\mathsf{L}\right], \tag{21}$$

where L is the number of active links over any strategy with transmission radius R. Letting (i,j) and (k,l) denote two distinct transmitter-receiver pairs, using the triangle inequality and the protocol model constraints, we have

$$\begin{aligned} d(j,l) &\geq d(k,j) - d(k,l) \\ &\geq (1+\Delta)R - d(k,l) \\ &\geq (1+\Delta)R - R = \Delta R. \end{aligned} \tag{22}$$

Hence, any two receivers must be separated by distance not smaller than ΔR. Equivalently, disks of radius $\frac{\Delta}{2}R$ around any receiver must be disjoint. Since there is at least a fraction $1/4$ of the area of such

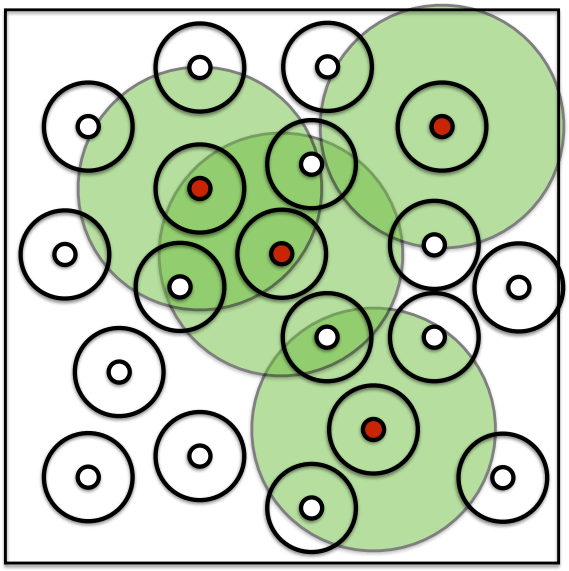


Fig. 6. Illustration of the fact that the number of small disks intersecting the union of the big disks centered at the active receivers is necessarily an upper bound to the number of active receivers.

disks inside the unit square containing our network, the number of such disjoint disks in the unit square is upper bounded by $\left\lceil \frac{16}{\pi\Delta^2 R^2} \right\rceil$.

We wish to upper bound the number of simultaneously active receivers L. In order to do so, consider the situation in Fig. 6, where the potentially active receivers (those that can receive according to the protocol model) are at the centers of mutually exclusive disks of radius $\frac{\Delta}{2}R$. Now, any of these receivers u can effectively receive only if $\mathcal{F}_g^u$ occurs. From (20), we have $\mathbb{P}(\mathcal{F}_g^u) \leq 1 - p^{\text{lb}}(g)$. Now, consider a disk of radius $(1+\Delta)R$ around each active receiver (shown as filled dots in Fig. 6), and let $U(R, \Delta, \mathsf{L})$ denote the union of all such disks. It is clear that the number of active receivers L is less than or equal to the number of small disks of radius $\frac{\Delta}{2}R$ with non-empty intersection with $U(R, \Delta, \mathsf{L})$. Since, as argued before, there are at most $\left\lceil \frac{16}{\pi\Delta^2 R^2} \right\rceil$ such disks, we can write

$$\mathsf{L} \leq \sum_{i=1}^{\left\lceil \frac{16}{\pi\Delta^2 R^2} \right\rceil} 1\left\{\text{disk } i \cap U(R, \Delta, \mathsf{L})\right\}. \tag{23}$$

Taking expectation of both sides of (23), and denoting the disks of radius $\frac{\Delta}{2}R$ centered around the receivers simply as "disk", we can write

$$\begin{aligned}\mathbb{E}[\mathsf{L}] &\leq \sum_{i=1}^{\frac{16n}{\Delta^2 g}} \mathbb{P}\left(\text{disk } i \cap U(R,\Delta,\mathsf{L})\right) \\ &\leq \frac{16n}{\Delta^2 g} \cdot \mathbb{P}\left(\text{Any disk} \cap U(R,\Delta,\mathsf{L})\right). \end{aligned} \tag{24}$$

Then we introduce the following lemma.

Lemma 2:

$$\mathbb{P}\left(\text{Any disk} \cap U(R,\Delta,\mathsf{L})\right) \leq \mathbb{P}\left(\exists \text{ an active receiver in a disk of radius } \left(1+\tfrac{3\Delta}{2}\right)R\right). \tag{25}$$

Proof: See Appendix F. ■

Using Lemma 2 in (24), we obtain

$$\begin{aligned}\mathbb{E}[\mathsf{L}] &\leq \frac{16}{\Delta^2}\left(\frac{n}{g}\right) \cdot \mathbb{P}\left(\exists \text{ an active receiver in a disk of radius } \left(1+\tfrac{3\Delta}{2}\right)R\right) \\ &\overset{(a)}{\leq} \frac{16}{\Delta^2}\left(\frac{n}{g}\right) \cdot \left(1-\left(p^{\mathrm{lb}}(g)\right)^{\left(1+\frac{3\Delta}{2}\right)^2 g}\right), \end{aligned} \tag{26}$$

where (a) follows from the fact that, recalling (16), the number of users in a disk of radius $\left(1+\frac{3\Delta}{2}\right)R$ is given by

$$n\pi\left(1+\frac{3\Delta}{2}\right)^2 R^2 = g\left(1+\frac{3\Delta}{2}\right)^2, \tag{27}$$

and the probability that no users in such disk find their requested content within the transmission range R can be lower bounded as

$$\mathbb{P}\left(\bigcap_{i=1}^{g\left(1+\frac{3\Delta}{2}\right)^2} (\mathcal{F}_g^i)^c\right) \geq \mathbb{P}\left(\bigcap_{i=1}^{g\left(1+\frac{3\Delta}{2}\right)^2} \{\mathsf{f}_i > gM\}\right) \geq \left(p^{\mathrm{lb}}(g)\right)^{\left(1+\frac{3\Delta}{2}\right)^2 g}.$$

Using (26) in (21), we obtain the sought upper bound $\mathcal{T}_{\mathrm{sum}}^{\mathrm{ub}}(g)$ on $\mathcal{T}_{\mathrm{sum}}^{*}(g)$ as

$$\mathcal{T}_{\mathrm{sum}}^{*}(g) \leq \mathcal{T}_{\mathrm{sum}}^{\mathrm{ub}}(g) \triangleq \frac{16C}{\Delta^2} \cdot \left[\left(1-\left(p^{\mathrm{lb}}(g)\right)^{\left(1+\frac{3\Delta}{2}\right)^2 g}\right)\frac{n}{g}\right]. \tag{28}$$

In order to discuss the different regimes of the outer bound, we start by considering the maximum throughput regime and the corresponding outage lower bound. This is obtained by maximizing $\mathcal{T}_{\mathrm{sum}}^{\mathrm{ub}}(g)$ in (28) with respect to g, and is given by the following result:

Lemma 3: As $m \to \infty$, the maximum of the quantity $\left[\left(1-\left(p^{\mathrm{lb}}(g)\right)^{\left(1+\frac{3\Delta}{2}\right)^2 g}\right)\frac{n}{g}\right]$ is given by $\frac{1}{\rho^*}\left(1-e^{-\zeta(\rho^*)}\right)\frac{n}{m^\alpha}$, where ρ^* is the solution of (8) with $\zeta(\rho)$ given by (9), and where the optimal g takes on the form $g^* = \rho^* m^\alpha$, with α given in (6).

Proof: See Appendix G. ■

Using Lemma 3, the resulting maximum (with respect to g) of the sum throughput upper bound is given by:

$$\mathcal{T}_{\mathrm{sum}}^{\mathrm{ub}}(g^*) \quad = \quad f_{\mathrm{ub}}(\rho^*)\frac{n}{m^\alpha}, \tag{29}$$

where $f_{\mathrm{ub}}(\rho)$ is defined in (10).

By replacing $g = g^*$ into (20), the corresponding value of the outage probability lower bound is given by

$$p^{\mathrm{lb}}(g^*) \quad = \quad 1 - \frac{\frac{1}{1-\gamma}\left(M\rho^* m^\alpha\right)^{1-\gamma} - \frac{1}{1-\gamma} + 1}{\frac{1}{1-\gamma}m^{1-\gamma} - \frac{1}{1-\gamma}}. \tag{30}$$

For large m, we have

$$p^{\mathrm{lb}}(g^*) = 1 - (M\rho^*)^{1-\gamma} m^{-\alpha} + o(m^{-\alpha}), \tag{31}$$

where we used the identity $(1-\gamma)(1-\alpha) = \alpha$.

At this point, we have essentially captured the throughput-outage tradeoff outer bound in the third line in expression (7) of Theorem 1. There is one small technical point that needs to be settled in order to obtain the desired result from (29) and (31), namely, we have to show that by introducing a perturbation of size $o(m^{-\alpha})$ in the outage probability lower bound $p^{\mathrm{lb}}(g^*)$, the corresponding perturbation of the throughput upper bound is $no(m^{-\alpha})$. This fact follows from the continuity of $\mathcal{T}_{\mathrm{sum}}^{\mathrm{ub}}(g)$ and $p^{\mathrm{lb}}(g)$ in g, and it is proved in Appendix J. After this perturbation argument, the throughput-outage point corresponding to the maximization of $\mathcal{T}_{\mathrm{sum}}^{\mathrm{ub}}(g)$ with respect to g shall be denoted by $((T^{\mathrm{ub}})^*, (p^{\mathrm{lb}})^*)$, with coordinates $(p^{\mathrm{lb}})^* = 1-(M\rho^*)^{1-\gamma}m^{-\alpha}$ and $(T^{\mathrm{ub}})^* = f_{\mathrm{ub}}(\rho^*)m^{-\alpha} + o(m^{-\alpha})$. The point $((T^{\mathrm{ub}})^*, (p^{\mathrm{lb}})^*)$ dominates the achievable throughput-outage tradeoff boundary $(T^*(p), p)$, for all $p \geq (p^{\mathrm{lb}})^*$, yielding the third line in expression (7) of Theorem 1.

Next, we characterize the other regimes of the outer bound on the throughput-outage tradeoff region by using (20) and (28), for different regimes of the disk size $g_R(m)$. It is clear from (20) that by increasing $g_R(m)$ beyond $g_R^*(m) = g^*$ given in Lemma 3, the outage probability lower bound decreases. We consider two cases: 1) $g_R(m) = \Theta\left(m^\alpha\right)$ with $g_R(m) > \rho^* m^\alpha$; 2) $\omega\left(m^\alpha\right) = g_R(m) \leq \min\{\frac{m}{M}, n\}$.

Case 1) In this case, we let $g = g_R(m) = \rho' m^\alpha$ with $\rho' > \rho^*$. Letting $m \to \infty$ in (20) and in (28), we obtain

$$1 - (M\rho')^{1-\gamma} m^{-\alpha} + o(m^{-\alpha}) \leq p^{\mathrm{lb}}(g) < 1 - (M\rho^*)^{1-\gamma} m^{-\alpha} + o(m^{-\alpha}),$$

and

$$\mathcal{T}^{\mathrm{ub}}_{\mathrm{sum}}(g) = f_{\mathrm{ub}}(\rho') \frac{n}{m^\alpha}. \tag{32}$$

With a derivation similar to what done in Appendix J, and not included in the paper for the sake of brevity, this yields part of second line in expression (7) of Theorem 1 (one of the two terms of the minimum).

Case 2) When $\omega\left(m^\alpha\right) = g_R(m) \leq \min\{\frac{m}{M}, n\}$, we use (20) and the probability bound as in (26) and write

$$\begin{aligned}
&\mathbb{P}\left(\exists \text{ an active receiver in a disk of radius } \left(1 + \tfrac{3\Delta}{2}\right) R\right) \\
&\leq 1 - \left(p^{\mathrm{lb}}(g)\right)^{\left(1+\frac{3\Delta}{2}\right)^2 g} \\
&= 1 - \left(1 - \frac{\frac{1}{1-\gamma}(Mg)^{1-\gamma} - \frac{1}{1-\gamma} + 1}{\frac{1}{1-\gamma} m^{1-\gamma} - \frac{1}{1-\gamma}}\right)^{\left(1+\frac{3\Delta}{2}\right)^2 g} \qquad (33) \\
&\leq 1 - o(1), \qquad (34)
\end{aligned}$$

where the last line follows from the fact that, writing the second term in (33) as

$$\left[\left(1 - M^{1-\gamma}\left(\frac{g}{m}\right)^{1-\gamma}(1 + o(1))\right)^{g}\right]^{\left(1+\frac{3\Delta}{2}\right)^2}, \tag{35}$$

we see that the condition for (35) to be non-vanishing in the limit for $g, m \to \infty$ is that

$$\left(\frac{g}{m}\right)^{1-\gamma} = \Theta(g^{-1}),$$

or, equivalently, that

$$g = \Theta(m^\alpha),$$

where $\alpha = \frac{1-\gamma}{2-\gamma}$ is the familiar quantity defined in (6). Hence, in the case $g = \omega(m^\alpha)$, the disk size g grows rapidly and the limit of (35) vanishes.

By using (34) into (24) with $g = g_R(m)$ we eventually obtain

$$\mathcal{T}^{\mathrm{ub}}_{\mathrm{sum}} = \frac{16C}{\Delta^2}\left(\frac{n}{g_R(m)}\right) + o\left(\frac{n}{g_R(m)}\right). \tag{36}$$

Moreover, from (20) we have

$$\begin{aligned} p^{\mathrm{lb}}(g) &= 1 - \frac{\frac{1}{1-\gamma}(Mg_R(m))^{1-\gamma} - \frac{1}{1-\gamma} + 1}{\frac{1}{1-\gamma}m^{1-\gamma} - \frac{1}{1-\gamma}} \\ &= 1 - \left(\frac{Mg_R(m)}{m}\right)^{1-\gamma} + o\left(\left(\frac{g_R(m)}{m}\right)^{1-\gamma}\right). \end{aligned} \tag{37}$$

Expressing $g_R(m)$ as a function of $p = p^{\mathrm{lb}}$, we find

$$g_R(m) = \frac{m}{M}(1-p)^{\frac{1}{1-\gamma}}.$$

Using this into (36) and following a perturbation argument similar to Appendix J, we find the desired form

$$\mathcal{T}^{\mathrm{ub}}_{\mathrm{sum}}(g) = T^{\mathrm{ub}}_{\mathrm{sum}}(p) = n\left(\frac{16CM}{\Delta^2 m(1-p)^{\frac{1}{1-\gamma}}} + o\left(\frac{1}{m(1-p)^{\frac{1}{1-\gamma}}}\right)\right), \tag{38}$$

which yields the first line and the second term in the minimum of the second line in expression (7) of Theorem 1.

By following into the same footsteps, Theorem 2 can be proved along the same lines with the only difference that, when there exists a positive constant ξ such that $\xi \leq \lim_{n\to\infty} \frac{m^\alpha}{n} \leq \frac{16}{\Delta^2\rho^*}$, the case $g_R(m) = \omega\left(m^\alpha\right)$ does not exist.

APPENDIX B
PROOF OF THEOREM 3

In the case $\lim_{n\to\infty} \frac{m^\alpha}{n} > \frac{16+\Delta^2}{\Delta^2\rho_2}$, an obvious upper bound of the sum throughput $\mathcal{T}^*_{\mathrm{sum}}(p)$ is provided by

$$\begin{aligned} T_{\mathrm{sum}} &= C \cdot \mathbb{E}[\mathsf{L}] \\ &\overset{(a)}{\leq} C\sum_{u=1}^{n}\sum_{f=1}^{Mn} P_r(f) = Cn\frac{H(\gamma,1,Mn)}{H(\gamma,1,m)} \\ &\overset{(b)}{\leq} Cn\frac{\frac{1}{1-\gamma}(Mn)^{1-\gamma} - \frac{1}{1-\gamma} + 1}{\frac{1}{1-\gamma}(m+1)^{1-\gamma} - \frac{1}{1-\gamma}} \\ &\leq n\left(CM^{1-\gamma}\frac{n^{1-\gamma}}{m^{1-\gamma}} + o\left(\frac{n^{1-\gamma}}{m^{1-\gamma}}\right)\right), \end{aligned} \tag{39}$$

where (a) is because we use a deterministic caching scheme (see Appendix G) which makes the network store the most n popular messages, and (b) follows from Lemma 1. Dividing by n, we obtain the upper bound

$$T^*(p) \leq T^{\mathrm{ub}}(p) \overset{\Delta}{=} CM^{1-\gamma}\frac{n^{1-\gamma}}{m^{1-\gamma}} + o\left(\frac{n^{1-\gamma}}{m^{1-\gamma}}\right). \tag{40}$$

Moreover, as n goes to infinity, the outage probability in this case can be computed as

$$p_o \geq 1 - \frac{H(\gamma, 1, Mn)}{H(\gamma, 1, m)} \geq 1 - M^{1-\gamma}\frac{n^{1-\gamma}}{m^{1-\gamma}} + o\left(\frac{n^{1-\gamma}}{m^{1-\gamma}}\right). \tag{41}$$

Again, following a perturbation argument similar to Appendix J), for $p \geq 1 - M^{1-\gamma}\frac{n^{1-\gamma}}{m^{1-\gamma}}$, we have $T^*(p) \leq T^{\text{ub}}(p)$ in (40). Otherwise, the problem is infeasible.

APPENDIX C
PROOF OF THEOREM 4

As mentioned in Section IV, we divide the network into clusters, each of which contains $g_c(m)$ nodes. In this case, let $\mathcal{F}^u_{g_c(m)}$ denote the event that user u can find the requested message inside its cluster of size $g_c(m)$. Letting $\mathbf{1}_u = 1\{\mathcal{F}^u_{g_c(m)}\}$, we define

$$p_u^c = \mathbb{E}[\mathbf{1}_u] = \mathbb{P}(\mathcal{F}^u_{g_c(m)}). \tag{42}$$

Our goal here is to find the caching distribution $P_c^*(f)$ that maximizes p_u^c. With independent random caching, the probability that a user u finds its request $\mathsf{f}_u = f$ in its cluster is given by $\mathbb{P}(\mathcal{F}^u_{g_c(m)}|\mathsf{f}_u = f) = 1 - (1 - P_c(f))^{M(g_c(m)-1)}$ (notice that we do not consider requests to files in the user own cache, since these do not generate any traffic). By the law of total probability, we can write

$$p_u^c = \sum_{f=1}^{m} P_r(f)\left((1 - (1 - P_c(f))^{Mg_c(m)-M}\right). \tag{43}$$

Letting $g_c(m) = g$ for simplicity of notation, and assuming $g > 2$, we have the convex optimization problem

$$\begin{aligned} \text{minimize} \quad & \sum_{f=1}^{m} P_r(f)(1 - P_c(f))^{Mg-M} \\ \text{subject to} \quad & \sum_{f=1}^{m} P_c(f) = 1, \;\; P_c(f) \geq 0 \quad \forall\, f \end{aligned} \tag{44}$$

The Lagrangian function for the problem is

$$\mathcal{L}(P_c, \xi) = \sum_{f=1}^{m} P_r(f)(1 - P_c(f))^{Mg-M} + \xi'\left(\sum_{f=1}^{m} P_c(f) - 1\right) \tag{45}$$

Taking the partial derivative with respect to $P_c(f)$ and using the KKT conditions [41] we obtain

$$P_c(f) = \left[1 - \left(\frac{\xi'}{P_r(f)M(g-1)}\right)^{1/(M(g-1)-1)}\right]^+ \tag{46}$$

It is immediate to see that the minimum is obtained when the sum probability constraint holds with equality. In order to solve for the Lagrangian multiplier that imposes the constraint with equality, it is

convenient to re-parameterize the problem by defining $(\frac{\xi'}{M(g-1)})^{\frac{1}{M(g-1)-1}} = \nu$ and $z_f = P_r(f)^{\frac{1}{M(g-1)-1}}$ where the coefficients z_f are non-increasing since $P_r(f)$ is non-increasing by assumption. Hence, we wish to solve

$$\sum_{f=1}^{m} \left[1 - \frac{\nu}{z_f}\right]^+ = 1$$

The unique solution must be found among the following conditions:

$$\begin{aligned} 1 - \frac{\nu}{z_1} = 1 \quad &\text{with} \quad \frac{\nu}{z_2} \geq 1 \\ 2 - \frac{\nu}{z_1} - \frac{\nu}{z_2} = 1 \quad &\text{with} \quad \frac{\nu}{z_3} \geq 1 \\ 3 - \frac{\nu}{z_1} - \frac{\nu}{z_2} - \frac{\nu}{z_3} = 1 \quad &\text{with} \quad \frac{\nu}{z_4} \geq 1 \\ &\vdots \\ m - \sum_{f=1}^{m} \frac{\nu}{z_f} = 1 \end{aligned} \tag{47}$$

which can be rewritten compactly as finding the unique index m^* for which the equation

$$\nu \left(\sum_{f=1}^{m^*} \frac{1}{z_f} \right) = m^* - 1 \tag{48}$$

has a solution in the interval for $\nu \geq z_{m^*+1}$ and $\nu \leq z_{m^*}$. Since we are guaranteed that such m^* exists, we can write

$$\nu(m^*) = \frac{m^* - 1}{\sum_{f=1}^{m^*} \frac{1}{z_f}} \tag{49}$$

From the conditions $\nu(m^*) \geq z_{m^*+1}$ and $\nu(m^*) \leq z_{m^*}$, we find that m^* is explicitly given as the unique integer in $\{1, 2, \cdots, m\}$ such that

$$m^* \geq 1 + z_{m^*+1} \sum_{f=1}^{m^*} \frac{1}{z_f}, \tag{50}$$

and

$$m^* \leq 1 + z_{m^*} \sum_{f=1}^{m^*} \frac{1}{z_f}. \tag{51}$$

Next, we wish to determine m^* as a function of $g = g_c(m)$ in the assumption that $g_c(m) \to \infty$ as $m \to \infty$. In order to do so, we shall evaluate the terms in the right-hand side of (50) and (51). Recalling the expression of z_f in terms of $P_r(f) = \frac{\kappa}{f^\gamma}$ (recall that we assume a Zipf distribution for the demands, with exponent $\gamma \in (0, 1)$), we have

$$z_{m^*+1} \sum_{f=1}^{m^*} \frac{1}{z_f} = \sum_{f=1}^{m^*} \left(\frac{f}{m^* + 1} \right)^{a'}, \tag{52}$$

and

$$z_{m^*}\sum_{f=1}^{m^*}\frac{1}{z_f}=\sum_{f=1}^{m^*}\left(\frac{f}{m^*}\right)^{a'}, \tag{53}$$

where we let $a'=\frac{\gamma}{M(g-1)-1}$ for brevity. We use the following integral lower and upper bounds

$$\frac{1}{(m^*+1)^{a'}}+\frac{1}{(m^*+1)^{a'}}\int_1^{m^*}x^{a'}dx\leq\sum_{f=1}^{m^*}\left(\frac{f}{m^*+1}\right)^{a'}\leq\frac{1}{(m^*+1)^{a'}}\int_1^{m^*+1}x^{a'}dx, \tag{54}$$

and

$$\frac{1}{(m^*)^{a'}}+\frac{1}{(m^*)^{a'}}\int_1^{m^*}x^{a'}dx\leq\sum_{f=1}^{m^*}\left(\frac{f}{m^*}\right)^{a'}\leq\frac{1}{(m^*)^{a'}}\int_1^{m^*+1}x^{a'}dx. \tag{55}$$

Solving the integrals, we obtain the lower bound (LB 1) and the upper bound (UB 1) in (54) in the form

$$\begin{aligned}\text{LB 1} &= \frac{a'}{a'+1}\frac{1}{(m^*+1)^{a'}}+\frac{m^*}{a'+1}\left(\frac{m^*}{m^*+1}\right)^{a'}\\ \text{UB 1} &= \frac{m^*+1}{a'+1}-\frac{1}{a'+1}\frac{1}{(m^*+1)^{a'}},\end{aligned}$$

and we obtain the lower bound (LB 2) and the upper bound (UB 2) in (55) in the form

$$\begin{aligned}\text{LB 2} &= \frac{a'}{a'+1}\frac{1}{(m^*)^{a'}}+\frac{m^*}{a'+1}\\ \text{UB 2} &= \frac{m^*+1}{a'+1}\left(\frac{m^*+1}{m^*}\right)^{a'}-\frac{1}{(m^*)^{a'}}\frac{1}{a'+1}.\end{aligned}$$

We let $m^*=c/a'$ for some constant c, and notice that $a'\downarrow 0$ as $g_c(m)\to\infty$ and that $\lim_{a'\downarrow 0}(1+c/a')^{a'}=1$, $\lim_{a'\downarrow 0}(1+a'/c)^{a'}=1$ and $\lim_{a'\downarrow 0}(c/a')^{a'}=1$. Hence, in the limit of $a'\downarrow 0$ we can write

$$\begin{aligned}\text{LB 1} &= \frac{c/a'}{a'+1}(1-\delta_1(a'))+\delta_2(a')\\ \text{UB 1} &= \frac{c/a'+1}{a'+1}-1+\delta_3(a'),\end{aligned}$$

and

$$\begin{aligned}\text{LB 2} &= \frac{c/a'}{a'+1}+\delta_4(a')\\ \text{UB 2} &= \frac{c/a'+1}{a'+1}(1+\delta_5(a'))-1+\delta_6(a'),\end{aligned}$$

where δ_i, $i=1,\cdots,6$ tend to zero from above as $a\downarrow 0$. It follows that

$$\frac{c/a'}{a'+1}(1-\delta_1(a'))+\delta_2(a')\leq z_{m^*+1}\sum_{f=1}^{m^*}\frac{1}{z_f}\leq\frac{c/a'+1}{a'+1}-1+\delta_3(a'),$$

and

$$\frac{c/a'}{a'+1}+\delta_4(a')\leq z_{m^*}\sum_{f=1}^{m^*}\frac{1}{z_f}\leq\frac{c/a'+1}{a'+1}(1+\delta_5(a'))-1+\delta_6(a'),$$

as $m^* = c/a'$ and $a \downarrow 0$. Replacing the common leading term in the LB 1, UB 1, LB 2 and UB 2 above into (50) and (51), we obtain

$$\frac{c}{a'} \gtrsim 1 + \frac{c/a'}{a'+1},$$

and

$$\frac{c}{a'} \lesssim 1 + \frac{c/a'}{a'+1},$$

which yields

$$\frac{c}{a'+1} \cong 1$$

Therefore, we obtain $c = 1$, which yields

$$m^* = \frac{1}{a'} = \frac{M(g_c(m)-1)-1}{\gamma} + O(1)$$

i.e., $m^* = \frac{M}{\gamma} g_c(m)$ to the leading order. Clearly, if $\frac{M}{\gamma} g_c(m) > m$, then $m^* = m$.

APPENDIX D
PROOF OF THEOREM 5

Recall that Theorem 5 deals with the small library regime $\lim_{n\to\infty} \frac{m^\alpha}{n} = 0$. We define the probability

$$p^c_{uu'} = \mathbb{E}[\mathbf{1}_u \mathbf{1}_{u'}] = \mathbb{P}(\mathcal{F}^u_{g_c(m)} \cap \mathcal{F}^{u'}_{g_c(m)}), \tag{56}$$

i.e., the probability that both user u and user u' can find the requested files in the corresponding cluster. We let

$$\mathsf{W} = \sum_{u=1}^{g_c(m)} \mathbf{1}_u, \tag{57}$$

denote the number of potential links in a cluster.

Given the random and independent caching placement Π^*_c and the random (or round robin) transmission policy Π^*_t as given at the beginning of Section IV, we let $T(p)$ denote achievable values of $\overline{T}_{\min}$ subject to the outage constraint $p_o \le p$. Also, we define $\overline{T}_{\text{sum}} = \sum_{u=1}^n \overline{T}_u$.

We provide first an outline of the proof and then dig into the details.

1) Under policies $P^*_c(f)$ and Π^*_t, we notice that both $\overline{T}_{\min}$ and p_o are uniquely determined by the cluster size $g_c(m)$. Hence, the maximum throughput is obtained by solving:

$$\begin{aligned} &\text{maximize} && \overline{T}_{\min} \\ &\text{subject to} && 0 < g_c(m) \le n. \end{aligned} \tag{58}$$

Since the exact solution of (58) is difficult to obtain, we instead compute a lower bound and, for the maximizing $g_c(m) = g_c^*(m)$, the corresponding value of the outage probability p_o^*.

2) It follows immediately form the definition that $T(p) = T(p_o^*)$ for all $p \geq p_o^*$ is achievable, by keeping $g_c(m) = g_c^*(m)$ and using the same caching and scheduling policies. This yields a lower bound on the achievable throughput-outage tradeoff when $p \geq p_o^*$.

3) In order to obtain a tradeoff for $p < p_o^*$, we increase $g_c(m)$ above $g_c^*(m)$. By letting $g_c(m)$ grow and calculating the corresponding value of p_o and (a lower bound on) $\overline{T}_{\min}$, we obtain a lower bound $T(p)$ for $p = p_o$ on the achievable throughput-outage tradeoff.

A. Achievable $T(p)$ when $p \geq p_o^*$

We first compute a lower bound on $\overline{T}_{\text{sum}}$ and the corresponding outage probability p_o for the caching and transmission policies Π_c^* with Π_t^*, with cluster size $g_c(m)$. Since the resulting system is symmetric with respect to any user, it follows that each user has exactly the same average throughput, such that $\overline{T}_{\min} = \frac{1}{n}\overline{T}_{\text{sum}}$. Then, we shall maximize the resulting (lower bound on) $\overline{T}_{\min}$ with respect to $g_c(m)$ in order to find $g_c^*(m)$, p_o^* and $T(p)$ for $p \geq p_o^*$. For simplicity of notations, in the following we ignore some of the smaller order terms as m and n goes to infinity.

The main tool to obtain a lower bound on $\overline{T}_{\text{sum}}$ is the Paley-Zygmund Inequality (see [42] and references therein). Letting again L denote the number of active links, we have

$$\begin{aligned} \overline{T}_{\text{sum}} &= C \cdot \mathbb{E}[\mathsf{L}] \\ &\overset{(a)}{=} C \cdot \mathbb{E}[\text{number of active clusters}], \end{aligned} \tag{59}$$

where (a) is because that in Π_t^*, only one transmission is allowed in each cluster. Moreover,

$$\begin{aligned} &\mathbb{E}[\text{number of active clusters}] \\ &\geq \frac{1}{K}\mathbb{E}[\text{number of good clusters}] \\ &= \frac{1}{K}\left(\text{total number of clusters in the network} \cdot \mathbb{P}(\mathsf{W} > 0)\right), \end{aligned} \tag{60}$$

where K is the TDMA reuse factor and we use the fact that a cluster is good if $\mathsf{W} > 0$.

From (60), we have that a lower bound of $\overline{T}_{\text{sum}}$ can b obtained by lower bounding $\mathbb{P}(\mathsf{W} > 0)$. The distribution of W is not obvious since the random variables $\mathbf{1}_u$ and $\mathbf{1}_{u'}$ are dependent when u and u' are in the same cluster and $u \neq u'$. Nevertheless, it is possible to compute the first and second moments of W. Then, with the help of the Paley-Zygmund Inequality, we can obtain a lower bound on $\mathbb{P}(\mathsf{W} > 0)$

which is good enough for our purposes. For completeness, the Paley-Zygmund Inequality is provided in the following lemma:

Lemma 4: Let X be a non-negative random variable such that $\mathbb{E}[X^2] < \infty$. Then for any $t \geq 0$ such that $t < \mathbb{E}[X]$, we have

$$\mathbb{P}(X > t) \geq \frac{(\mathbb{E}[X] - t)^2}{\mathbb{E}[X^2]}. \tag{61}$$

□

By using Lemma 4 with $t = 0$ and $X = \mathsf{W}$, we get

$$\mathbb{P}(\mathsf{W} > 0) \geq \frac{\mathbb{E}[\mathsf{W}]^2}{\mathbb{E}[\mathsf{W}^2]}. \tag{62}$$

Therefore, our goal is to find a lower bound for $\mathbb{E}[\mathsf{W}]$ and an upper bound $\mathbb{E}[\mathsf{W}^2]$ under the optimal caching distribution P_c^*, given by Theorem 4. First, we focus on $\mathbb{E}[\mathsf{W}]$. Using the expression $\mathbb{E}[\mathsf{W}] = \sum_{u=1}^{g_c(m)} p_u^c = g_c(m) p_u^c$, we shall focus on the computation of p_u^c as follows:

$$\begin{aligned}
p_u^c &= \sum_{f=1}^{m^*} P_r(f) \left(1 - \left(\frac{\nu}{z_f}\right)^{M(g_c(m)-1)}\right) \\
&\overset{(a)}{\leq} \sum_{f=1}^{m^*} P_r(f) \left(1 - \left(\frac{z_{m^*+1}}{z_f}\right)^{M(g_c(m)-1)}\right) \\
&= \sum_{f=1}^{m^*} P_r(f) \left(1 - \left(\frac{P_r(m^*+1)}{P_r(f)}\right)^{\frac{M(g_c(m)-1)}{M(g_c(m)-1)-1}}\right) \\
&= \sum_{f=1}^{m^*} P_r(f) - \sum_{f=1}^{m^*} P_r(f) \left(\frac{P_r(m^*+1)}{P_r(f)}\right)^{\frac{M(g_c(m)-1)}{M(g_c(m)-1)-1}} \\
&= \sum_{f=1}^{m^*} P_r(f) - \sum_{f=1}^{m^*} P_r(f) \left(\frac{P_r(m^*+1)}{P_r(f)}\right) \left(\frac{P_r(m^*+1)}{P_r(f)}\right)^{\frac{1}{M(g_c(m)-1)-1}} \\
&= \sum_{f=1}^{m^*} P_r(f) - P_r(m^*+1) \sum_{f=1}^{m^*} \left(\frac{f}{m^*+1}\right)^{\frac{\gamma}{M(g_c(m)-1)-1}} \\
&= \frac{H(\gamma, 1, m^*)}{H(\gamma, 1, m)} - \frac{(m^*+1)^{(-\gamma)}}{H(\gamma, 1, m)} \sum_{f=1}^{m^*} \left(\frac{f}{m^*+1}\right)^{\frac{\gamma}{M(g_c(m)-1)-1}},
\end{aligned} \tag{63}$$

where (a) is because $\nu \geq z_{m^*+1}$ (see Theorem 4 and its proof in Section C). Similarly, we have

$$\begin{aligned} p_u^c &= \sum_{i=1}^{m^*} P_r(f) \left(1 - \left(\frac{\nu}{z_f}\right)^{M(g_c(m)-1)}\right) \\ &\overset{(a)}{\geq} \sum_{i=1}^{m^*} P_r(f) \left(1 - \left(\frac{z_{m^*}}{z_f}\right)^{M(g_c(m)-1)}\right) \\ &= \frac{H(\gamma,1,m^*)}{H(\gamma,1,m)} - \frac{(m^*)^{(-\gamma)}}{H(\gamma,1,m)} \sum_{f=1}^{m^*} \left(\frac{f}{m^*}\right)^{\frac{\gamma}{M(g_c(m)-1)-1}}, \end{aligned} \tag{64}$$

where (a) is because $\nu \leq z_{m^*}$ (again, see Theorem 4 and its proof in Section C).

By (63), (64) and Lemma 1, we have

$$\begin{aligned} p_u^c &\leq \frac{1}{\frac{1}{1-\gamma}(m+1)^{1-\gamma} - \frac{1}{1-\gamma}} \left(\left(\frac{1}{1-\gamma} m^{*1-\gamma} - \frac{1}{1-\gamma} + 1\right) - m^{*(-\gamma)} \sum_{f=1}^{m^*} \left(\frac{f}{m^*+1}\right)^{\frac{\gamma}{M(g_c(m)-1)}} \right) \\ &\leq \frac{1-\gamma}{(m+1)^{1-\gamma}-1} \cdot \left(\frac{1}{1-\gamma} m^{*1-\gamma} - m^{*(-\gamma)} \sum_{f=1}^{m^*} \left(\frac{f}{m^*+1}\right)^{\frac{\gamma}{M(g_c(m)-1)}} - \frac{\gamma}{1-\gamma} \right) \\ &\overset{(a)}{=} \frac{1-\gamma}{(m+1)^{1-\gamma}} \cdot \left(\frac{1}{1-\gamma} \left(\frac{M}{\gamma} g_c(m)\right)^{1-\gamma} - \left(\frac{M}{\gamma} g_c(m)\right)^{(-\gamma)} \right. \\ &\quad \left. \cdot \sum_{f=1}^{\frac{M}{\gamma} g_c(m)} \left(\frac{f}{\frac{M}{\gamma} g_c(m)+1}\right)^{\frac{\gamma}{M(g_c(m)-1)-1}} - \frac{\gamma}{1-\gamma} \right) \\ &\leq \gamma^{\gamma-1} \left(\frac{M g_c(m)}{m+1}\right)^{1-\gamma} - (1-\gamma)\gamma^{\gamma} \frac{(M g_c(m))^{-\gamma}}{(m+1)^{1-\gamma}} \left(\frac{1}{\frac{M}{\gamma} g_c(m)+1}\right)^{\frac{\gamma}{M(g_c(m)-1)-1}} \\ &\quad \cdot \left(1 + \int_1^{\frac{M}{\gamma} g_c(m)} x^{\frac{\gamma}{M(g_c(m)-1)-1}} dx\right) - \frac{\gamma}{(m+1)^{1-\gamma}} \\ &= \gamma^{\gamma-1} \left(\frac{M g_c(m)}{m+1}\right)^{1-\gamma} - (1-\gamma)\gamma^{\gamma} \frac{(M g_c(m))^{-\gamma}}{(m+1)^{1-\gamma}} \left(\frac{1}{\frac{M g_c(m)}{\gamma}+1}\right)^{\frac{\gamma}{M(g_c(m)-1)-1}} \\ &\quad \cdot \left(\frac{\frac{\gamma}{M(g_c(m)-1)-1}}{\frac{\gamma}{M(g_c(m)-1)-1}+1} + \frac{\left(\frac{M g_c(m)}{\gamma}\right)^{\frac{\gamma}{M(g_c(m)-1)-1}}}{\frac{\gamma}{M(g_c(m)-1)-1}+1} \frac{M g_c(m)}{\gamma} \right) - \frac{\gamma}{(m+1)^{1-\gamma}} \\ &= \gamma^{\gamma} \left(\frac{M g_c(m)}{m}\right)^{1-\gamma} + o\left(\left(\frac{M g_c(m)}{m}\right)^{1-\gamma}\right), \end{aligned} \tag{65}$$

where (a) is because $m^* = \frac{Mg_c(m)}{\gamma}$. and

$$\begin{aligned}
p_u^c &\geq \frac{1}{\frac{1}{1-\gamma}m^{1-\gamma} - \frac{1}{1-\gamma} + 1} \left(\left(\frac{1}{1-\gamma}(m^*+1)^{1-\gamma} - \frac{1}{1-\gamma} \right) - m^{*(-\gamma)} \sum_{f=1}^{m^*} \left(\frac{f}{m^*} \right)^{\frac{\gamma}{M(g_c(m)-1)}} \right) \\
&\geq \frac{1-\gamma}{m^{1-\gamma}-\gamma} \cdot \left(\frac{1}{1-\gamma} m^{*1-\gamma} - m^{*(-\gamma)} \sum_{f=1}^{m^*} \left(\frac{f}{m^*} \right)^{\frac{\gamma}{M(g_c(m)-1)}} - \frac{1}{1-\gamma} \right) \\
&\overset{(a)}{=} \frac{1-\gamma}{m^{1-\gamma}-\gamma} \left(\frac{1}{1-\gamma} \left(\frac{Mg_c(m)}{\gamma} \right)^{1-\gamma} - \left(\frac{Mg_c(m)}{\gamma} \right)^{(-\gamma)} \right. \\
&\quad \left. \cdot \sum_{f=1}^{\frac{Mg_c(m)}{\gamma}} \left(\frac{f}{\frac{Mg_c(m)}{\gamma}} \right)^{\frac{\gamma}{M(g_c(m)-1)-1}} - \frac{1}{1-\gamma} \right) \\
&\geq \gamma^{\gamma-1} \left(\frac{Mg_c(m)}{m} \right)^{1-\gamma} \left(1 + \frac{\gamma}{m^{1-\gamma}-\gamma} \right) - (1-\gamma)\gamma^{\gamma} \left(\frac{Mg_c(m)}{m} \right)^{-\gamma} \\
&\quad \cdot \left(\frac{m^{-\gamma}}{m^{1-\gamma}-\gamma} \right) \left(\frac{\gamma}{Mg_c(m)} \right)^{\frac{\gamma}{M(g_c(m)-1)-1}} \int_1^{\frac{Mg_c(m)}{\gamma}+1} x^{\frac{\gamma}{M(g_c(m)-1)-1}} dx - \frac{1}{m^{1-\gamma}-\gamma} \\
&= \gamma^{\gamma-1} \left(\frac{Mg_c(m)}{m} \right)^{1-\gamma} \left(1 + \frac{\gamma}{m^{1-\gamma}-\gamma} \right) - (1-\gamma)\gamma^{\gamma} \left(\frac{Mg_c(m)}{m} \right)^{-\gamma} \\
&\quad \cdot \left(\frac{m^{-\gamma}}{m^{1-\gamma}-\gamma} \right) \left(\frac{\gamma}{Mg_c(m)} \right)^{\frac{\gamma}{M(g_c(m)-1)-1}} \frac{1}{\frac{\gamma}{M(g_c(m)-1)-1}+1} \\
&\quad \cdot \left(\left(\frac{Mg_c(m)}{\gamma} + 1 \right)^{\frac{\gamma}{M(g_c(m)-1)-1}} \frac{Mg_c(m)}{\gamma} + \left(\frac{Mg_c(m)}{\gamma} + 1 \right)^{\frac{\gamma}{M(g_c(m)-1)-1}} - 1 \right) \\
&\quad - \frac{1}{m^{1-\gamma}-\gamma} \\
&= \gamma^{\gamma} \left(\frac{Mg_c(m)}{m} \right)^{1-\gamma} + o\left(\left(\frac{Mg_c(m)}{m} \right)^{1-\gamma} \right). \qquad (66)
\end{aligned}$$

where (a) is because $m^* = \frac{Mg_c(m)}{\gamma}$.

Therefore, by using (65) and (66), we obtain

$$p_u^c = \gamma^{\gamma} \left(\frac{Mg_c(m)}{m} \right)^{1-\gamma} + o\left(\left(\frac{Mg_c(m)}{m} \right)^{1-\gamma} \right). \qquad (67)$$

By using (67), we have

$$\mathbb{E}[\mathsf{W}] = \gamma^{\gamma} g_c(m) \left(\frac{Mg_c(m)}{m} \right)^{1-\gamma} + o\left(g_c(m) \left(\frac{Mg_c(m)}{m} \right)^{1-\gamma} \right). \qquad (68)$$

Now, since here we deal with an achievability strategy and we can choose the clustering strategy at wish, we choose $g_c(m) = c_2 m^{\alpha}$. By Theorem 4, it follows that $m^* = c_1 m^{\alpha}$ with $\frac{c_1}{c_2} = \frac{M}{\gamma}$. Clearly, this

requires that $n \geq g_c(m) = c_2 m^\alpha$ for all sufficiently large n. Then, by using (68), as $m \to \infty$, we have

$$\begin{aligned}
\mathbb{E}[\mathsf{W}] &= \gamma^\gamma g_c(m) \left(\frac{M g_c(m)}{m}\right)^{1-\gamma} + o\left(g_c(m)\left(\frac{M g_c(m)}{m}\right)^{1-\gamma}\right) \\
&= \gamma^\gamma c_2 m^\alpha \left(\frac{M c_2 m^\alpha}{m}\right)^{1-\gamma} + o\left(c_2 m^\alpha \left(\frac{M c_2 m^\alpha}{m}\right)^{1-\gamma}\right) \\
&= \gamma c_1^{1-\gamma} c_2 - o(1). \qquad (69)
\end{aligned}$$

Next, we compute $\mathbb{E}[\mathsf{W}^2]$. Since

$$\begin{aligned}
\mathbb{E}[\mathsf{W}^2] &= \mathbb{E}\left[\left(\sum_{u=1}^{g_c(m)} \mathbf{1}_u\right)^2\right] \\
&= \mathbb{E}\left[\sum_{u=1}^{g_c(m)} \mathbf{1}_u\right] + \sum_{u=1}^{g_c(m)} \sum_{u'=1, u\neq u'}^{g_c(m)} \mathbb{E}[\mathbf{1}_u \mathbf{1}_{u'}] \\
&= g_c(m) p_u^c + g_c(m)(g_c(m)-1) p_{uu'}^c, \qquad (70)
\end{aligned}$$

then under the optimal caching distribution P_c^*, we need to compute $p_{uu'}^c$.

Let B_u^f be the event that user u requests file f and can find message f in its cluster, such that $\mathcal{F}_{g_c(m)}^u = \bigcup_{f=1}^m B_u^f$. Then, we can write

$$\begin{aligned}
p_{uu'}^c &= \mathbb{P}(\{\mathbf{1}_u = 1\} \cap \{\mathbf{1}_{u'} = 1\}) \\
&= \mathbb{P}\left(\left(\cup_{i=1}^m B_u^i\right) \cap \left(\cup_{j=1}^m B_{u'}^j\right)\right) \\
&= \mathbb{P}\left(\cup_{i=1}^m \cup_{j=1}^m \left(B_u^i \cap B_{u'}^j\right)\right) \\
&\overset{(a)}{=} \sum_{i=1}^m \sum_{j=1}^m \mathbb{P}\left(B_u^i \cap B_{u'}^j\right) \\
&= \sum_{i=1}^m \sum_{j=1}^m \mathbb{P}\left(B_u^i\right) \mathbb{P}\left(B_{u'}^j | B_u^i\right) \\
&= \sum_{i=1}^m \sum_{j=1, j\neq i}^m \mathbb{P}\left(B_u^i\right) \mathbb{P}\left(B_{u'}^j | B_u^i\right) + \sum_{i=1}^m \mathbb{P}\left(B_u^i\right) \mathbb{P}\left(B_{u'}^i | B_u^i\right) \\
&\leq \sum_{i=1}^m \sum_{j=1, j\neq i}^m \left(P_r(i)(1-(1-P_c(i))^{M(g_c(m)-1)})\right)\left(P_r(j)(1-(1-P_c(j))^{M(g_c(m)-1)-1})\right) \\
&\quad + \sum_{i=1}^m \left(P_r(i)(1-(1-P_c(i))^{M(g_c(m)-1)})\right) P_r(i)
\end{aligned}$$

$$= \sum_{i=1}^{m}\left(P_r(i)(1-(1-P_c(i))^{M(g_c(m)-1)})\right)\sum_{j=1,j\neq i}^{m}\left(P_r(j)(1-(1-P_c(j))^{M(g_c(m)-1)-1})\right)$$
$$+\sum_{i=1}^{m}\left(P_r(i)(1-(1-P_c(i))^{M(g_c(m)-1)})\right)P_r(i), \quad (71)$$

where (a) is because that $B_u^i \cap B_{u'}^j$ are disjoint for different pairs of (i,j). Replacing $P_c(f) = P_c^*(f)$ in (71), we can continue as

$$p_{uu'}^c \leq \sum_{i=1}^{m^*}\left(P_r(i)(1-(1-P_c^*(i))^{M(g_c(m)-1)})\right)\sum_{j=1,j\neq i}^{m^*}\left(P_r(j)(1-(1-P_c^*(j))^{M(g_c(m)-1)-1})\right)$$
$$+\sum_{i=1}^{m^*}\left(P_r(i)(1-(1-P_c^*(i))^{M(g_c(m)-1)})\right)P_r(i)$$
$$\leq \left(\sum_{i=1}^{m^*}\left(P_r(i)(1-(1-P_c^*(i))^{M(g_c(m)-1)})\right)\right)^2+\sum_{i=1}^{m^*}\left(P_r(i)^2(1-(1-P_c^*(i))^{M(g_c(m)-1)})\right)$$
$$\overset{(a)}{=} {p_u^c}^2+\sum_{i=1}^{m^*}\left(P_r(i)^2(1-(1-P_c^*(i))^{M(g_c(m)-1)})\right), \quad (72)$$

where (a) is because that $p_u^c = \sum_{i=1}^{m^*}\left(P_r(i)(1-(1-P_c^*(i))^{M(g_c(m)-1)})\right)$.

The second term in (72) can be upper bounded by the following lemma.

Lemma 5: $\sum_{i=1}^{m^*}\left(P_r(i)^2(1-(1-P_c^*(i))^{M(g_c(m)-1)})\right)$ upper bounded by $o\left({p_u^c}^2\right)$.

Proof: See Appendix H. ■

At this point we are ready to obtain a lower bound on $\mathbb{P}(\mathsf{W}>0)$ via Lemma 4 and (62). From (70) we can write

$$\mathbb{E}[\mathsf{W}^2] \leq g_c(m)p_u^c + g_c(m)^2 p_{uu'}^c$$
$$\leq g_c(m)p_u^c + g_c(m)^2\left({p_u^c}^2 + o\left({p_u^c}^2\right)\right). \quad (73)$$

Then, Lemma 4, (69) and (73) yield

$$\mathbb{P}(\mathsf{W}>0) \geq \frac{\mathbb{E}[\mathsf{W}]^2}{\mathbb{E}[\mathsf{W}^2]}$$
$$\geq \frac{(\gamma c_1^{1-\gamma}c_2)^2}{g_c(m)p_u^c + g_c(m)^2\left({p_u^c}^2 + o\left({p_u^c}^2\right)\right)}$$
$$\geq \frac{(\gamma c_1^{1-\gamma}c_2)^2}{\gamma c_1^{1-\gamma}c_2 + (\gamma c_1^{1-\gamma}c_2)^2 + o(1)}$$
$$\overset{(a)}{=} \frac{\gamma^\gamma M^{1-\gamma}c_2^{2-\gamma}}{1+\gamma^\gamma M^{1-\gamma}c_2^{2-\gamma}} + o(1), \quad (74)$$

where (a) is because that we pick $\frac{c_1}{c_2} = \frac{M}{\gamma}$.

By using (60), we obtain

$$\begin{aligned}\mathbb{E}[\text{number of good clusters}] &= \frac{n}{g_c(m)} \cdot \mathbb{P}(\mathsf{W} > 0) \\ &\geq \frac{n}{c_2 m^\alpha} \cdot \frac{\gamma^\gamma M^{1-\gamma} c_2^{2-\gamma}}{1+\gamma^\gamma M^{1-\gamma} c_2^{2-\gamma}} + o\left(\frac{n}{m^\alpha}\right) \\ &= \frac{n}{m^\alpha} \frac{\gamma^\gamma M^{1-\gamma} c_2^{1-\gamma}}{1+\gamma^\gamma M^{1-\gamma} c_2^{2-\gamma}} + o\left(\frac{n}{m^\alpha}\right) \\ &= \frac{n}{m^\alpha} \frac{a c_2^{1-\gamma}}{1+a c_2^{2-\gamma}} + o\left(\frac{n}{m^\alpha}\right), \end{aligned} \tag{75}$$

where $a = \gamma^\gamma M^{1-\gamma}$. Since $m^* = \frac{M}{\gamma} g_c(m) = \frac{c_2 M}{\gamma} m^\alpha$ by Theorem 4, then as $m \to \infty$, by using (65) and (66), the corresponding average outage probability is given by

$$\begin{aligned} p_o &= 1 - p_u^c \\ &= 1 - \gamma^\gamma M^{1-\gamma} c_2^{1-\gamma} m^{-\alpha} + o\left(m^{-\alpha}\right) \\ &= 1 - a c_2^{1-\gamma} m^{-\alpha} + o\left(m^{-\alpha}\right). \end{aligned} \tag{76}$$

Therefore, we have

$$\overline{T}_{\text{sum}} \geq \frac{C}{K} \frac{a c_2^{1-\gamma}}{1+a c_2^{2-\gamma}} \frac{n}{m^\alpha} + o\left(\frac{n}{m^\alpha}\right). \tag{77}$$

By the symmetry of the system and of the caching and transmission policies Π_c^* and Π_l^*, the achievable throughput is lower bounded by

$$\overline{T}_{\min} = \frac{1}{n}\overline{T}_{\text{sum}} \geq \frac{C}{K} \frac{a c_2^{1-\gamma}}{1+a c_2^{2-\gamma}} \frac{1}{m^\alpha} + o\left(\frac{1}{m^\alpha}\right). \tag{78}$$

Next, we wish to find c_2 that maximizes the coefficient $\frac{a c_2^{1-\gamma}}{1+a c_2^{2-\gamma}}$ in the throughput lower bound (78). Setting the derivative to zero and looking for a maximum point, we find the unique solution $c_2 = b = \left(\frac{1-\gamma}{a}\right)^{\frac{1}{2-\gamma}}$. Let $D = \frac{a b^{1-\gamma}}{1+a b^{2-\gamma}}$, by using (78), we have

$$\overline{T}_{\min} \geq \frac{CD}{K} \frac{1}{m^\alpha} + o\left(\frac{1}{m^\alpha}\right), \tag{79}$$

with outage probability

$$\begin{aligned} p_o &= 1 - p_u^c \\ &= 1 - a b^{1-\gamma} m^{-\alpha} + o\left(m^{-\alpha}\right). \end{aligned} \tag{80}$$

Letting $p_o^* = 1 - a b^{1-\gamma} m^{-1/\alpha}$, following a perturbation argument similar to Appendix J, we have that for all $p \geq p_o^*$,

$$T(p) = \frac{CD}{K} \frac{1}{m^\alpha} + o\left(m^{-\alpha}\right), \tag{81}$$

is achievable. Thus, we have proved the last regime in (14) in Theorem 5.

B. Achievable $T(p)$ for $p < p_o^*$

By choosing a throughput-suboptimal value $c_2 = \rho_2 > b$ in (76) and (78), we have that for $p_o = 1 - a\rho_2^{1-\gamma} m^{-\alpha} \leq p \leq p_o^*$, then

$$T(p) = \frac{CB}{K}\frac{1}{m^\alpha} + o\left(m^{-\alpha}\right), \tag{82}$$

with $B = \frac{a\rho_2^{1-\gamma}}{1+a\rho_2^{2-\gamma}}$, is achievable. This yields the third regime in Theorem 5.

Next, we turn our attention to the case of $p_o = 1 - \omega\left(m^{-\alpha}\right)$. This is obtained by increasing the cluster size in order to decrease the outage probability and correspondingly decrease the throughput. As before, we find expressions for p_o and lower bounds on $\overline{T}_{\min}$ as a function of $g_c(m)$. We consider two cases for the value of $g_c(m)$. One is when $g_c(m) = \omega\left(\frac{n}{m^\alpha}\right)$ and $g_c(m) \leq \gamma m/M$. The other is when $g_c(m) = \rho_1 m/M$, where $\rho_1 \geq \gamma$.

1) Case $g_c(m) = \omega\left(\frac{n}{m^\alpha}\right)$ and $g_c(m) \leq \gamma m/M$: In this case, the cluster size is so large that $\mathbb{P}(\mathsf{W} > 0) \to 1$ as $m \to \infty$. In order to show this, we shall show that for arbitrary $\varepsilon_1 > 0$, with high probability $\mathsf{W} \in [(1-\varepsilon_1)\mathbb{E}[\mathsf{W}], (1+\varepsilon_1)\mathbb{E}[\mathsf{W}]]$ with $\mathbb{E}[\mathsf{W}] \to \infty$ as $m \to \infty$. This will be proved using Chebyshev's Inequality, which requires the computation of $\mathbb{E}[\mathsf{W}]$ and $\mathrm{Var}[\mathsf{W}]$. By using (68), we obtain $\mathbb{E}[\mathsf{W}]$. Since $g_c(m) = \omega\left(m^\alpha\right)$, then $\lim_{m\to\infty} \mathbb{E}[\mathsf{W}] = \infty$.

Next, we need to compute

$$\begin{aligned}
\mathrm{Var}[\mathsf{W}] &= \mathbb{E}[\mathsf{W}^2] - \mathbb{E}[\mathsf{W}]^2 \\
&= \mathbb{E}\left[\sum_{u=1}^{g_c(m)} \mathbf{1}_u\right] + \sum_{u=1}^{g_c(m)} \sum_{u'=1, u\neq u'}^{g_c(m)} \mathbb{E}[\mathbf{1}_u \mathbf{1}_{u'}] - \left(\sum_{u=1}^{g_c(m)} \mathbb{E}[\mathbf{1}_u]\right)^2 \\
&= g_c(m)p_u^c + g_c(m)(g_c(m)-1)p_{uu'}^c - g_c(m)^2 {p_u^c}^2 \\
&= g_c(m)(p_u^c - p_{uu'}^c) + g_c(m)^2(p_{uu'}^c - {p_u^c}^2).
\end{aligned} \tag{83}$$

We focus now on the term $p_{uu'}^c$, which is given by the following lemma.

Lemma 6: $p_{uu'}^c$ is given by:

$$p_{uu'}^c \leq \gamma^{2\gamma}\left(\frac{Mg_c(m)}{m}\right)^{2(1-\gamma)} + o\left(\left(\frac{Mg_c(m)}{m}\right)^{2(1-\gamma)}\right). \tag{84}$$

Proof: See Appendix I. ■

Therefore, as $m \to \infty$, by using Lemma 6, we have

$$\begin{aligned}
\mathrm{Var}[\mathsf{W}] &= g_c(m)(p_u^c - p_{uu'}^c) + g_c(m)^2(p_{uu'}^c - p_u^{c\,2}) \\
&\leq g_c(m)\left(\gamma^\gamma \left(\frac{Mg_c(m)}{m}\right)^{1-\gamma} - \gamma^{2\gamma}\left(\frac{Mg_c(m)}{m}\right)^{2(1-\gamma)}\right) \\
&\quad + g_c(m)^2\left(\gamma^{2\gamma}\left(\frac{Mg_c(m)}{m}\right)^{2(1-\gamma)} - \gamma^{2\gamma}\left(\frac{Mg_c(m)}{m}\right)^{2(1-\gamma)}\right) \\
&\quad + o\left(g_c(m)\left(\frac{Mg_c(m)}{m}\right)^{1-\gamma}\right) \\
&= \gamma^\gamma g_c(m)\left(\frac{Mg_c(m)}{m}\right)^{1-\gamma} + o\left(g_c(m)\left(\frac{Mg_c(m)}{m}\right)^{1-\gamma}\right). \qquad (85)
\end{aligned}$$

Thus, by using (68) and (85) into Chebyshev's Inequality, it is not difficult to show that, for any $\varepsilon_1 > 0$,

$$\mathbb{P}\left(|\mathsf{W} - \mathbb{E}[\mathsf{W}]| \leq \varepsilon_1 \mathbb{E}[\mathsf{W}]\right) \geq 1 - o\left(1\right). \qquad (86)$$

as $m \to \infty$.

Since, as observed before, $\lim_{m\to\infty} \mathbb{E}[\mathsf{W}] = \infty$, we conclude that for any $0 < \gamma < 1$, as $m \to \infty$, $\mathbb{P}(\mathsf{W} > 0) = 1 - o(1)$. It follows that all clusters are good, such that

$$\overline{T}_{\mathrm{sum}} = \frac{C}{K}\frac{n}{g_c(m)} + o\left(\frac{n}{g_c(m)}\right). \qquad (87)$$

As $m \to \infty$, by using (65) and (66), the corresponding outage probability is given by

$$\begin{aligned}
p_o &= 1 - p_u^c \\
&= 1 - \gamma^\gamma\left(\frac{Mg_c(m)}{m}\right)^{1-\gamma} + o\left(\left(\frac{Mg_c(m)}{m}\right)^{1-\gamma}\right). \qquad (88)
\end{aligned}$$

By the usual symmetry argument, we have

$$\overline{T}_{\mathrm{min}} = \frac{1}{n}\overline{T}_{\mathrm{sum}} = \frac{C}{K}\frac{1}{g_c(m)} + o\left(\frac{1}{g_c(m)}\right). \qquad (89)$$

Finally, letting $p = p_o$, we can solve for $g_c(m) = \frac{1}{\gamma^{\frac{\gamma}{1-\gamma}}}\frac{m}{M}(1-p)^{\frac{1}{1-\gamma}} + o\left(\frac{m}{M}(1-p)^{\frac{1}{1-\gamma}}\right)$. By using (89) and letting $A = \gamma^{\frac{\gamma}{1-\gamma}}$, with the similar perturbation argument shown in Appendix J, we have that when $p = 1 - \gamma^\gamma\left(\frac{Mg_c(m)}{m}\right)^{1-\gamma}$, then

$$T(p) = \frac{CA}{K}\frac{M}{m(1-p)^{\frac{1}{1-\gamma}}} + o\left(\frac{M}{m(1-p)^{\frac{1}{1-\gamma}}}\right) \qquad (90)$$

is achievable. This settles the second regime of Theorem 1.

2) Case $g_c(m) = \rho_1 m/M$, where $\rho_1 \geq \gamma$: In this case, by using Theorem 4, we have $m^* = m$. We can obtain that $\mathbb{P}(\mathsf{W} > 0) = 1 - o(1)$ as $m \to \infty$. Thus, we have

$$\overline{T}_{\min} = \frac{1}{n}\frac{C}{K}\frac{n}{g_c(m)} = \frac{C}{K}\frac{M}{\rho_1 m} + o\left(\frac{M}{m}\right). \tag{91}$$

The corresponding outage probability is computed next. Here, we need to find a different bounding technique other than the one we used before. In this case, we directly plug $\nu = \frac{m^*-1}{\sum_{j=1}^{m^*}\frac{1}{z_{f_j}}}$ into p_u^c and use the integral approximations of summations to obtain the lower bound of p_u^c, instead of using that fact that $\nu \leq z_{m^*}$ and $\nu \geq z_{m^*+1}$ as we used before. The reason is that in this case, $m^* = m$ as shown by Theorem 4, which means that $m^* \neq \frac{Mg_c(m)}{\gamma}$ when $\rho_1 > \gamma$, this makes z_{m^*} and z_{m^*+1} not good approximations anymore.

Operating along these lines, p_u^c can be computed as

$$\begin{aligned}
p_u^c &= \sum_{f=1}^{m^*} P_r(f)\left(1 - \left(\frac{\nu}{z_f}\right)^{M(g_c(m)-1)}\right) \\
&= \sum_{f=1}^{m} P_r(f)\left(1 - \left(\frac{\nu}{z_f}\right)^{M(g_c(m)-1)}\right) \\
&= 1 - \nu^{M(g_c(m)-1)}\sum_{f=1}^{m}\frac{P_r(f)}{z_f^{M(g_c(m)-1)}} \\
&= 1 - \nu^{M(g_c(m)-1)}\sum_{f=1}^{m}\frac{P_r(f)}{P_r(f)^{\frac{M(g_c(m)-1)}{M(g_c(m)-1)-1}}} \\
&= 1 - \left(\frac{m-1}{\sum_{i=1}^{m}\frac{1}{z_i}}\right)^{M(g_c(m)-1)}\sum_{f=1}^{m} P_r(f)^{-\frac{1}{M(g_c(m)-1)-1}} \\
&= 1 - \left(\frac{m-1}{\sum_{i=1}^{m} P_r(i)^{-\frac{1}{M(g_c(m)-1)-1}}}\right)^{M(g_c(m)-1)}\sum_{f=1}^{m} P_r(f)^{-\frac{1}{M(g_c(m)-1)-1}} \\
&= 1 - (m-1)^{M(g_c(m)-1)}\left(\sum_{f=1}^{m} P_r(f)^{-\frac{1}{g_c(m)-2}}\right)^{-(g_c(m)-2)} \\
&= 1 - (m-1)^{M(g_c(m)-1)}\left(\sum_{f=1}^{m}\left(\frac{f^{-\gamma}}{H(\gamma,1,m)}\right)^{-\frac{1}{M(g_c(m)-1)-1}}\right)^{-(M(g_c(m)-1)-1)} \\
&= 1 - \frac{(m-1)^{M(g_c(m)-1)}}{H(\gamma,1,m)}\frac{1}{\left(\sum_{f=1}^{m} f^{\frac{\gamma}{M(g_c(m)-1)-1}}\right)^{M(g_c(m)-1)-1}}.
\end{aligned} \tag{92}$$

The lower bound of p_u^c is given by

$$\begin{aligned}
p_u^c &\geq 1-\frac{(m-1)^{M(g_c(m)-1)}}{\frac{1}{1-\gamma}(m+1)^{1-\gamma}-\frac{1}{1-\gamma}}\cdot\frac{1}{\left(1+\int_1^m x^{\frac{\gamma}{M(g_c(m)-1)-1}}dx\right)^{M(g_c(m)-1)-1}} \\
&= 1-\frac{(m-1)^{M(g_c(m)-1)}}{\frac{1}{1-\gamma}(m+1)^{1-\gamma}-\frac{1}{1-\gamma}}\cdot\frac{1}{\left(1+\frac{1}{1+\frac{\gamma}{M(g_c(m)-1)-1}}(m^{\frac{\gamma}{M(g_c(m)-1)-1}+1}-1)\right)^{M(g_c(m)-1)-1}} \\
&= 1-(1-\gamma)\frac{(m-1)^{M(g_c(m)-1)}}{(m+1)^{1-\gamma}-1}\frac{1}{\left(\frac{1}{1+\frac{\gamma}{M(g_c(m)-1)-1}}m^{\frac{\gamma}{M(g_c(m)-1)-1}+1}+1-\frac{1}{\frac{\gamma}{M(g_c(m)-1)-1}+1}\right)^{M(g_c(m)-1)-1}} \\
&= 1-(1-\gamma)\frac{(m-1)^{M(g_c(m)-1)}}{m^{1-\gamma}}\frac{m^{1-\gamma}}{(m+1)^{1-\gamma}-1}\frac{1}{m^{M(g_c(m)-1)-1+\gamma}} \\
&\quad\cdot\frac{m^{M(g_c(m)-1)-1+\gamma}}{\left(\frac{1}{1+\frac{\gamma}{M(g_c(m)-1)-1}}m^{\frac{\gamma}{M(g_c(m)-1)-1}+1}+1-\frac{1}{\frac{\gamma}{M(g_c(m)-1)-1}+1}\right)^{M(g_c(m)-1)-1}} \\
&= 1-(1-\gamma)\frac{(m-1)^{M(g_c(m)-1)}}{m^{1-\gamma}}\frac{m^{1-\gamma}}{(m+1)^{1-\gamma}-1}\frac{1}{m^{M(g_c(m)-1)-1+\gamma}}\frac{1}{\left(\frac{1}{1+\frac{\gamma}{M(g_c(m)-1)-1}}\right)^{M(g_c(m)-1)-1}} \\
&\quad\cdot\frac{m^{M(g_c(m)-1)-1+\gamma}}{\left(m^{\frac{\gamma}{M(g_c(m)-1)-1}+1}+\frac{\gamma}{M(g_c(m)-1)-1}\right)^{M(g_c(m)-1)-1}} \\
&= 1-(1-\gamma)\left(1-\frac{1}{m}\right)^{M(g_c(m)-1)}\frac{1}{\left(1-\frac{\gamma}{M(g_c(m)-1)-1+\gamma}\right)^{\frac{M(g_c(m)-1)-1+\gamma}{\gamma}\frac{M(g_c(m)-1)-1}{M(g_c(m)-1)-1+\gamma}\gamma}} \\
&\quad\cdot\frac{m^{1-\gamma}}{(m+1)^{1-\gamma}-1}\frac{m^{M(g_c(m)-1)-1+\gamma}}{\left(m^{\frac{\gamma}{M(g_c(m)-1)-1}+1}+\frac{\gamma}{M(g_c(m)-1)-1}\right)^{M(g_c(m)-1)-1}} \\
&= 1-(1-\gamma)\left(1-\frac{1}{m}\right)^{M(\rho_1 m/M-1)}\frac{1}{\left(1-\frac{\gamma}{M(g_c(m)-1)-1+\gamma}\right)^{\frac{M(g_c(m)-1)-1+\gamma}{\gamma}\frac{M(g_c(m)-1)-1}{M(g_c(m)-1)-1+\gamma}\gamma}}(1+o(1)) \\
&= 1-(1-\gamma)\left(\frac{1}{e}\right)^{\rho_1}\frac{1}{\left(\frac{1}{e}\right)^{\gamma}}(1+o(1)) \\
&= 1-(1-\gamma)\left(\frac{1}{e}\right)^{\rho_1-\gamma}(1+o(1))\,. \qquad (93)
\end{aligned}$$

Thus, we have

$$\begin{aligned}
p_o &= 1-p_u^c \\
&\leq 1-\left(1-(1-\gamma)\left(\frac{1}{e}\right)^{\rho_1-\gamma}(1+o(1))\right) \\
&= (1-\gamma)\left(\frac{1}{e}\right)^{\rho_1-\gamma}(1+o(1))\,. \qquad (94)
\end{aligned}$$

Therefore, letting $p = (1-\gamma)e^{\gamma-\rho_1}$, and following a perturbation argument similar to Appendix J, we have that the throughput

$$T(p) = \frac{C}{K}\frac{M}{\rho_1 m} + o\left(\frac{M}{m}\right) \tag{95}$$

is achievable. This settles the first regime of Theorem 5.

APPENDIX E
PROOF OF LEMMA 1

When $\gamma \neq 1$, then, since $\frac{1}{x^\gamma}$ is an decreasing function, we have

$$\begin{aligned} H(\gamma, x, y) &= \sum_{i=x}^{y} \frac{1}{i^\gamma} \geq \int_a^{b+1} \frac{1}{x'^\gamma} dx' \\ &= \frac{1}{1-\gamma}(y+1)^{1-\gamma} - \frac{1}{1-\gamma}x^{1-\gamma}, \end{aligned} \tag{96}$$

and

$$\begin{aligned} H(\gamma, x, y) &= \sum_{i=x}^{y} \frac{1}{i^\gamma} = \frac{1}{x^\gamma} + \sum_{i=x-1}^{y} \frac{1}{i^\gamma} \\ &\leq \int_{x-1+1}^{y} \frac{1}{x'^\gamma} dx' + \frac{1}{x^\gamma} \\ &= \frac{1}{1-\gamma}y^{1-\gamma} - \frac{1}{1-\gamma}x^{1-\gamma} + \frac{1}{x^\gamma}. \end{aligned} \tag{97}$$

When $\gamma = 1$, similarly, since $\frac{1}{x}$ is an decreasing function, we have

$$\begin{aligned} H(\gamma, x, y) &= \sum_{i=x}^{y} \frac{1}{i} \geq \int_x^{y+1} \frac{1}{x'} dx' \\ &= \log(y+1) - \log(x), \end{aligned} \tag{98}$$

and

$$\begin{aligned} H(\gamma, x, y) &= \sum_{i=x}^{y} \frac{1}{i} = \frac{1}{x} + \sum_{i=x-1}^{y} \frac{1}{i} \\ &\leq \int_{x-1+1}^{y} \frac{1}{x'} dx' + \frac{1}{x} \\ &= \log(y) - \log(x) + \frac{1}{x}. \end{aligned} \tag{99}$$

APPENDIX F
PROOF OF LEMMA 2

Recall that we denote the disks of radius $\frac{\Delta}{2}R$ centered around the receivers as "disk", our goal is to show that

$$\mathbb{P}\left(\text{Any disk} \cap U(R,\Delta,\mathsf{L})\right) \leq \mathbb{P}\left(\exists \text{ an active receiver in a disk of radius } \left(1+\tfrac{3\Delta}{2}\right)R\right). \tag{100}$$

which is equivalent to show that

$$\begin{aligned}&\{\text{Any disk} \cap U(R,\Delta,\mathsf{L})\}\\ &\subseteq \{\exists \text{ an active receiver in a disk of radius } \left(1+\tfrac{3\Delta}{2}\right)R\}.\end{aligned} \tag{101}$$

To see (101), we first consider a simple illustration which is easier to explain, but it is not accurate. As shown in Fig. 7, the network is divided into squarelets whose diagonal are $\frac{\Delta}{2}R$. These squarelets are the analogue to the the sectors with radius of $\frac{\Delta}{2}R$ (a quarter of disk with radius $\frac{\Delta}{2}R$). Now we want to see which events can cause a squarelet to intersect with $U(R,\Delta,\mathsf{L})$, which means that the area of this squarelet is consumed due to communicating links according to the protocol model. From Fig. 7, we can see that if the link from user u' to u is activated, then the upper bound of the maximum area this link can consume is the area of all the blue squarelets. If we consider squarelet A and let user v be a receiver, we can see that A cannot intersect with $U(R,\Delta,\mathsf{L})$ if there is no any active receiver in a disk centered at v, with radius $\left(1+\frac{3\Delta}{2}\right)R$. Therefore, if there is at least one active receiver in a disk centered at v, with radius $\left(1+\frac{3\Delta}{2}\right)R$, then it is possible that A can intersect with $U(R,\Delta,\mathsf{L})$.

Now we prove this lemma accurately. From the arguments in Appendix A, we know that all the disks with radius of $\frac{\Delta}{2}R$ have to be disjoint. Moreover, there is at least a fraction $\frac{1}{4}$ of the area of such disks inside the network. Therefore, to obtain an upper bound of the maximum concurrent transmissions, we maximumly pack such sectors[7] inside the network as shown in Fig 8 (Of course, Fig 8 shows an over optimistic way of packing, since we cannot guarantee all the disks with radius $\frac{\Delta}{2}R$ are disjoint. However, at least all the sectors are disjoint.). Notice that

$$\{\text{Any disk} \cap U(R,\Delta,\mathsf{L})\} \subseteq \{\text{Any sector} \cap U(R,\Delta,\mathsf{L})\}. \tag{102}$$

Now we consider each such sector as an analogue of the squarelet considered before. This shows that if the receiver u is activated, then the upper bound of the maximum number of sectors that can intersect

[7]In the following, we denote the sector that is $\frac{1}{4}$ of the disk with radius of $\frac{\Delta}{2}R$ as "sector".

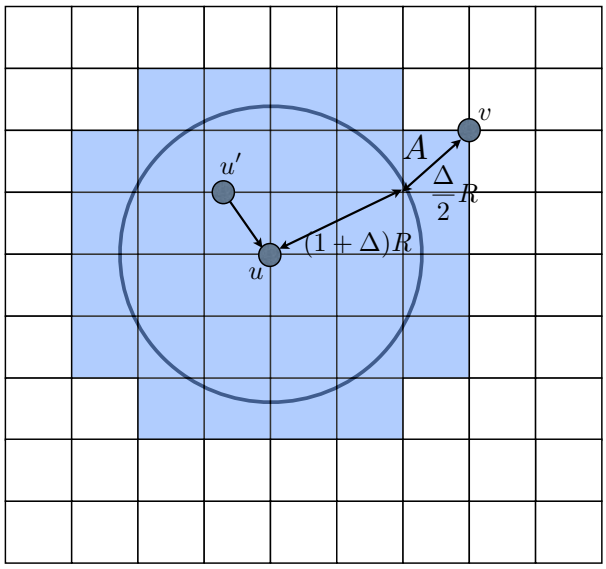


Fig. 7. In this figure, u' is a transmitter and u is a receiver. v is another receiver corresponding to another transmitter. The diagonal of each squarelet is $\frac{\Delta}{2}R$. The maximum area that are consumed by receiver u is the disk centered at u, with radius $(1+\Delta)R$. The blue squarelets are the maximum activated squarelets that are caused by the active receiver u. A indicates the squarelet containing receiver v.

with $U(R,\Delta,\mathsf{L})$ are the blue sectors. Now pick a arbitrary node v, if there is no any active receiver inside a disk centered at v of radius $\left(1+\frac{3\Delta}{2}\right)R$, then the sector A cannot intersect with $U(R,\Delta,\mathsf{L})$, which means that if there is at least one active receiver inside a disk centered at v of radius $\left(1+\frac{3\Delta}{2}\right)R$, then the sector A may intersect with $U(R,\Delta,\mathsf{L})$. Since v is arbitrary, then

$$\begin{aligned}&\{\text{Any sector}\cap U(R,\Delta,\mathsf{L})\}\\ &\subseteq\{\exists\text{ an active receiver in a disk of radius }\left(1+\tfrac{3\Delta}{2}\right)R\}.\end{aligned}\tag{103}$$

By using (102) and (103), (101) is proved.

APPENDIX G
PROOF OF LEMMA 3

Using (20), we are interested in the quantity

$$\left(1-(p^{\text{lb}}(g))^{\left(1+\frac{3\Delta}{2}\right)^2 g}\right)\frac{n}{g} \;=\; \left(1-\left(1-\frac{\frac{1}{1-\gamma}(Mg)^{1-\gamma}-\frac{1}{1-\gamma}+1}{\frac{1}{1-\gamma}m^{1-\gamma}-\frac{1}{1-\gamma}}\right)^{\left(1+\frac{3\Delta}{2}\right)^2 g}\right)\frac{n}{g}.\tag{104}$$

We consider three regimes for g, namely, $g=o(m^\alpha)$, $g=\omega(m^\alpha)$ and $g=\Theta(m^\alpha)=\rho m^\alpha$.

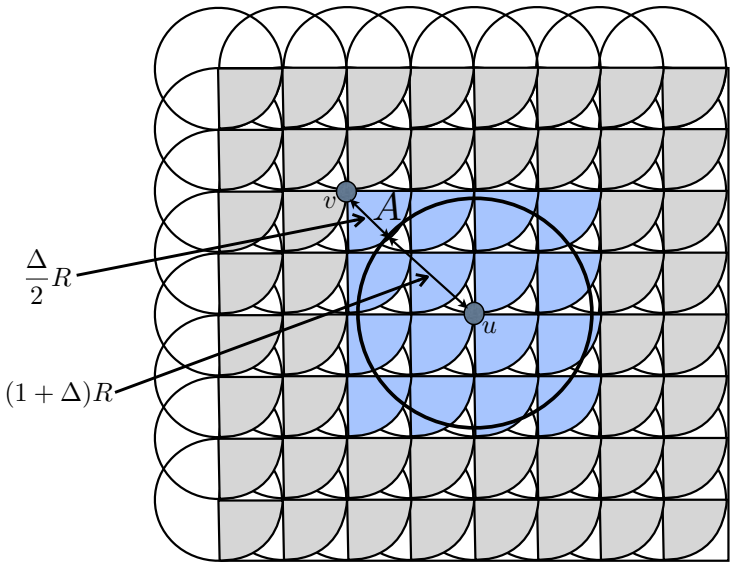


Fig. 8. In this figure, u and v are receivers. The radius of each grey sector is $\frac{\Delta}{2}R$. Each grey sector is $\frac{1}{4}$ of each disk with radius $\frac{\Delta}{2}R$. The maximum area that are consumed by receiver u is the disk centered at u, with radius $(1+\Delta)R$. The blue sectors are the maximum activated sectors that are caused by the active receiver u. A indicates the sector containing receiver v.

When $g = o(m^\alpha)$, by using (104), we have

$$\begin{aligned}\left(1-(p^{\mathrm{lb}}(g))^{\left(1+\frac{3\Delta}{2}\right)^2 g}\right)\frac{n}{g} &\overset{(a)}{\leq} \left(1-\left(1-\left(1+\frac{3\Delta}{2}\right)^2 g\frac{\frac{1}{1-\gamma}(Mg)^{1-\gamma}-\frac{1}{1-\gamma}+1}{\frac{1}{1-\gamma}m^{1-\gamma}-\frac{1}{1-\gamma}}\right)\right)\frac{n}{g}\\ &= \left(1+\frac{3\Delta}{2}\right)^2 M^{1-\gamma}n\left(\frac{g}{m}\right)^{1-\gamma}+o\left(n\left(\frac{g}{m}\right)^{1-\gamma}\right)\\ &= o\left(\frac{n}{m^\alpha}\right), \end{aligned} \tag{105}$$

where (a) is because that $(1-x)^t \geq 1-tx$ for $0 \leq x \leq 1$ and $t \geq 1$.

When $g = \omega(m^\alpha)$, by using (104), we obtain

$$\left(1-(p^{\mathrm{lb}}(g))^{\left(1+\frac{3\Delta}{2}\right)^2 g}\right)\frac{n}{g} \overset{(a)}{\leq} \frac{n}{g} = o\left(\frac{n}{m^\alpha}\right), \tag{106}$$

where (a) is because that $\left(1-(p^{\mathrm{lb}}(g))^{\left(1+\frac{3\Delta}{2}\right)^2 g}\right) \leq 1$.

When $g = \rho m^\alpha$, by using (104), we get

$$\begin{aligned}\left(1-(p^{\mathrm{lb}}(g))^{\left(1+\frac{3\Delta}{2}\right)^2 g}\right)\frac{n}{g} &= \frac{n}{\rho m^\alpha}\left(1-\left(1-\rho^{1-\gamma}M^{1-\gamma}m^{(\alpha-1)(1-\gamma)}\right)^{\left(1+\frac{3\Delta}{2}\right)^2\rho m^\alpha}\right)\\ &\overset{(a)}{=} \frac{n}{\rho m^\alpha}\left(1-\left(1-\rho^{1-\gamma}M^{1-\gamma}m^{-\alpha}\right)^{\left(1+\frac{3\Delta}{2}\right)^2\rho m^\alpha}\right)\\ &= \frac{n}{\rho m^\alpha}\left(1-\left(\left(1-\rho^{1-\gamma}M^{1-\gamma}m^{-\alpha}\right)^{\rho^{-(1-\gamma)}M^{-(1-\gamma)}m^\alpha}\right)^{\left(1+\frac{3\Delta}{2}\right)^2\rho^{2-\gamma}M^{1-\gamma}}\right)\end{aligned}$$

$$\begin{aligned}
&\overset{(b)}{=} \frac{1}{\rho}\left(1-\exp\left(-\left(1+\frac{3\Delta}{2}\right)^2\rho^{2-\gamma}M^{1-\gamma}\right)\right)\frac{n}{m^\alpha}\\
&= \left(1+\frac{3\Delta}{2}\right)^{\frac{2}{2-\gamma}}\frac{1}{\left(1+\frac{3\Delta}{2}\right)^{\frac{2}{2-\gamma}}\rho}\\
&\quad\cdot\left(1-\exp\left(-\left(\left(1+\frac{3\Delta}{2}\right)^{\frac{2}{2-\gamma}}\rho\right)^{2-\gamma}M^{1-\gamma}\right)\right)\frac{n}{m^\alpha}\\
&\overset{(c)}{=} \left(1+\frac{3\Delta}{2}\right)^{\frac{2}{2-\gamma}}\frac{1}{\tilde{\rho}}\left(1-\exp\left(-\tilde{\rho}^{2-\gamma}M^{1-\gamma}\right)\right)\frac{n}{m^\alpha}, \qquad (107)
\end{aligned}$$

where (a) follows by using $(\alpha-1)(1-\gamma)=-\alpha$; (b) follows because $\lim_{x\to\infty}(1-x^{-1})^x=e^{-1}$; (c) is obtained by defining $\tilde{\rho}=\left(1+\frac{3\Delta}{2}\right)^{\frac{2}{2-\gamma}}\rho$.

We conclude that $\left(1-(p^{\mathrm{lb}}(g))^{\left(1+\frac{3\Delta}{2}\right)^2 g}\right)\frac{n}{g}=O\left(\frac{n}{m^\alpha}\right)$ and, when $g=\rho m^\alpha$, then $\left(1-(p^{\mathrm{lb}}(g))^{\left(1+\frac{3\Delta}{2}\right)^2 g}\right)\frac{n}{g}=\Theta\left(\frac{n}{m^\alpha}\right)$. Now we compute the optimal constant $\tilde{\rho}$, which is shown in the following lemma.

Lemma 7: The optimal value of $\tilde{\rho}^*$ to maximize $\left(1-(p^{\mathrm{lb}}(g))^{\left(1+\frac{3\Delta}{2}\right)^2 g}\right)\frac{n}{g}$ is the solution of

$$\tilde{\rho}^{2-\gamma}M^{1-\gamma}=\log\left(1+(2-\gamma)\tilde{\rho}^{2-\gamma}M^{1-\gamma}\right).$$

Moreover, the solution satisfies $\tilde{\rho}^{*2-\gamma}M^{1-\gamma}>\alpha$, and Eq. $x=\log\left(1+(2-\gamma)x\right)$ is a fixed point equation and has a non-negative solution for $x>\alpha$.

Proof: From (107), we know that to maximize $\left(1-(p^{\mathrm{lb}}(g))^{\left(1+\frac{3\Delta}{2}\right)^2 g}\right)\frac{n}{g}$ we need to maximize $\frac{1}{\tilde{\rho}}\left(1-\exp(-\tilde{\rho}^{2-\gamma}M^{1-\gamma})\right)$. Differentiating this expression with respect to $\tilde{\rho}$, and equating to zero, we find

$$\tilde{\rho}^{2-\gamma}M^{1-\gamma}=\log\left(1+(2-\gamma)\tilde{\rho}^{2-\gamma}M^{1-\gamma}\right). \qquad (108)$$

This proves the first part of Lemma 7.

Then, by letting $x=\tilde{\rho}^{2-\gamma}M^{1-\gamma}$, we get

$$x=\log\left(1+(2-\gamma)x\right). \qquad (109)$$

Let $f(x)=\log\left(1+(2-\gamma)x\right)-x$. We observe that if $f(x)=0$, then there are two roots, one is 0 which must b excluded since $x=\rho^{2-\gamma}M^{1-\gamma}>0$ and the other root is greater than 0. Differentiating with respect to x, we find

$$\begin{aligned}
\frac{d}{dx}f(x) &= \frac{2-\gamma}{1+(2-\gamma)x}-1\\
&= \frac{(2-\gamma)(1-x)-1}{1+(2-\gamma)x}. \qquad (110)
\end{aligned}$$

We observe that $\frac{d}{dx}f(x) < 0$ for $x > \alpha$, $\frac{d}{dx}f(x) > 0$ for $x < \alpha$, and $\frac{d}{dx}f(x) = 0$ for $x = \alpha$. Thus, $f(x)$ achieves its maximum value when $x = \alpha$.

Now we can see that

$$\begin{aligned} f(\alpha) &= \log(1 + (2-\gamma)\alpha) - \alpha \\ &= \log(2-\gamma) - \alpha > 0, \end{aligned} \tag{111}$$

when $0 \leq \gamma < 1$. Thus, the positive root of $f(x) = 0$ is greater than α. This proves the second part of Lemma 7.

Let $\phi(x) = \log(1 + (2-\gamma)x)$, then if $\phi(x)$ is a contraction from $\mathbb{R}$ to $\mathbb{R}$, then we can show that $\phi(x) = x$ is a fixed point equation and can be solved by iterations numerically [43]. Therefore, we need to show when $x > \alpha$, $\phi(x)$ is a contraction from $\mathbb{R}$ to $\mathbb{R}$.

Let $x, y \in \mathbb{R}$. With out loss of generality we assume $x > y$. When $x > \alpha$, we get

$$\begin{aligned} |\phi(x) - \phi(y)| &= |\log(1 + (2-\gamma)x) - \log(1 + (2-\gamma)y)| \\ &= \left|\log\left(\frac{1 + (2-\gamma)x}{1 + (2-\gamma)y}\right)\right| \\ &= \left|\log\left(1 + \frac{2-\gamma}{1 + (2-\gamma)y}(x-y)\right)\right| \\ &\overset{(a)}{<} \frac{2-\gamma}{1 + (2-\gamma)y}|x-y| \\ &\overset{(b)}{<} k|x-y|, \end{aligned} \tag{112}$$

where $k = \frac{2-\gamma}{1+(2-\gamma)y} < 1$. (a) is because that $\log(1+x) < x$, when $x \neq 0$. (b) is because when $y > \alpha$, then $k = \frac{2-\gamma}{1+(2-\gamma)y} < 1$. Thus $\phi(x)$ is a contraction from $\mathbb{R}$ to $\mathbb{R}$ when $x > \alpha$. Therefore, we conclude that $\phi(x) = x$ is a fixed point equation for $x > \alpha$. ■

APPENDIX H
PROOF OF LEMMA 5

We have

$$\begin{aligned} \sum_{i=1}^{m^*}\left(P_r(i)^2(1-(1-P_c^*(i))^{M(g_c(m)-1)})\right) &\overset{(a)}{\leq} \sum_{i=1}^{m^*} P_r(i)^2\left(1 - \left(\frac{P_r(m^*+1)}{P_r(i)}\right)^{\frac{M(g_c(m)-1)}{M(g_c(m)-1)-1}}\right) \\ &= \sum_{i=1}^{m^*} P_r(i)^2 - \sum_{i=1}^{m^*} P_r(i)^2\left(\frac{P_r(m^*+1)}{P_r(i)}\right)^{\frac{M(g_c(m)-1)}{M(g_c(m)-1)-1}} \\ &= \sum_{i=1}^{m^*} P_r(i)^2 - P_r(m^*+1)\sum_{i=1}^{m^*} P_r(i)\left(\frac{P_r(m^*+1)}{P_r(i)}\right)^{\frac{1}{M(g_c(m)-1)-1}} \end{aligned}$$

$$= \frac{H(2\gamma,1,m^*)}{H(\gamma,1,m)^2} - \frac{(m^*+1)^{-\gamma}}{H(\gamma,1,m)^2}\left(\sum_{i=1}^{m^*} i^{-\gamma}\left(\frac{i}{m^*+1}\right)^{\frac{\gamma}{M(g_c(m)-1)-1}}\right), \quad (113)$$

where (a) is because $\nu \geq z_{m*+1}$.

Now, in order to compute an upper bound on (113), we consider the first and the second term separately. In order to upper bound the first term, we have to consider the cases of $\gamma \neq \frac{1}{2}$ and $\gamma = \frac{1}{2}$. For $\gamma \neq \frac{1}{2}$, by using Lemma 1, the first term in (113) can be upper bounded as:

$$\begin{aligned}
\frac{H(2\gamma,1,m^*)}{H(\gamma,1,m)^2} &\leq \frac{\frac{1}{1-2\gamma}m^{*1-2\gamma} - \frac{1}{1-2\gamma} + 1}{\left(\frac{1}{1-\gamma}(m+1)^{1-\gamma} - \frac{1}{1-\gamma}\right)^2} \\
&= \frac{(1-\gamma)^2}{1-2\gamma}\frac{m^{*1-2\gamma} - 2\gamma}{((m+1)^{1-\gamma}-1)^2} \\
&\leq \frac{(1-\gamma)^2}{1-2\gamma}\frac{m^{*1-2\gamma} - 2\gamma}{(m^{1-\gamma}-1)^2} \\
&= \frac{(1-\gamma)^2}{(1-2\gamma)}\frac{\left(\frac{Mg_c(m)}{\gamma}\right)^{1-2\gamma} - 2\gamma}{(m^{1-\gamma}-1)^2}. \quad (114)
\end{aligned}$$

For $\gamma = \frac{1}{2}$, by using Lemma 1, the first term in (113) can be upper bounded as:

$$\begin{aligned}
\frac{H(2\gamma,1,m^*)}{H(\gamma,1,m)^2} &= \frac{H(1,1,m^*)}{H(\frac{1}{2},1,m)^2} \\
&\leq \frac{\log(m^*)+1}{\left(\frac{1}{1-\frac{1}{2}}(m+1)^{1-\frac{1}{2}} - \frac{1}{1-\frac{1}{2}}\right)^2} \\
&= \frac{\log(Mg_c(m)) + \log 2 + 1}{\left(2(m+1)^{\frac{1}{2}} - 2\right)^2}. \quad (115)
\end{aligned}$$

By letting $g_c(m) = \frac{c_1\gamma m^\alpha}{M}$, we have

$$\frac{H(2\gamma,1,m^*)}{H(\gamma,1,m)^2} \leq \frac{\frac{1}{3}\log(m) + \log(c_1) + 1}{(2m^{1-\gamma}-2)^2}. \quad (116)$$

The second term in (113), for any $\gamma < 1$, can be lower bounded as:

$$\begin{aligned}
&\frac{(m^*+1)^{-\gamma}}{H(\gamma,1,m)^2}\left(\sum_{i=1}^{m^*} i^{-\gamma}\left(\frac{i}{m^*+1}\right)^{\frac{\gamma}{M(g_c(m)-1)-1}}\right) \\
\geq\ & \frac{(m^*+1)^{-\gamma}}{\left(\frac{1}{1-\gamma}m^{1-\gamma} - \frac{1}{1-\gamma} + 1\right)^2}\frac{1}{(m^*+1)^{\frac{\gamma}{M(g_c(m)-1)-1}}}\sum_{i=1}^{m^*} i^{-\frac{M(g_c(m)-1)-2}{M(g_c(m)-1)-1}\gamma} \\
\geq\ & \frac{(m^*+1)^{-\gamma}}{\left(\frac{1}{1-\gamma}m^{1-\gamma} - \frac{\gamma}{1-\gamma}\right)^2}\frac{1}{(m^*+1)^{\frac{\gamma}{M(g_c(m)-1)-1}}} \\
&\cdot \int_1^{m^*+1} x^{-\frac{M(g_c(m)-1)-2}{M(g_c(m)-1)-1}\gamma} dx
\end{aligned}$$

$$
\begin{aligned}
&= \frac{(m^*+1)^{-\gamma}}{\left(\frac{1}{1-\gamma}m^{1-\gamma}-\frac{\gamma}{1-\gamma}\right)^2}\frac{1}{(m^*+1)^{\frac{\gamma}{M(g_c(m)-1)-1}}} \\
&\quad \cdot\frac{1}{1-\frac{M(g_c(m)-1)-2}{M(g_c(m)-1)-1}\gamma}\left((m^*+1)^{1-\frac{M(g_c(m)-1)-2}{M(g_c(m)-1)-1}\gamma}-1\right) \\
&= (1-\gamma)\frac{(m^*+1)^{1-2\gamma}}{(m^{1-\gamma}-\gamma)^2}\left(1-\frac{\gamma-\frac{M(g_c(m)-1)-2}{M(g_c(m)-1)-1}\gamma}{1-\frac{M(g_c(m)-1)-2}{M(g_c(m)-1)-1}\gamma}\right) \\
&\quad -\frac{1}{1-\frac{M(g_c(m)-1)-2}{M(g_c(m)-1)-1}\gamma}\frac{(m^*+1)^{-\frac{M(g_c(m)-1)-2}{M(g_c(m)-1)-1}\gamma}}{\left(\frac{1}{1-\gamma}m^{1-\gamma}-\frac{\gamma}{1-\gamma}\right)^2} \\
&\geq (1-\gamma)\frac{(m^*+1)^{1-2\gamma}}{(m^{1-\gamma}-\gamma)^2}-o\left(\frac{(m^*+1)^{1-2\gamma}}{(m^{1-\gamma}-\gamma)^2}\right) \\
&= (1-\gamma)\frac{\left(\frac{M}{\gamma}g_c(m)+1\right)^{1-2\gamma}}{(m^{1-\gamma}-\gamma)^2}-o\left(\frac{\left(\frac{M}{\gamma}g_c(m)+1\right)^{1-2\gamma}}{(m^{1-\gamma}-\gamma)^2}\right) \qquad (117)
\end{aligned}
$$

In order to obtain the scaling behavior of (113) we consider the cases of $\gamma<\frac{1}{2}$, $\gamma>\frac{1}{2}$ and $\gamma=\frac{1}{2}$.

For $\gamma<\frac{1}{2}$, let $g_c(m)=\frac{c_1\gamma m^\alpha}{M}$, by using Lemma 1, (114) and (117), we have

$$
\begin{aligned}
&\sum_{i=1}^{m^*}\left(P_r(i)^2(1-(1-P_c^*(i))^{M(g_c(m)-1)})\right) \\
&\leq \frac{(1-\gamma)^2}{1-2\gamma}\frac{c_1^{1-2\gamma}m^{\frac{(1-\gamma)(1-2\gamma)}{2-\gamma}}}{(m^{1-\gamma}-1)^2}-(1-\gamma)\frac{(c_1m^{\frac{1-\gamma}{2-\gamma}})^{1-2\gamma}}{m^{2(1-\gamma)}}-\frac{2\gamma}{(m^{1-\gamma}-1)^2}+o\left(\frac{(c_1m^{\frac{1-\gamma}{2-\gamma}}+1)^{1-2\gamma}}{(m^{1-\gamma}-\gamma)^2}\right) \\
&= \frac{(1-\gamma)^2}{1-2\gamma}c_1^{1-2\gamma}m^{-\frac{3(1-\gamma)}{2-\gamma}}\left(1+\frac{1}{m^{1-\gamma}-1}\right)^2-(1-\gamma)c_1^{1-2\gamma}m^{-\frac{3(1-\gamma)}{2-\gamma}} \\
&\quad -\frac{2\gamma}{(m^{1-\gamma}-1)^2}+o\left(\frac{(c_1m^{\frac{1-\gamma}{2-\gamma}}+1)^{1-2\gamma}}{(m^{1-\gamma}-\gamma)^2}\right) \\
&= \frac{\gamma(1-\gamma)}{1-2\gamma}c_1^{1-2\gamma}m^{-\frac{3(1-\gamma)}{2-\gamma}}+o\left(m^{-\frac{3(1-\gamma)}{2-\gamma}}\right). \qquad (118)
\end{aligned}
$$

For $\gamma>\frac{1}{2}$, let $g_c(m)=\frac{c_1\gamma m^\alpha}{M}$, by using Lemma 1, (114) and (117), we have

$$
\begin{aligned}
&\sum_{i=1}^{m^*}\left(P_r(i)^2(1-(1-P_c^*(i))^{M(g_c(m)-1)})\right) \\
&\leq \frac{(1-\gamma)^2}{1-2\gamma}\frac{c_1^{1-2\gamma}m^{\frac{(1-\gamma)(1-2\gamma)}{2-\gamma}}-2\gamma}{(m^{1-\gamma}-1)^2}-(1-\gamma)\frac{(c_1m^{\frac{1-\gamma}{2-\gamma}})^{1-2\gamma}}{m^{2(1-\gamma)}}+o\left(\frac{(m^*+1)^{1-2\gamma}}{(m^{1-\gamma}-\gamma)^2}\right) \\
&\leq \frac{2\gamma(1-\gamma)^2}{2\gamma-1}m^{-2(1-\gamma)}-(1-\gamma)c_1^{1-2\gamma}m^{-\frac{3(1-\gamma)}{2-\gamma}}+o\left(\frac{(m^*+1)^{1-2\gamma}}{(m^{1-\gamma}-\gamma)^2}\right) \\
&= \frac{2\gamma(1-\gamma)^2}{2\gamma-1}m^{-2(1-\gamma)}-O\left(m^{-\frac{3(1-\gamma)}{2-\gamma}}\right). \qquad (119)
\end{aligned}
$$

This settles the scaling behavior of the term $\sum_{i=1}^{m^*}\left(P_r(i)^2(1-(1-P_c^*(i))^{M(g_c(m)-1)})\right)$ for $\gamma \neq \frac{1}{2}$.

For the case $\gamma = \frac{1}{2}$, let $g_c(m) = \frac{c_1 \gamma m^\alpha}{M}$, we use (115) and (117) to obtain

$$\begin{aligned}
&\sum_{i=1}^{m^*}\left(P_r(i)^2(1-(1-P_c^*(i))^{M(g_c(m)-1)})\right) \\
&= \frac{1}{12}\frac{\log(m)+\log(c_1)+1}{\left(m^{\frac{1}{2}}-1\right)^2} - \frac{1}{2m} + o\left(\frac{1}{m}\right) \\
&= \frac{1}{12}\frac{\log m}{m} + O\left(\frac{1}{m}\right). \qquad (120)
\end{aligned}$$

From (67), and by using (119) and (119), we obtain the desired result.

APPENDIX I
PROOF OF LEMMA 6

By using (72) and (113), we have two cases of γ to consider, namely, $\gamma \neq \frac{1}{2}$ and $\gamma = \frac{1}{2}$. When $\gamma \neq \frac{1}{2}$, by using (65), (66), (114) and (117), we have

$$\begin{aligned}
p_{uu'}^c &\leq \left(\gamma^\gamma\left(\frac{Mg_c(m)}{m}\right)^{1-\gamma} + o\left(\left(\frac{Mg_c(m)}{m}\right)^{1-\gamma}\right)\right)^2 + \frac{(1-\gamma)^2}{1-2\gamma}\frac{\left(\frac{M}{\gamma}g_c(m)\right)^{1-2\gamma}-2\gamma}{(m^{1-\gamma}-1)^2} \\
&\quad -(1-\gamma)\frac{\left(\frac{M}{\gamma}g_c(m)+1\right)^{1-2\gamma}}{(m^{1-\gamma}-\gamma)^2} + o\left(\frac{\left(\frac{M}{\gamma}g_c(m)+1\right)^{1-2\gamma}}{(m^{1-\gamma}-\gamma)^2}\right) \\
&= \gamma^{2\gamma}\left(\frac{Mg_c(m)}{m}\right)^{2(1-\gamma)} + \frac{(1-\gamma)^2}{(1-2\gamma)}\frac{\left(\frac{Mg_c(m)}{\gamma}\right)^{1-2\gamma}-2\gamma}{(m^{1-\gamma}-1)^2} \\
&\quad -\frac{\frac{1-\gamma}{\gamma^{1-2\gamma}}\left(Mg_c(m)+\gamma\right)^{1-2\gamma}}{(m^{1-\gamma}-1)^2} + o\left(\left(\frac{Mg_c(m)}{m}\right)^{2(1-\gamma)}\right) \\
&\leq \gamma^{2\gamma}\left(\frac{Mg_c(m)}{m}\right)^{2(1-\gamma)} + o\left(\left(\frac{Mg_c(m)}{m}\right)^{2(1-\gamma)}\right). \qquad (121)
\end{aligned}$$

When $\gamma = \frac{1}{2}$, by using (65), (66), (115) and (120), we have

$$\begin{aligned}
p_{uu'}^c &\leq \left(\left(\frac{1}{2}\right)^{\frac{1}{2}}\left(\frac{Mg_c(m)}{m}\right)^{1-\frac{1}{2}} + o\left(\frac{Mg_c(m)}{m}\right)^{\frac{1}{2}}\right)^2 + \frac{\log\left(\frac{M}{\frac{1}{2}}g_c(m)\right)+1}{\left(2(m+1)^{\frac{1}{2}}-2\right)^2} \\
&\quad -\frac{1}{2}\frac{\left(\frac{M}{\frac{1}{2}}g_c(m)\right)^{1-2\frac{1}{2}}}{m^{2(1-\frac{1}{2})}} + o\left(\frac{\left(\frac{M}{\frac{1}{2}}g_c(m)\right)^{1-2\frac{1}{2}}}{m^{2(1-\frac{1}{2})}}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}\left(\frac{Mg_c(m)}{m}\right) + \frac{\log(Mg_c(m)) + \log 2 + 1}{\left(2(m+1)^{\frac{1}{2}} - 2\right)^2} - \frac{1}{2m} \\
&\quad + o\left(\frac{Mg_c(m)}{m}\right) \\
&\leq \frac{1}{2}\left(\frac{Mg_c(m)}{m}\right) + o\left(\frac{Mg_c(m)}{m}\right). \qquad (122)
\end{aligned}
$$

Thus, (121) and (122) give the desired result.

APPENDIX J
CONTINUITY AND PERTURBATIONS

In this section, under the condition that

$$T^*_{\rm sum}(p) \leq f_{\rm ub}(\rho^*)\frac{n}{m^\alpha}, \qquad (123)$$

and

$$p^{\rm lb}(g^*) = 1 - (M\rho^*)^{1-\gamma}m^{-\alpha} + o(m^{-\alpha}), \qquad (124)$$

we want to show that

$$T^*_{\rm sum}(p) \leq f_{\rm ub}(\rho^*)\frac{n}{m^\alpha} + no(m^{-\alpha}), \qquad (125)$$

where $p = p^{\rm lb}(g^*) = 1 - (M\rho^*)^{1-\gamma}m^{-\alpha}$.

From calculus, We know that

$$
\begin{aligned}
&T^{\rm ub}_{\rm sum}(1 - (M\rho^*)^{1-\gamma}m^{-\alpha}) \\
&= T^{\rm ub}_{\rm sum}(1 - (M\rho^*)^{1-\gamma}m^{-\alpha} + o(m^{-\alpha})) \qquad (126) \\
&+ \left.\frac{\frac{dT^{\rm ub}_{\rm sum}}{dg}}{\frac{dp^{\rm lb}}{dg}}\right|_{g=\rho^* m^{-\alpha}} \times o(m^{-\alpha}) + o(o(m^{-\alpha})) \\
&= f_{\rm ub}(\rho^*)\frac{n}{m^\alpha} + \left.\frac{\frac{dT^{\rm ub}_{\rm sum}}{dg}}{\frac{dp^{\rm lb}}{dg}}\right|_{g=\rho^* m^{-\alpha}} \times o(m^{-\alpha}) + o(o(m^{-\alpha})). \qquad (127)
\end{aligned}
$$

Thus, the goal is to compute $\left.\frac{\frac{dT^{\rm ub}_{\rm sum}}{dg}}{\frac{dp^{\rm lb}}{dg}}\right|_{g=\rho^* m^{-\alpha}}$, which requires to compute $\frac{dT^{\rm ub}_{\rm sum}}{dg}$ and $\frac{dp^{\rm lb}}{dg}$. To obtain $\frac{dT^{\rm ub}_{\rm sum}}{dg}$, we first need to compute the derivative in the form of $F(x) = f(x)^{g(x)}$, which is given by

$$
\begin{aligned}
\frac{dF(x)}{dx} &= F(x)\left(g'(x)\log f(x) + \frac{g(x)}{f(x)}f'(x)\right) \\
&= f(x)^{g(x)}\left(g'(x)\log f(x) + \frac{g(x)}{f(x)}f'(x)\right). \qquad (128)
\end{aligned}
$$

Then, by denoting $g_R(m)$ as g, we obtain

$$\begin{aligned}
\frac{dT_{\mathrm{sum}}^{\mathrm{ub}}}{dg} &= \frac{\partial}{\partial g}\frac{16}{\Delta^2}\cdot\left(\left(1-(p^{\mathrm{lb}}(g))^{\left(1+\frac{3\Delta}{2}\right)^2 g}\right)\frac{n}{g}\right) \\
&= \frac{16}{\Delta^2}\left(-\frac{n}{g}\frac{\partial}{\partial g}\left((p^{\mathrm{lb}}(g))^{\left(1+\frac{3\Delta}{2}\right)^2 g}\right)+\left(1-(p^{\mathrm{lb}}(g))^{\left(1+\frac{3\Delta}{2}\right)^2 g}\right)\frac{\partial}{\partial g}\left(\frac{n}{g}\right)\right) \\
&= \frac{16}{\Delta^2}\left(-\frac{n}{g}(p^{\mathrm{lb}}(g))^{\left(1+\frac{3\Delta}{2}\right)^2 g}\left(1+\frac{3\Delta}{2}\right)^2\left(\log(p^{\mathrm{lb}}(g))+\frac{g}{p^{\mathrm{lb}}(g)}\left(\frac{\partial p^{\mathrm{lb}}(g)}{\partial g}\right)\right)\right. \\
&\quad \left.+\left(1-(p^{\mathrm{lb}}(g))^{\left(1+\frac{3\Delta}{2}\right)^2 g}\right)\frac{\frac{dn}{dg}g-n}{g^2}\right) \\
&= \frac{16}{\Delta^2}\left(-\frac{n}{g}(p^{\mathrm{lb}}(g))^{\left(1+\frac{3\Delta}{2}\right)^2 g}\left(1+\frac{3\Delta}{2}\right)^2\left(\log(p^{\mathrm{lb}}(g))+\frac{g}{p^{\mathrm{lb}}(g)}\left(\frac{\partial p^{\mathrm{lb}}(g)}{\partial g}\right)\right)\right. \\
&\quad \left.+\left(1-(p^{\mathrm{lb}}(g))^{\left(1+\frac{3\Delta}{2}\right)^2 g}\right)\frac{-n}{g^2}\right). \qquad (129)
\end{aligned}$$

Then, we get

$$\begin{aligned}
\frac{\partial p^{\mathrm{lb}}(g)}{\partial g} &= \frac{\partial}{\partial g}\left(1-\frac{\frac{1}{1-\gamma}(Mg)^{1-\gamma}-\frac{1}{1-\gamma}+1}{\frac{1}{1-\gamma}m^{1-\gamma}-\frac{1}{1-\gamma}}\right) \\
&= -\frac{(Mg)^{-\gamma}M\left(\frac{1}{1-\gamma}m^{1-\gamma}-\frac{1}{1-\gamma}\right)-\left(\frac{1}{1-\gamma}(Mg)^{1-\gamma}-\frac{1}{1-\gamma}+1\right)m^{-\gamma}\cdot\frac{dm}{dg}}{\left(\frac{1}{1-\gamma}m^{1-\gamma}-\frac{1}{1-\gamma}\right)^2} \\
&= -\frac{(Mg)^{-\gamma}M}{\frac{1}{1-\gamma}m^{1-\gamma}-\frac{1}{1-\gamma}}. \qquad (130)
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
\frac{\frac{dT_{\mathrm{sum}}^{\mathrm{ub}}}{dg}}{\frac{dp^{\mathrm{lb}}}{dg}} &= \frac{16}{\Delta^2}\frac{\left(-\frac{n}{g}(p^{\mathrm{lb}}(g))^{\left(1+\frac{3\Delta}{2}\right)^2 g}\left(1+\frac{3\Delta}{2}\right)^2\left(\log(p^{\mathrm{lb}}(g))+\frac{g}{p^{\mathrm{lb}}(g)}\left(\frac{\partial p^{\mathrm{lb}}(g)}{\partial g}\right)\right)\right)}{-\frac{(Mg)^{-\gamma}M}{\frac{1}{1-\gamma}m^{1-\gamma}-\frac{1}{1-\gamma}}} \\
&\quad+\frac{16}{\Delta^2}\frac{\left(p^{\mathrm{lb}}(g)\right)^{\left(1+\frac{3\Delta}{2}\right)^2 g}\frac{-n}{g^2}}{-\frac{(Mg)^{-\gamma}M}{\frac{1}{1-\gamma}m^{1-\gamma}-\frac{1}{1-\gamma}}} \\
&= \frac{16}{\Delta^2}\frac{-\frac{n}{g}(p^{\mathrm{lb}}(g))^{\left(1+\frac{3\Delta}{2}\right)^2 g}\left(1+\frac{3\Delta}{2}\right)^2\log(p^{\mathrm{lb}}(g))}{-\frac{(Mg)^{-\gamma}M}{\frac{1}{1-\gamma}m^{1-\gamma}-\frac{1}{1-\gamma}}} \\
&\quad+\frac{16}{\Delta^2}\frac{-\frac{n}{g}(p^{\mathrm{lb}}(g))^{\left(1+\frac{3\Delta}{2}\right)^2 g}\left(1+\frac{3\Delta}{2}\right)^2\frac{g}{p^{\mathrm{lb}}(g)}\left(\frac{\partial p^{\mathrm{lb}}(g)}{\partial g}\right)}{-\frac{(Mg)^{-\gamma}M}{\frac{1}{1-\gamma}m^{1-\gamma}-\frac{1}{1-\gamma}}} \\
&\quad+\frac{16}{\Delta^2}\frac{\left(1-(p^{\mathrm{lb}}(g))^{\left(1+\frac{3\Delta}{2}\right)^2 g}\right)\frac{-n}{g^2}}{-\frac{(Mg)^{-\gamma}M}{\frac{1}{1-\gamma}m^{1-\gamma}-\frac{1}{1-\gamma}}}. \qquad (131)
\end{aligned}$$

By letting $m \to \infty$, we obtain

$$
\begin{aligned}
&\left.\frac{\frac{dT_{\mathrm{sum}}^{\mathrm{ub}}}{dg}}{\frac{dp^{\mathrm{lb}}}{dg}}\right|_{g=\rho^* m^{-\alpha}} \\
&= \frac{16}{\Delta^2}\left(1+\frac{3\Delta}{2}\right)^2 \frac{1}{\rho^*}\frac{M^{\gamma-1}}{1-\gamma}\rho^{*\gamma}\left(e^{-\zeta(\rho^*)}+o(1)\right)\frac{n}{m^\alpha}\log(p^{\mathrm{lb}}(g))m^{\alpha\gamma}m^{1-\gamma} \\
&\quad - \frac{16}{\Delta^2}\left(1+\frac{3\Delta}{2}\right)^2 \frac{1}{\rho^*}\left(e^{-\zeta(\rho^*)}+o(1)\right)\frac{n}{m^\alpha}\frac{\rho^* m^\alpha}{p^{\mathrm{lb}}(g)} \\
&\quad + \frac{16}{\Delta^2}\left(1-e^{-\zeta(\rho^*)}+o(1)\right)\frac{1}{\rho^{*2}}\frac{n}{m^{2\alpha}}\frac{M^{\gamma-1}}{1-\gamma}\rho^{*\gamma}m^{\alpha\gamma}m^{1-\gamma} \\
&= \frac{16}{\Delta^2}\left(1+\frac{3\Delta}{2}\right)^2 \frac{M^{\gamma-1}}{1-\gamma}\rho^{*\gamma-1}\left(e^{-\zeta(\rho^*)}+o(1)\right)\frac{n}{m^\alpha}(-(1-p^{\mathrm{lb}}(g))+O\left((1-p^{\mathrm{lb}}(g))^2\right)m^{\alpha\gamma}m^{1-\gamma} \\
&\quad - \frac{16}{\Delta^2}\left(1+\frac{3\Delta}{2}\right)^2 \left(e^{-\zeta(\rho^*)}+o(1)\right)\frac{n}{p^{\mathrm{lb}}(g)} \\
&\quad + \frac{16}{\Delta^2}\left(1-e^{-\zeta(\rho^*)}+o(1)\right)\frac{M^{\gamma-1}}{1-\gamma}\rho^{*\gamma-2}\frac{n}{m^{2\alpha}}m^{\alpha\gamma}m^{1-\gamma} \\
&= -\frac{16}{\Delta^2}\left(1+\frac{3\Delta}{2}\right)^2 \frac{M^{\gamma-1}}{1-\gamma}\rho^{*\gamma-1}\left(e^{-\zeta(\rho^*)}+o(1)\right) \\
&\quad \cdot\left(M^{1-\gamma}\rho^{*1-\gamma}\frac{n}{m^\alpha}m^{\alpha(1-\gamma)-(1-\gamma)}m^{\alpha\gamma}m^{1-\gamma}+O\left(\frac{n}{m^\alpha}m^{\alpha\gamma}m^{1-\gamma}\frac{1}{m^{2\alpha}}\right)\right) \\
&\quad - \frac{16}{\Delta^2}\left(1+\frac{3\Delta}{2}\right)^2 \left(e^{-\zeta(\rho^*)}+o(1)\right)\frac{n}{p^{\mathrm{lb}}(g)} \\
&\quad + \frac{16}{\Delta^2}\left(1-e^{-\zeta(\rho^*)}+o(1)\right)\frac{M^{\gamma-1}}{1-\gamma}\rho^{*\gamma-2}\frac{n}{m^{2\alpha}}m^{\alpha\gamma}m^{1-\gamma} \\
&= -\frac{16}{\Delta^2}\left(1+\frac{3\Delta}{2}\right)^2 \frac{1}{1-\gamma}\left(e^{-\zeta(\rho^*)}+o(1)\right)\left(1+O\left(\frac{1}{m^\alpha}\right)\right)n \\
&\quad - \frac{16}{\Delta^2}\left(1+\frac{3\Delta}{2}\right)^2 \left(e^{-\zeta(\rho^*)}+o(1)\right)n \\
&\quad + \frac{16}{\Delta^2}\left(1-e^{-\zeta(\rho^*)}+o(1)\right)\frac{M^{\gamma-1}}{1-\gamma}\rho^{*\gamma-2}n \\
&= \frac{16}{\Delta^2}\left(\frac{M^{\gamma-1}}{1-\gamma}\rho^{*\gamma-2}\left(1-e^{-\zeta(\rho^*)}+o(1)\right)\right. \\
&\quad \left. - \left(1+\frac{3\Delta}{2}\right)^2\left(e^{-\zeta(\rho^*)}+o(1)\right) - \left(1+\frac{3\Delta}{2}\right)^2\frac{1}{1-\gamma}\left(e^{-\zeta(\rho^*)}+o(1)\right)+O\left(\frac{1}{m^\alpha}\right)\right)n \\
&= \frac{16}{\Delta^2}\left(\frac{M^{\gamma-1}}{1-\gamma}\rho^{*\gamma-2}\left(1-e^{-\zeta(\rho^*)}\right)\right. \\
&\quad \left. - \left(1+\frac{3\Delta}{2}\right)^2 e^{-\zeta(\rho^*)} - \left(1+\frac{3\Delta}{2}\right)^2\frac{1}{1-\gamma}e^{-\zeta(\rho^*)}+o\left(1\right)\right)n \\
&= O(n). \qquad (132)
\end{aligned}
$$

Thus, we obtain

$$
\begin{aligned}
&T_{\mathrm{sum}}^{\mathrm{ub}}(1-(M\rho^*)^{1-\gamma}m^{-\alpha}) \\
&= T_{\mathrm{sum}}^{\mathrm{ub}}(1-(M\rho^*)^{1-\gamma}m^{-\alpha}+o(m^{-\alpha}))+\left.\frac{\frac{dT_{\mathrm{sum}}^{\mathrm{ub}}}{dg}}{\frac{dp^{\mathrm{lb}}}{dg}}\right|_{g=\rho^* m^{-\alpha}} \times o(m^{-\alpha})+o(o(m^{-\alpha})) \\
&= f_{\mathrm{ub}}(\rho^*)\frac{n}{m^\alpha}+\left.\frac{\frac{dT_{\mathrm{sum}}^{\mathrm{ub}}}{dg}}{\frac{dp^{\mathrm{lb}}}{dg}}\right|_{g=\rho^* m^{\alpha}} \times o(m^{-\alpha})+o(o(m^{-\alpha})) \\
&= f_{\mathrm{ub}}(\rho^*)\frac{n}{m^\alpha}+O(n)\cdot o(m^{-\alpha})+o(o(m^{-\alpha})) \\
&\leq f_{\mathrm{ub}}(\rho^*)\frac{n}{m^\alpha}+no(m^{-\alpha}). \qquad (133)
\end{aligned}
$$

REFERENCES

[1] Cisco, "The Zettabyte Era-Trends and Analysis," 2013.

[2] F-L. Luo, *Mobile Multimedia Broadcasting Standards: Technology and Practice*, Springer Verlag, 2008.

[3] U. Reimers, "Digital video broadcasting," *Communications Magazine, IEEE*, vol. 36, no. 6, pp. 104–110, 1998.

[4] U. Ladebusch and C.A. Liss, "Terrestrial dvb (dvb-t): A broadcast technology for stationary portable and mobile use," *Proceedings of the IEEE*, vol. 94, no. 1, pp. 183–193, 2006.

[5] O.Y. Bursalioglu, M. Fresia, G. Caire, and H.V. Poor, "Lossy multicasting over binary symmetric broadcast channels," *Signal Processing, IEEE Transactions on*, vol. 59, no. 8, pp. 3915–3929, 2011.

[6] Y. Li, E. Soljanin, and P. Spasojević, "Three schemes for wireless coded broadcast to heterogeneous users," *Physical Communication*, 2012.

[7] W.H.R. Equitz and T.M. Cover, "Successive refinement of information," *Information Theory, IEEE Transactions on*, vol. 37, no. 2, pp. 269–275, 1991.

[8] O.Y. Bursalioglu, M. Fresia, G. Caire, and H.V. Poor, "Lossy joint source-channel coding using raptor codes," *International Journal of Digital Multimedia Broadcasting*, vol. 2008, 2008.

[9] M.R. Chari, F. Ling, A. Mantravadi, R. Krishnamoorthi, R. Vijayan, G.K. Walker, and R. C, "Flo physical layer: An overview," *Broadcasting, IEEE Transactions on*, vol. 53, no. 1, pp. 145–160, 2007.

[10] V.K. Goyal, "Multiple description coding: Compression meets the network," *Signal Processing Magazine, IEEE*, vol. 18, no. 5, pp. 74–93, 2001.

[11] Y. Wang, A.R. Reibman, and S. Lin, "Multiple description coding for video delivery," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 57–70, 2005.

[12] R. Ahlswede, "On multiple descriptions and team guessing," *Information Theory, IEEE Transactions on*, vol. 32, no. 4, pp. 543–549, 1986.

[13] E. Nygren, R. K. Sitaraman, and J. Sun, "The akamai network: a platform for high-performance internet applications," *ACM SIGOPS Operating Systems Review*, vol. 44, no. 3, pp. 2–19, 2010.

[14] N. Golrezaei, A.F. Molisch, and A.G. Dimakis, "Base station assisted device-to-device communications for high-throughput wireless video networks," *IEEE Communications Magazine, in press.*, 2012.

[15] N. Golrezaei, K. Shanmugam, A. G Dimakis, A. F Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," *CoRR*, vol. abs/1109.4179, 2011.

[16] M.A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *arXiv preprint arXiv:1209.5807*, 2012.

[17] M.A. Maddah-Ali and U. Niesen, "Decentralized caching attains order-optimal memory-rate tradeoff," *arXiv preprint arXiv:1301.5848*, 2013.

[18] X. Wu, S. Tavildar, S. Shakkottai, T. Richardson, J. Li, R. Laroia, and A. Jovicic, "Flashlinq: A synchronous distributed scheduler for peer-to-peer ad hoc networks," in *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*. IEEE, 2010, pp. 514–521.

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

[20] S.R. Kulkarni and P. Viswanath, "A deterministic approach to throughput scaling in wireless networks," *Information Theory, IEEE Transactions on*, vol. 50, no. 6, pp. 1041–1049, 2004.

[21] S. Shakkottai, X. Liu, and R. Srikant, "The multicast capacity of large multihop wireless networks," *IEEE/ACM Transactions on Networking (TON)*, vol. 18, no. 6, pp. 1691–1700, 2010.

[22] U. Niesen, P. Gupta, and D. Shah, "The balanced unicast and multicast capacity regions of large wireless networks," *Information Theory, IEEE Transactions on*, vol. 56, no. 5, pp. 2249–2271, 2010.

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

[24] S. Gitzenis, GS Paschos, and L. Tassiulas, "Asymptotic laws for joint content replication and delivery in wireless networks," *Arxiv preprint arXiv:1201.3095*, 2012.

[25] N. Golrezaei, A.G. Dimakis, and A.F. Molisch, "Wireless device-to-device communications with distributed caching," in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*. IEEE, 2012, pp. 2781–2785.

[26] Y. Sánchez de la Fuente, T. Schierl, C. Hellge, T. Wiegand, D. Hong, D. De Vleeschauwer, W. Van Leekwijck, and Y. Le Louédec, "Improved caching for http-based video on demand using scalable video coding," in *Consumer Communications and Networking Conference (CCNC), 2011 IEEE*. IEEE, 2011, pp. 595–599.

[27] Y. Sánchez de la Fuente, T. Schierl, C. Hellge, T. Wiegand, D. Hong, D. De Vleeschauwer, W. Van Leekwijck, and Y. Le Louédec, "idash: improved dynamic adaptive streaming over http using scalable video coding," in *Proceedings of the second annual ACM conference on Multimedia systems*. ACM, 2011, pp. 257–264.

[28] R. Pantos, "Http live streaming," 2012.

[29] T. Stockhammer, "Dynamic adaptive streaming over http–: standards and design principles," in *Proceedings of the second annual ACM conference on Multimedia systems*. ACM, 2011, pp. 133–144.

[30] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: Evidence and implications," in *INFOCOM'99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*. IEEE, 1999, vol. 1, pp. 126–134.

[31] M. Zink, K. Suh, Y. Gu, and J. Kurose, "Watch global, cache local: YouTube network traffic at a campus network-measurements and implications," *Proceeding of the 15th SPIE/ACM Multimedia Computing and Networking*, 2008.

[32] M. Zink, K. Suh, Y. Gu, and J. Kurose, "Characteristics of youtube network traffic at a campus network-measurements, models, and implications," *Computer Networks*, vol. 53, no. 4, pp. 501–514, 2009.

[33] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: Evidence and implications," in *INFOCOM 99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies*. IEEE, 1999, vol. 1, pp. 126–134.

[34] A.F. Molisch, *Wireless communications*, John Wiley & Sons, 2011.

[35] L. Juhn and L. Tseng, "Harmonic broadcasting for video-on-demand service," *Broadcasting, IEEE Transactions on*, vol. 43, no. 3, pp. 268–271, 1997.

[36] J-F Pâris, S.W. Carter, and D.E. Long, "Efficient broadcasting protocols for video on demand," in *Modeling, Analysis and Simulation of Computer and Telecommunication Systems, 1998. Proceedings. Sixth International Symposium on*. IEEE, 1998, pp. 127–132.

[37] J-F Pâris, S.W. Carter, and D.E. Long, "A low bandwidth broadcasting protocol for video on demand," in *Computer Communications and Networks, 1998. Proceedings. 7th International Conference on*. IEEE, 1998, pp. 690–697.

[38] L. Engebretsen and M. Sudan, "Harmonic broadcasting is bandwidth-optimal assuming constant bit rate," *Networks*, vol. 47, no. 3, pp. 172–177, 2006.

[39] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *arXiv preprint arXiv:1305.5216*, 2013.

[40] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless d2d networks," *arXiv preprint:1405.5336*, 2014.

[41] S.P. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge Univ Pr, 2004.

[42] A. Özgür, "Operating regimes of large wireless networks," *Foundations and Trends® in Networking*, vol. 5, no. 1, pp. 1–107, 2010.

[43] W. Rudin, *Principles of mathematical analysis*, vol. 3, McGraw-Hill New York, 1976.