Popularity-Aware Caching Increases the Capacity of Wireless Networks

Li Qiu and Guohong Cao
Department of Computer Science and Engineering
The Pennsylvania State University
E-mail: {lyq5023, gcao}@cse.psu.edu

***Abstract*—In wireless ad hoc networks, due to the interference between concurrent transmissions, the per node capacity generally decreases with the increasing number of nodes in the network. Caching can help improve the network capacity, as it shortens the content transmission distance and reduces the communication interference. However, current researches on the capacity of wireless ad hoc networks with caching generally assume that content popularity follows uniform distribution. They ignore the fact that contents in reality have skewed popularity, which may lead to totally different capacity results. In this paper, we evaluate how the distribution of the content popularity affects the network capacity, and derive different capacity scaling laws based on the skewness of the content popularity. Our results suggest that for wireless networks with caching, when contents have skewed popularity, increasing the number of nodes monotonically increases the per node capacity.**

I. Introduction

Wireless ad hoc networks enable communications among mobile nodes without any infrastructure support, as nodes themselves relay and forward packets for each other. Due to the interference between concurrent transmissions, the per node capacity generally decreases with the increasing number of nodes in the network. This was first found by Gupta and Kumar [1]. They showed that for a wireless ad hoc network with n nodes, each node can at most transmit at a rate of $\Theta(\frac{1}{\sqrt{n}})$ to its destination, even with optimal scheduling of transmissions from various nodes.

Later, Grossglauser and Tse [2] examined the capacity of a mobile ad hoc network. They proved that with node mobility, the per node capacity can be kept constant even when the number of nodes in the network grows. In their proposed scheme, two nodes wait until they move close enough to transmit data. In this way, the network capacity can be significantly improved, since the transmissions are limited to nearby neighbors and the consequent interference will affect much less number of nodes. However, the delay of their proposed scheme is extremely long, because nodes need to wait until they are close enough to transmit.

Caching can be used to improve network capacity. With caching, contents are stored close to the users, which shortens the transmission distance and improves the network capacity. Although theoretical study of caching has attracted considerable attention, most existing research focuses on caching in the Internet [3]–[6], and the results can not be directly applied to wireless ad hoc networks. It was only recently that researchers have become interested in studying the fundamental performance limits of caching in wireless ad hoc networks. Liu *et al.* [7] have derived capacity upper bounds for two specific content access schemes, and examined how the network capacity is affected by the cache size and the number of nodes. In [8], the authors proved that for wireless networks with caching, the per node capacity will remain constant even when the number of nodes grows. However, these works simply assume that content popularity follows a uniform distribution. They ignore the fact that in reality some contents are accessed much more frequently than others, which requires a more complicated caching strategy to maximize the network capacity, and may lead to totally different scaling laws of network capacity.

In this paper, we quantify the effect of popularity-aware caching on the capacity of wireless ad hoc networks. To maximize the network capacity given the content popularity, we first study the optimal caching strategy; i.e., how frequently contents with various popularity should be cached so that the network capacity is maximized. Based on the optimal caching strategy, we then evaluate the effect of content popularity on network capacity, and derive different capacity scaling laws for networks with different content popularity skewness. For all the different capacity scaling laws, we analytically investigate how the network capacity is affected by various parameters, including the number of nodes (n), the cache size (s), and the number of unique content (m). The main contributions of the paper are summarized as follows:

- We propose the optimal popularity-aware caching strategy that maximizes the per node capacity for wireless ad hoc networks.
- We derive different capacity scaling laws based on the skewness of the content popularity. Basically, as the distribution of the content popularity changes from uniform distribution to more skewed distributions, the network capacity increases from $\Theta(\sqrt{\frac{s}{m}})$ to roughly $\Theta(\sqrt{s})$.
- We find that for wireless ad hoc networks with caching, when contents have skewed popularity, increasing the number of nodes monotonically increases the per node capacity.

The rest of the paper is organized as follows. Section II reviews existing work. Section III introduces the network model. Section IV gives the problem formulation. In Section

This work was supported in part by the National Science Foundation (NSF) under grant CNS-1320278, CNS-1526425, and by Network Science CTA under grant W911NF-09-2-0053.

V, we derive the capacity scaling laws. We analyze the effects of various parameters in Section VI. We present the numerical results in Section VII. We conclude the paper in Section VIII.

II. Related Work

Caching in wireless networks is a traditional topic that has been studied from various aspects. In [9], Yin and Cao have designed and evaluated cooperative caching schemes to support data access in wireless ad hoc networks. In [10], the authors have implemented and examined a caching scheme for wireless P2P networks. Fiore *et al.* [11] have proposed an algorithm to help users decide whether a content should be cached, so as to reduce the data redundancy among neighbors. However, none of them has analytically investigated the fundamental performance limits of caching in wireless networks.

Recently, researchers have become interested in studying the fundamental property of caching in wireless networks [7], [8], [12], [13]. Ji *et al.* [12] have presented capacity upper and lower bounds for a wireless D2D network, where the content transmission is restricted to be single-hop. Liu *et al.* [7] have derived capacity upper bounds for a wireless ad hoc network under two specific content access schemes. In [8], the authors have derived the asymptotic bounds of network capacity, which suggests that for contents with uniform popularity, the per node capacity will remain constant even when the number of nodes grows. Yet, in all these works, their results are under the assumption that content access pattern is either arbitrary or follows a simple uniform distribution, and the fundamental problem of how the distribution of content popularity affects the scaling laws of the network capacity has not been explored.

There exists some research on optimal caching when the content popularity is known. In [14], Cohen and Shenker studied the problem of how the contents with various popularity should be cached in P2P networks, so that the number of searches to retrieve the content is minimized. Jin and Wang [15] proposed techniques to determine how contents should be cached at various nodes in wireless networks. However, in both papers, the authors only proposed caching schemes, the effect of caching on network capacity was not investigated.

In [16], Gitzenis *et al.* have examined the asymptotic laws for joint replication and delivery in wireless networks. Unlike this work and prior works [1], [2], they have derived the minimum throughput on each link so that every node can satisfy one request per second. However, whether such throughput can be supported by the network is unknown, since the transmissions on various links may interfere with each other. Hence, their throughput results are not as practical as our per node capacity results. Furthermore, our results are more interesting which suggests that the per node capacity will not diminish even when the number of nodes grows.

III. Preliminaries

A. Network Model

We consider a wireless network consisting of n nodes that are independently and uniformly distributed on the surface of a unit sphere. As in [1], we analyze the network capacity when nodes are located on the surface of sphere S^2 rather than on a disk so as to eliminate the edge effects; i.e., nodes near the edge have much fewer neighbors than nodes near the center.

Assume each node can cache s bits of contents, and all nodes employ the same amount of power $\mathcal{P}$ for transmission. All nodes transmit over a common wireless channel which can support W bits per second. According to the physical model in [1], when a node i sends data to node j, the transmission rate can reach W bits per second, if the signal-to-interference-plus-noise ratio (SINR) at node j is greater than or equal to β, where β is the minimum SINR for successful reception.

B. Content Access and Zipf Distribution

Let m denote the number of unique content throughout the network. To simplify the analysis, we assume each content has one bit. These m contents are cached throughout the network by various nodes. As each node has a cache size of s, for any node i, it caches s out of m contents locally.

We assume the three parameters n, s and m are independent. In wireless networks with caching, generally each node can only cache a small portion of the m unique contents, in this way, $m \gg s$. On the other hand, to guarantee that at least one copy of each content exists in the network, the total cache size of all nodes should be greater than or equal to m, i.e., $ns \geq m$. Thus, we assume $\frac{m}{n} \leq s \ll m$.

At each node, there is always a content request, and a new request will arrive after the previous request has been served. We assume the content requests are independent. The popularity of the contents follows the Zipf distribution, i.e., the probability that the i-th most popular content being requested is proportional to $\frac{1}{i^\gamma}$. This content access pattern has been used by previous studies, and existing works [17], [18] have shown that the content access pattern in the Internet follows Zipf distribution. We consider the case when the content popularity changes slowly or even remains unchanged. Typical examples include music files, photos from nearby events, and uploaded short videos, which will remain popular for a long period of time. There are m unique contents, and the probability that the i-th most popular content will be requested, denoted as ρ_i, will be $\rho_i = 1/(i^\gamma H_{m,\gamma})$, where $H_{m,\gamma}$ is the generalized harmonic number, given by

$$H_{m,\gamma} = \sum_{j=1}^{m} \frac{1}{j^\gamma}. \tag{1}$$

C. Capacity

As in [1], [2], network capacity describes node's capability to transmit or retrieve contents. In a wireless network with caching, we define the *network capacity* as the number of bits each node utilizes per second to satisfy content requests, which includes the amount of contents it receives from others, and the contents in the local cache that have been used to serve requests.

We study the scaling behavior of the per node capacity C based on four parameters: the number of nodes n, the number of unique content m, the cache size s, and the Zipf parameter γ. The network capacity is studied under the assumption, $n \to \infty$, $m \to \infty$ and $\frac{m}{n} \leq s \ll m$.

IV. Problem Formulation

When content popularity is known, the network capacity depends on the caching strategy, i.e., for content i with popularity ρ_i, how many nodes in the network should cache that content? Let $\mathbf{P} = \{p_i\}_{1\le i\le m}$ denote the density of the contents, where p_i is the density of content i (i.e., the fraction of content i among all ns contents). Apparently, the sum of the densities of all contents must be less than or equal to 1:

$$\sum_{i=1}^{m} p_i \le 1. \tag{2}$$

It is meaningless for any node to cache more than one copy of the same content locally, thus p_i must satisfy

$$p_i \le \frac{1}{s}, \text{ for } i = 1, ..., m. \tag{3}$$

To guarantee that all contents can be found in the network, it is required that each content has at least one copy, that is

$$p_i \ge \frac{1}{ns}, \text{ for } i = 1, ..., m. \tag{4}$$

Given the content density P, the distributed algorithms that actually enable nodes to cache contents according to P have been considered in [15], [19], and it is not the focus of this paper. To connect the density P to the network capacity, we review one important result from [1]. In the paper, the authors proved that for n nodes uniformly distributed on the surface of a sphere, the per node capacity is upper bounded by $\frac{W}{L\sqrt{n}}$, where L is the average transmission distance between the source nodes and destination nodes. This result implies that the network capacity will increase as nodes retrieve contents from closer neighbors. To obtain the highest capacity, we should find the densities P that can minimize the average transmission distance L, where P is subject to constraints (2), (3) and (4).

Consider a content i with popularity ρ_i and density p_i. Since n nodes with cache size s can in total cache ns contents, i is cached at nsp_i nodes in the network (note that each node caches at most one copy of i locally). For the remaining $n - nsp_i$ nodes that have not cached i, they have to retrieve i from their neighbors. Thus, each node that caches i will on average be responsible for requests of i from $\frac{1}{p_i s} - 1$ other nodes. As all n nodes are uniformly and independently distributed, for any node j that caches i, based on [8], its average distance to itself and its closest $\frac{1}{p_i s} - 1$ neighbors is approximately:

$$L_i \approx \frac{2}{3}\arccos\left(1 - \frac{2}{n}\left(\frac{1}{p_i s} - 1\right)\right) \ge \sqrt{\frac{1}{p_i sn} - \frac{1}{n}} \ge \sqrt{\frac{1}{p_i sn}} - \sqrt{\frac{1}{n}}.$$

Therefore, the average transmission distance of content i is at least $\sqrt{\frac{1}{p_i sn}} - \sqrt{\frac{1}{n}}$. Applying the inequality to all m contents, the transmission distance L averaged over all contents will be

$$L = \sum_{i=1}^{m} L_i * \rho_i \ge \sum_{i=1}^{m}\left(\sqrt{\frac{1}{np_i s}} - \sqrt{\frac{1}{n}}\right)\rho_i. \tag{5}$$

Based on the above result, if L^* is the optimal value of the following nonlinear program, then L^* will be a lower bound for the average transmission distance L.

$$\begin{aligned} \min_{\mathrm{P}\in\Omega} \quad & L = f(\mathrm{P}) = \sum_{i=1}^{m}\left(\frac{1}{\sqrt{p_i sn}} - \frac{1}{\sqrt{n}}\right)\rho_i \\ \text{s.t.} \quad & g(\mathrm{P}) = \sum_{i=1}^{m} p_i - 1 \le 0 \end{aligned} \tag{6}$$

where $\Omega = [1/ns, 1/s]^m$. The above convex nonlinear program is referred to as the **primal** problem, and it has the same constraints (2), (3) and (4).

To obtain L^*, let us consider the dual of the above nonlinear program:

$$\begin{aligned} \max \quad & d = D(u) \\ \text{s.t.} \quad & u \ge 0, \end{aligned} \tag{7}$$

where the dual objective function $D(u)$ is given by

$$\begin{aligned} D(u) &= \min_{\mathrm{P}\in\Omega}\left(f(\mathrm{P}) + u \times g(\mathrm{P})\right) \\ &= \min_{\mathrm{P}\in\Omega}\left(\sum_{i=1}^{m}\left(\frac{1}{\sqrt{p_i sn}} - \frac{1}{\sqrt{n}}\right)\rho_i + u\left(\sum_{i=1}^{m} p_i - 1\right)\right). \end{aligned}$$

The above nonlinear program is referred to as the **dual** problem.

Definition 1. *A pair (P^*, u^*) with $\mathrm{P}^* \in \Omega$ and $u^* \ge 0$ satisfies the global optimality conditions for the primal problem, if*

$$(i) \qquad f(\mathrm{P}^*) + u^* g(\mathrm{P}^*) = \min_{\mathrm{P}\in\Omega}\left(f(\mathrm{P}) + u^* g(\mathrm{P})\right) \tag{8}$$

$$(ii) \qquad u^* g(\mathrm{P}^*) = 0 \tag{9}$$

$$(iii) \qquad g(\mathrm{P}^*) \le 0 \tag{10}$$

Lemma 1. *If a pair (P^*, u^*) satisfies the global optimality condition given in Definition 1, P^* is optimal in the primal problem.*

Proof. The detailed proof can be found in [20]. Basically, given Eq. (9), $f(\mathrm{P}^*) = D(u^*)$. Weak duality theorem [20] states that $f(\mathrm{P}) \ge D(u)$ for any feasible P and u, therefore P^* is optimal to the primal problem. □

Solving (8), (9) and (10) gives us P^* which optimizes the primal problem, and $L^* = f(\mathrm{P}^*)$. Recall that the per node capacity is upper bounded by $\Theta(\frac{W}{L\sqrt{n}})$, and L is known to be larger or equal to L^*, then $\Theta(\frac{W}{L^*\sqrt{n}})$ gives an upper bound on the network capacity.

V. Capacity Upper Bound

A. Generalized Harmonic Number

Before solving the nonlinear program, we first give a short discussion of the generalized harmonic number $H_{m,\gamma}$ defined in Eq. (1). It is easy to see that $H_{m,\gamma}$ satisfies

$$\int_1^{m+1}\frac{1}{x^\gamma}\mathrm{d}x \le H_{m,\gamma} = \sum_{j=1}^{m}\frac{1}{j^\gamma} \le 1 + \int_1^m \frac{1}{x^\gamma}\mathrm{d}x.$$

Accordingly,

$$\begin{cases} \log(m+1) \le H_{m,\gamma} \le \log m + 1, & \text{if } \gamma = 1 \\ \dfrac{(m+1)^{1-\gamma} - 1}{1-\gamma} \le H_{m,\gamma} \le \dfrac{m^{1-\gamma} - 1}{1-\gamma} + 1, & \text{otherwise} \end{cases} \tag{11}$$

As $m \to \infty$, $H_{m,\gamma}$ converges if and only if $\gamma > 1$.

B. Preliminaries

The partial derivative of $f(\mathrm{P}) + u^* g(\mathrm{P})$ over any p_i is

$$\frac{\partial\left(f(\mathrm{P}) + u^* g(\mathrm{P})\right)}{\partial p_i} = u^* - \frac{1}{2}\frac{\rho_i p_i^{-3/2}}{\sqrt{sn}}. \tag{12}$$

The above derivative monotonically increases with p_i, and as $u^* \ge 0$ (the constraint of the dual), the function $f(\mathrm{P}) + u^* g(\mathrm{P})$ will first decrease with p_i, and then increase with p_i when p_i

is larger than $(\rho_i/2u^*)^{2/3}$. Based on Eq. (8), P* minimizes $f(\mathrm{P})+u^*g(\mathrm{P})$, and $\frac{1}{sn} \le p_i \le \frac{1}{s}$ $(1 \le i \le m)$, hence p_i^* is

$$p_i^* = \begin{cases} \frac{1}{s}, & \text{if } (\frac{\rho_i}{2u^*\sqrt{sn}})^{\frac{2}{3}} \ge \frac{1}{s} \\ (\frac{\rho_i}{2u^*\sqrt{sn}})^{\frac{2}{3}}, & \text{if } \frac{1}{sn} < (\frac{\rho_i}{2u^*\sqrt{sn}})^{\frac{2}{3}} < \frac{1}{s} \\ \frac{1}{sn}, & \text{if } (\frac{\rho_i}{2u^*\sqrt{sn}})^{\frac{2}{3}} \le \frac{1}{sn} \end{cases} \quad (13)$$

Due to the convexity of the primal problem and the condition $\frac{m}{n} \le s \ll m$, P* exists and is unique. To simplify the notations, let μ be the number of contents with density $\frac{1}{s}$, i.e., when $(\frac{\rho_1}{2u^*\sqrt{sn}})^{\frac{2}{3}} \ge \frac{1}{s}$, μ is the non-negative integer satisfying

$$\left(\frac{\rho_{\mu+1}}{2u^*\sqrt{sn}}\right)^{\frac{2}{3}} < \frac{1}{s} \le \left(\frac{\rho_\mu}{2u^*\sqrt{sn}}\right)^{\frac{2}{3}}. \quad (14)$$

When μ is reasonably large, $(\frac{\rho_\mu}{2u^*\sqrt{sn}})^{\frac{2}{3}} \approx \frac{1}{s}$, and

$$\mu \approx \left(\frac{1}{\sqrt{n}}\frac{s}{2u^*H_{m,\gamma}}\right)^{\frac{1}{\gamma}}. \quad (15)$$

Let λ be the number of contents with density $\frac{1}{sn}$, then when $(\frac{\rho_m}{2u^*\sqrt{sn}})^{\frac{2}{3}} \le \frac{1}{sn}$, λ is the non-negative integer that satisfies

$$\left(\frac{\rho_{m-\lambda+1}}{2u^*\sqrt{sn}}\right)^{\frac{2}{3}} \le \frac{1}{sn} < \left(\frac{\rho_{m-\lambda}}{2u^*\sqrt{sn}}\right)^{\frac{2}{3}}. \quad (16)$$

If λ is not too close to m, $(\frac{\rho_{m-\lambda}}{2u^*\sqrt{sn}})^{\frac{2}{3}} \approx \frac{1}{sn}$, and

$$m-\lambda \approx \left(\frac{sn}{2u^*H_{m,\gamma}}\right)^{\frac{1}{\gamma}}. \quad (17)$$

In this way, among all m contents, the most popular μ contents each has a density of $\frac{1}{s}$, and the most unpopular λ contents each has a density of $\frac{1}{sn}$, and the density for each of the remaining contents is $(\frac{\rho_i}{2u^*\sqrt{sn}})^{\frac{2}{3}}$. The constraint of the primal problem must be tight, take these density results back to Eq. (9) leads to:

$$\frac{\mu}{s} + \left(\frac{1}{2u^*\sqrt{sn}}\right)^{\frac{2}{3}} \sum_{i=\mu+1}^{m-\lambda} \left(\frac{\frac{1}{i^\gamma}}{H_{m,\gamma}}\right)^{\frac{2}{3}} + \frac{\lambda}{sn} = 1. \quad (18)$$

In the next subsection, we solve the above equation where μ and λ are restricted by inequalities (14) and (16), respectively.

C. Solution to the Primal Problem

We solve the primal problem based on various values of Zipf parameter γ: $\gamma < \frac{3}{2}$, $\gamma = \frac{3}{2}$ and $\gamma > \frac{3}{2}$.

(i) $\gamma < \frac{3}{2}$

Lemma 2. *In case of $\gamma \le \frac{3}{2}$, $\mu = 0$.*

Proof. We use contradiction to prove $\mu = 0$. Assume $\mu \ge 1$, based on inequality (14), for the most popular content, we have $(\frac{\rho_1}{2u^*\sqrt{sn}})^{2/3} \ge \frac{1}{s}$, which leads to:

$$\left(\frac{1}{2u^*\sqrt{sn}}\right)^{\frac{2}{3}} \ge \frac{(H_{m,\gamma})^{\frac{2}{3}}}{s}. \quad (19)$$

Consequently,

$$\frac{\mu}{s} + \sum_{i=\mu+1}^{m-\lambda}\left(\frac{\rho_i}{2u^*\sqrt{sn}}\right)^{\frac{2}{3}} \ge \sum_{i=1}^{m-\lambda}\left(\frac{\rho_i}{2u^*\sqrt{sn}}\right)^{\frac{2}{3}} \ge \frac{(H_{m,\gamma})^{2/3}}{s}\frac{H_{m-\lambda,2\gamma/3}}{(H_{m-\lambda,\gamma})^{2/3}} \ge \frac{H_{m-\lambda,2\gamma/3}}{s}.$$

Combining inequality (19) and (16), when $\lambda \ge 1$, $(m-\lambda+1)^{\frac{2\gamma}{3}} \ge n$, which means $m-\lambda+1 \ge n^{3/2\gamma}$; otherwise (i.e., $\lambda = 0$), $m-\lambda+1 = m+1$. Combining the two cases of $\lambda \ge 1$ and $\lambda = 0$, we have $(m-\lambda+1) \ge \min(m+1, n^{3/2\gamma})$. As $m \to \infty$ and $n \to \infty$, $m-\lambda \to \infty$. Since $2\gamma/3 \le 1$, $H_{m-\lambda,2\gamma/3}$ goes to infinity, and

$$\frac{\mu}{s} + \sum_{i=\mu+1}^{m-\lambda}\left(\frac{\rho_i}{2u^*}\right)^{\frac{2}{3}} \gg 1.$$

The above inequality contradicts Eq. (18), which suggests that our assumption of $\mu \ge 1$ when $\gamma \le \frac{3}{2}$ cannot be true. □

For the value of λ we have the following lemma.

Lemma 3. *When $\gamma < \frac{3}{2}$, $\lambda = 0$ if and only if $m < \left(1-\frac{2}{3}\gamma\right)sn$.*

Proof. Based on Lemma 2, when $\gamma < \frac{3}{2}$, $\mu = 0$. By assuming $\lambda = 0$, Eq. (18) becomes

$$\left(\frac{1}{2u^*\sqrt{sn}}\right)^{\frac{2}{3}} \sum_{i=1}^{m}\left(\frac{\frac{1}{i^\gamma}}{H_{m,\gamma}}\right)^{\frac{2}{3}} = 1. \quad (20)$$

Since no content has a density of $\frac{1}{sn}$, according to Eq. (13), even for the most unpopular content, we have

$$\left(\frac{\rho_m}{2u^*\sqrt{sn}}\right)^{\frac{2}{3}} = \left(\frac{1}{2u^*\sqrt{sn}}\right)^{\frac{2}{3}}\left(\frac{\frac{1}{m^\gamma}}{H_{m,\gamma}}\right)^{\frac{2}{3}} > \frac{1}{sn}. \quad (21)$$

Combining Eq. (20) and inequality (21) leads to

$$\left(\frac{1}{2u^*\sqrt{sn}}\right)^{\frac{2}{3}}\left(\frac{\frac{1}{m^\gamma}}{H_{m,\gamma}}\right)^{\frac{2}{3}} = \frac{\frac{1}{m^{2\gamma/3}}}{H_{m,2\gamma/3}} = \frac{1-\frac{2}{3}\gamma}{m} > \frac{1}{sn},$$

which is the same as $m < \left(1-\frac{2}{3}\gamma\right)sn$.

On the other hand, when both Eq. (20) and inequality (21) hold, it is guaranteed that $\lambda = 0$. Hence, when $\gamma < \frac{3}{2}$, $\lambda = 0$ if and only if $m < \left(1-\frac{2}{3}\gamma\right)sn$. □

Based on the above two lemmas, we have the following:

Proposition 1. *When $\gamma < \frac{3}{2}$, a lower bound on the average transmission distance is*

$$L^* = \begin{cases} \frac{1}{\sqrt{sn}}\frac{(H_{m,2\gamma/3})^{\frac{3}{2}}}{H_{m,\gamma}}, & \text{if } m < (1-\frac{2}{3}\gamma)ns \\ \frac{(m-\lambda)^{1-\gamma}}{(1-\frac{2}{3}\gamma)H_{m,\gamma}} + 1 - \frac{H_{m-\lambda,\gamma}}{H_{m,\gamma}}, & \text{otherwise} \end{cases}$$

where $m-\lambda = (sn-m)(\frac{3}{2\gamma}-1)$.

The proof of this proposition can be found in the Appendix.

(ii) $\gamma = \frac{3}{2}$

Similar to Proposition 1, we have the following:

Proposition 2. *When $\gamma = \frac{3}{2}$, a lower bound on the average transmission distance is*

$$L^* = \begin{cases} \frac{1}{\sqrt{sn}}\frac{(H_{m,1})^{\frac{3}{2}}}{H_{m,3/2}}, & \text{if } m\ln m < ns \\ \frac{2(\kappa^{-\frac{1}{2}} - m^{-\frac{1}{2}}) + \kappa^{-\frac{1}{2}}\ln\kappa}{H_{m,3/2}}, & \text{otherwise} \end{cases}$$

where κ satisfies $\kappa(\ln\kappa - 1) = sn - m$.

The proof of this proposition is similar to the proof of Proposition 1. The basic idea is to combine the solution of Eq. (18) when $\gamma = \frac{3}{2}$ with the objective of the primal

problem, and the details are omitted due to space limitation.

(iii) $\gamma > \frac{3}{2}$

Proposition 3. *When $\gamma > \frac{3}{2}$, a lower bound on the average transmission distance is*

$$L^* = \begin{cases} \frac{1}{\sqrt{sn}} \frac{\left(H_{m,2\gamma/3}\right)^{\frac{3}{2}}}{H_{m,\gamma}}, & \text{if } s < \chi_1 \text{ and } sn > \chi_2 \\ \frac{1}{\sqrt{sn-m}} \frac{\left(H_{m,2\gamma/3}\right)^{\frac{3}{2}}}{H_{m,\gamma}}, & \text{if } s < \chi_3 \text{ and } sn \le \chi_2 \\ \frac{\left(\frac{2\gamma-3}{2\gamma}\left(s-\frac{m}{n}\right)\right)^{1-\gamma}}{\sqrt{n}H_{m,\gamma}(\frac{2}{3}\gamma-1)}, & \text{if } s \ge \chi_3 \text{ and } sn \le \chi_4 \\ \frac{\left(\frac{2\gamma-3}{2\gamma}s\right)^{1-\gamma}}{\sqrt{n}H_{m,\gamma}(\frac{2}{3}\gamma-1)}, & \text{if } s \ge \chi_1 \text{ and } sn > \chi_4 \end{cases}$$

where $\chi_1 = H_{m,2\gamma/3}$, $\chi_2 = m^{2\gamma/3}H_{m,2\gamma/3}$, $\chi_3 = \frac{m}{n} + \frac{2\gamma}{2\gamma-3}$ and $\chi_4 = \frac{2\gamma}{2\gamma-3}\frac{m}{n^{3/2\gamma-1}} + m$.

Similarly, the proof can be obtained by combing the solution of Eq. (18) given $\gamma > \frac{3}{2}$ with the objective of the primal problem, and the details are omitted here for conciseness.

D. Scaling Laws of Network Capacity

Based on previous results on the average transmission distance, we now present an upper bound of the per node capacity. In [1], the authors have shown that the per node capacity is upper bounded by $\frac{\sqrt{8}}{\sqrt{\pi}}\frac{W}{\beta^{\frac{1}{\alpha}}-1}\frac{1}{L\sqrt{n}}$, where L is the average transmission distance. Combining this result with the lower bounds on average transmission distance given in propositions 1, 2 and 3, we have the following theorem.

Theorem 1. *Under our network and content access model, an upper bound C on the per node capacity (with constants $\frac{\sqrt{8}}{\sqrt{\pi}}\frac{W}{\beta^{\frac{1}{\alpha}}-1}$ omitted) is given as follows.*

(i) $\gamma < \frac{3}{2}$

If $m < (1-\frac{2}{3}\gamma)sn$:

$$C = \begin{cases} \sqrt{\frac{s}{m}}\frac{\left(1-\frac{2}{3}\gamma\right)^{\frac{3}{2}}}{1-\gamma}, & \text{if } \gamma < 1 \\ 3^{-\frac{3}{2}}\sqrt{\frac{s}{m}}\ln m, & \text{if } \gamma = 1 \\ H_{m,\gamma}\frac{\sqrt{s}}{m^{\frac{3}{2}-\gamma}}\left(1-\frac{2}{3}\gamma\right)^{\frac{3}{2}}, & \text{if } 1 < \gamma < \frac{3}{2} \end{cases}$$

If $m \ge (1-\frac{2}{3}\gamma)sn$:

$$C = \begin{cases} \frac{1}{\sqrt{n}}\frac{1}{1-\frac{1}{2}(\frac{3}{2\gamma}-1)^{-\gamma}\left(\frac{sn}{m}-1\right)^{1-\gamma}}, & \text{if } \gamma < 1 \\ \frac{1}{\sqrt{n}}\frac{\ln m}{\ln\frac{2m}{sn-m}+3}, & \text{if } \gamma = 1 \\ \frac{1}{\sqrt{n}}\frac{H_{m,\gamma}\frac{\gamma-1}{m^{1-\gamma}}}{\frac{1}{2}(\frac{3}{2\gamma}-1)^{-\gamma}(\frac{sn}{m}-1)^{1-\gamma}-1}, & \text{if } 1 < \gamma < \frac{3}{2} \end{cases}$$

(ii) $\gamma = \frac{3}{2}$

$$C = \begin{cases} \frac{\sqrt{s}}{(\ln m)^{\frac{3}{2}}}H_{m,3/2}, & \text{if } m\ln m < sn \\ \frac{1}{\sqrt{n}}\frac{H_{m,3/2}}{2(\kappa^{-\frac{1}{2}}-m^{-\frac{1}{2}})+\kappa^{-\frac{1}{2}}\ln\kappa}, & \text{if } m\ln m \ge sn \end{cases}$$

where κ satisfies $\kappa(\ln\kappa - 1) = sn - m$.

(iii) $\gamma > \frac{3}{2}$

$$C = \begin{cases} \sqrt{s}\frac{H_{m,\gamma}}{\left(H_{m,2\gamma/3}\right)^{\frac{3}{2}}}, & \text{if } s < \chi_1 \text{ and } sn > \chi_2 \\ \sqrt{s-\frac{m}{n}}\frac{H_{m,\gamma}}{\left(H_{m,2\gamma/3}\right)^{\frac{3}{2}}}, & \text{if } s < \chi_3 \text{ and } sn \le \chi_2 \\ \left(s-\frac{m}{n}\right)^{\gamma-1}\frac{\frac{2}{3}\gamma H_{m,\gamma}}{\left(\frac{2\gamma-3}{2\gamma}\right)^{-\gamma}}, & \text{if } s \ge \chi_3 \text{ and } sn \le \chi_4 \\ s^{\gamma-1}\frac{\frac{2}{3}\gamma H_{m,\gamma}}{\left(\frac{2\gamma-3}{2\gamma}\right)^{-\gamma}}, & \text{if } s \ge \chi_1 \text{ and } sn > \chi_4 \end{cases}$$

where $H_{m,\gamma}$ is the generalized harmonic number, and the values of χ_1, χ_2, χ_3 and χ_4 are given in Proposition 3.

Proof. Let $\bar{p}_i$ denote the fraction of nodes that have not cached content i locally, and let $\bar{L}_i$ represent the average distance for those nodes to retrieve i from others. Then, the average distance for nodes to retrieve contents from others (averaged over all contents), denoted by $\bar{L}$, is

$$\bar{L} = \frac{\sum_{i=1}^{m}\rho_i\bar{p}_i\bar{L}_i}{\sum_{i=1}^{m}\rho_i\bar{p}_i}.$$

According to [1], the number of bits each node can receive from others per second is upper bounded by $\frac{1}{\bar{L}\sqrt{n}}$. The probability that a requested content is not cached locally is $\bar{p} = \sum_{i=1}^{m}\rho_i\bar{p}_i$. Therefore, besides the bits received from others, the number of bits obtained from local cache to serve requests is $\frac{1-\bar{p}}{\bar{p}}\frac{1}{\bar{L}\sqrt{n}}$. Combining the bits received from others and the bits obtained from local cache, we can get an upper bound of network capacity as follows:

$$C = \frac{1}{\bar{L}\sqrt{n}} \times \left(1 + \frac{1-\bar{p}}{\bar{p}}\right) = \frac{1}{\sqrt{n}\sum_{i=1}^{m}\rho_i\bar{p}_i\bar{L}_i}. \quad (22)$$

Recall that L_i is the transmission distance of content i averaged over all nodes (including the nodes that have cached i locally), then $L_i = \bar{L}_i\bar{p}_i + 0(1-\bar{p}_i)$, for $i = 1, ..., m$. The transmission distance averaged over all contents is $L = \sum_{i=1}^{m}L_i\rho_i = \sum_{i=1}^{m}\bar{L}_i\bar{p}_i\rho_i$. Then we can rewrite Eq. (22) as

$$C = \frac{1}{L\sqrt{n}}.$$

Combining the above equation with propositions 1, 2 and 3 yields the results in the theorem. □

VI. Analytical Results

A. Capacity and the Number of Nodes

In this subsection, we analyze how the capacity of wireless networks with caching changes with the number of nodes. According to Theorem 1, given a specific γ, the capacity result C is a piecewise function based on n, s and m. We first show C is continuous for any $\gamma > 0$.

Lemma 4. *For any $\gamma > 0$, the network capacity function C is continuous on $n \ge \frac{m}{s}$.*

Proof. To show that C is continuous, we just need to prove C is continuous at the critical points (i.e., the point that connects two sub-functions of the piecewise function).

(i) $\gamma < \frac{3}{2}$

In this case, there is only one critical point $n_0 = \frac{m}{s(1-2\gamma/3)}$. The comparisons of the right-hand limit and the actual value of the capacity at the critical point for different γ are given as follows:

$$\lim_{n\to n_0^+} C = \sqrt{\frac{s}{m}}\frac{(1-\frac{2}{3}\gamma)^{\frac{3}{2}}}{1-\gamma} = \frac{1}{\sqrt{n_0}}\frac{1}{1-\frac{1}{2}\frac{1}{3/2\gamma-1}}, \quad \gamma < 1$$

$$\lim_{n\to n_0^+} C = 3^{-\frac{3}{2}}\sqrt{\frac{s}{m}}\ln m = \sqrt{\frac{s}{3m}}\frac{\ln m}{\ln\frac{2m}{3m-m}+3}, \quad \gamma = 1$$

$$\lim_{n\to n_0^+} C = \frac{H_{m,\gamma}\sqrt{s}}{m^{\frac{3}{2}-\gamma}}(1-\frac{2}{3}\gamma)^{\frac{3}{2}} = \frac{1-\frac{2}{3}\gamma}{\sqrt{n_0}}\frac{H_{m,\gamma}}{m^{1-\gamma}}, \quad 1<\gamma<\frac{3}{2}.$$

(ii) $\gamma = \frac{3}{2}$

At the only critical point $n_0 = \frac{m\ln m}{s}$, we have

$$\lim_{n\to n_0^+} C = \frac{\sqrt{s}H_{m,3/2}}{(\ln m)^{\frac{3}{2}}} = \frac{\sqrt{s}}{\sqrt{m\ln m}}\frac{H_{m,3/2}}{m^{-\frac{1}{2}}\ln m}, \quad \text{if } \gamma = \frac{3}{2}.$$

(iii) $\gamma > \frac{3}{2}$

There are two critical points. At the first critical point $n_0 = \frac{m^{2\gamma/3}H_{m,2\gamma/3}}{s}$, we have $m \ll n_0$, and the right-hand limit and the actual value are equal

$$\lim_{n\to n_0^+} C = \frac{\sqrt{s}H_{m,\gamma}}{\left(H_{m,2\gamma/3}\right)^{\frac{3}{2}}} = \lim_{\frac{m}{n}\to 0}\frac{\sqrt{s-\frac{m}{n}}H_{m,\gamma}}{\left(H_{m,2\gamma/3}\right)^{\frac{3}{2}}}, \quad \text{if } \gamma > \frac{3}{2}.$$

The other critical point denoted by n_0', satisfies $n_0' = \frac{m}{s}(1+\frac{2\gamma}{2\gamma-3}{n_0'}^{1-3/2\gamma})$. At n_0', we have $m \ll n_0'$ and the right-hand limit and the actual value are equal

$$\lim_{n\to {n_0'}^+} C = \frac{\frac{2}{3}\gamma s^{\gamma-1}H_{m,\gamma}}{\left(\frac{2\gamma-3}{2\gamma}\right)^{-\gamma}} = \lim_{\frac{m}{n}\to 0}\frac{\frac{2}{3}\gamma\left(s-\frac{m}{n}\right)^{\gamma-1}H_{m,\gamma}}{\left(\frac{2\gamma-3}{2\gamma}\right)^{-\gamma}}, \quad \text{if } \gamma > \frac{3}{2}.$$

Since in all three cases, the right-hand limit and the actual value are always equal at the critical points, C is continuous. □

Regarding the monotonicity of the capacity C, we have the following lemma.

Lemma 5. *For any $\gamma > 0$, the capacity C monotonically increases with the number of nodes n.*

Proof. We have shown in Lemma 4 that C is continuous, to show C monotonically increases, we just need to prove each sub-function of C monotonically increases with n.

(i) $\gamma < 1$

When $m < (1-\frac{2}{3}\gamma)sn$, C is independent of n. We just need to prove C monotonically increases with n when $m \geq (1-\frac{2}{3}\gamma)sn$. Let $t = \sqrt{n}(1-\frac{1}{2}(\frac{3}{2\gamma-1})^{-\gamma}(\frac{sn}{m}-1)^{1-\gamma})$, which is the denominator of C. Then

$$\frac{\mathrm{d}t}{\mathrm{d}n} = \frac{n^{-\frac{1}{2}}}{4}\tilde{t}, \tag{23}$$

where $\tilde{t} = \frac{1}{2} - \frac{1}{4}((\frac{3}{2\gamma}-1)(\frac{sn}{m}-1))^{-\gamma}(\frac{sn}{m}-1+\frac{2sn}{m}(1-\gamma))$. Let $x = (\frac{sn}{m}-1)$, then $0 \leq x = \frac{sn}{m}-1 \leq \frac{2\gamma}{3-2\gamma}$. The derivative of $\tilde{t}$ over x is

$$\frac{\mathrm{d}\tilde{t}}{\mathrm{d}x} = -(3/2\gamma-1)^{-\gamma}(1-\gamma)((3-2\gamma)x-2\gamma)x^{-\gamma-1} \geq 0.$$

As $\frac{\mathrm{d}\tilde{t}}{\mathrm{d}x} \geq 0$, $\tilde{t}$ monotonically increases with x, and the maximum value of $\tilde{t}$ is 0 when $x = \frac{2\gamma}{3-2\gamma}$. Combining $\tilde{t} \leq 0$ with Eq. (23) leads to $\frac{\mathrm{d}t}{\mathrm{d}n} \leq 0$. Recall $C = \frac{1}{t}$, then $\frac{\mathrm{d}C}{\mathrm{d}n} \geq 0$, and we have C monotonically increases with n.

(ii) $\gamma = 1$

When $\gamma = 1$ and $m < (1-\frac{2}{3}\gamma)sn$, the capacity C is independent of n. Hence, to show C monotonically increases, we only need to prove C increases with n when $m \geq (1-\frac{2}{3}\gamma)sn$. Let t denote the denominator of C, $t = \sqrt{n}(\ln\frac{2m}{sn-m}+3)$. Then

$$\begin{aligned}\frac{\mathrm{d}t}{\mathrm{d}n} &= \frac{1}{2}n^{-\frac{1}{2}}\left(\ln\frac{2m}{sn-m}+3\right) - \frac{\sqrt{n}s}{sn-m}\\ &= \frac{1}{2\sqrt{n}}\left(\ln\frac{2m}{sn-m}+1-\frac{2m}{sn-m}\right).\end{aligned}$$

As $m \geq \frac{1}{3}sn$, $\frac{2m}{sn-m} \geq 1$. Also note $x - \ln x \geq 1$ holds for all $x \geq 1$, thus $\frac{\mathrm{d}t}{\mathrm{d}n} \leq 0$. In this way, t monotonically decreases with n, and C increases with n.

(iii) $1 < \gamma < \frac{3}{2}$

When $m < (1-\frac{2}{3}\gamma)sn$, C is independent of n. When $m \geq (1-\frac{2}{3}\gamma)sn$, the numerator of C is independent of n, hence we only focus on the denominator of C, and let $t = \sqrt{n}(\frac{1}{2}(\frac{3}{2\gamma-1})^{-\gamma}(\frac{sn}{m}-1)^{1-\gamma}-1)$. The proof here is similar to the proof of case $\gamma < 1$, as t in both cases only differ by a negative sign. Consequently, the only difference in the proof is when showing $\frac{\mathrm{d}\tilde{t}}{\mathrm{d}x} \geq 0$. Previously we have $1-\gamma > 0$ for $\gamma < 1$, and now we have $-(1-\gamma) > 0$ for $1 < \gamma < \frac{3}{2}$.

(iv) $\gamma = \frac{3}{2}$

When $sn > m\ln m$, the capacity C is independent of n. For $sn \leq m\ln m$, let t denote the denominator of C, then

$$\frac{\mathrm{d}t}{\mathrm{d}n} = \frac{1}{\sqrt{n}}((\kappa^{-\frac{1}{2}} - m^{-\frac{1}{2}}) + \frac{\kappa^{-\frac{1}{2}}}{2}\ln\kappa) - \sqrt{n}(\frac{\kappa^{-\frac{3}{2}}}{2}\ln\kappa)\frac{\mathrm{d}\kappa}{\mathrm{d}n}. \tag{24}$$

Recall that κ satisfies $\kappa(\ln\kappa - 1) = sn - m$, then we have

$$\begin{aligned}\frac{\mathrm{d}(\kappa\ln\kappa-\kappa)}{\mathrm{d}\kappa}\frac{\mathrm{d}\kappa}{\mathrm{d}n} &= \frac{\mathrm{d}(sn-m)}{\mathrm{d}n}\\ \frac{\mathrm{d}\kappa}{\mathrm{d}n} &= \frac{s}{\ln\kappa}.\end{aligned} \tag{25}$$

Combining Eq. (24) with Eq. (25) leads to

$$\begin{aligned}\frac{\mathrm{d}t}{\mathrm{d}n} &= \frac{1}{2\sqrt{n}}(2\kappa^{-\frac{1}{2}} - 2m^{-\frac{1}{2}} + \kappa^{-\frac{1}{2}}\ln\kappa - ns\kappa^{-\frac{3}{2}})\\ &= \frac{1}{2\sqrt{n}}(2\kappa^{-\frac{1}{2}} - 2m^{-\frac{1}{2}} - m\kappa^{-\frac{3}{2}}).\end{aligned}$$

As $sn \leq m\ln m$ and $\kappa(\ln\kappa - 1) = sn - m$, $\kappa \leq m$. Then the above derivative $\frac{\mathrm{d}t}{\mathrm{d}n} \leq 0$, t decreases with n, and C monotonically increases with n when $\gamma = \frac{3}{2}$.

(v) $\gamma > \frac{3}{2}$

In this case, based on Theorem 1, the capacity C which is a piecewise function has four sub-functions. Apparently, for the first and forth sub-function, the network capacity is independent of n. In the other two sub-functions, only the term $(s-\frac{m}{n})$ is affected by n, since $-\frac{m}{n}$ increases as n increases, C increases with n. □

This result suggests that the network capacity will not decrease when the number of nodes increases. More interestingly, in some cases, it is even possible for the per node capacity to increase as n increases.

B. Effects of Local Cache

In this subsection, we analyze the probability that a request is served locally, when contents have been optimally cached.

Let φ denote the probability that a request is served directly by local cache, we have the following lemma.

Lemma 6. *When contents have been optimally cached to maximize the network capacity, for any $\gamma \leq \frac{3}{2}$, $\varphi \to 0$.*

Proof. Given P^*, the probability that requests being satisfied by local cache is $\varphi = \sum_{i=1}^{m} \rho_i p_i^* s$. Based on Lemma 2, $\mu = 0$ for $\gamma \leq \frac{3}{2}$. As $\mu = 0$, Eq. (18) is

$$1 = \sum_{i=1}^{m-\lambda} (\frac{\rho_i}{2u^*\sqrt{sn}})^{\frac{2}{3}} + \frac{\lambda}{sn} \geq \sum_{i=1}^{m} (\frac{\rho_i}{2u^*\sqrt{sn}})^{\frac{2}{3}}. \quad (26)$$

Since $\sum_{i=1}^{m} \rho_i^{\frac{2}{3}} = \frac{H_{m,2\gamma/3}}{(H_{m,\gamma})^{2/3}}$, the above inequality (26) is equivalent to $\frac{(H_{m,\gamma})^{2/3}}{H_{m,2\gamma/3}} \geq \frac{1}{(2u^*\sqrt{sn})^{2/3}}$. Combining the previous inequality and the value of p_i^* given in Eq. (13), and let $\tilde{p_i^*} = \frac{(\rho_i H_{m,\gamma})^{2/3}}{H_{m,2\gamma/3}}$, we can get

$$\tilde{p_1^*} - p_1^* \geq \tilde{p_2^*} - p_2^* \geq ... \geq \tilde{p_m^*} - p_m^*. \quad (27)$$

Based on the definition of Zipf distribution, the popularity of contents satisfies $\rho_1 \geq \rho_2 \geq ... \geq \rho_m > 0$. Combining this content popularity inequality with inequality (27), and as $\sum_{i=1}^{m} \tilde{p_i^*} - p_i^* = 0$, we have $\sum_{i=1}^{m} \rho_i(\tilde{p_i^*} - p_i^*)s \geq 0$, which leads to:

$$\sum_{i=1}^{m} \rho_i \tilde{p_i^*} s \geq \sum_{i=1}^{m} \rho_i p_i^* s.$$

The left-hand side of the above inequality is

$$\sum_{i=1}^{m} \rho_i \tilde{p_1^*} s = \sum_{i=1}^{m} \frac{s}{i^\gamma H_{m,\gamma}} \frac{1}{i^{\frac{2}{3}\gamma} H_{m,2\gamma/3}} = \frac{sH_{m,5\gamma/3}}{H_{m,\gamma}H_{m,2\gamma/3}}.$$

Since $\frac{sH_{m,5\gamma/3}}{H_{m,\gamma}H_{m,2\gamma/3}} \to 0$ for any $\gamma \leq \frac{3}{2}$, $\sum_{i=1}^{m} \rho_i p_i^* s = \varphi \to 0$. □

The above lemma states that when $\gamma \leq \frac{3}{2}$, only a small portion of requests are served locally, and the effect of the local cache is negligible. On the other hand, when $\gamma > \frac{3}{2}$, the content access is mainly focused on a few popular contents which may have been cached locally, and the local cache will have a more significant effect on network capacity. In the following lemma, we give two sufficient conditions for a constant portion of requests being served by local cache.

Lemma 7. *When contents have been optimally cached to maximize the network capacity, for any $\gamma > \frac{3}{2}$, if $\mu \neq 0$ or $\mu = \lambda = 0$, there exists a constant c independent of m, s and n, such that $\varphi \geq c$.*

Proof. **(i) $\mu \neq 0$**

In case of $\mu \geq 1$, $p_1 = \frac{1}{s}$ and every node caches the most popular content locally. Then, requests for that content is always served by the local cache, and φ must satisfy

$$\varphi > \rho_1 = \frac{1}{H_{m,\gamma}} > \frac{1}{1 + \frac{1}{\gamma - 1}} = \frac{\gamma - 1}{\gamma}.$$

(ii) $\mu = \lambda = 0$

When both λ and μ are 0, we have

$$\varphi = \sum_{i=1}^{m} \rho_i p_i^* s = \frac{sH_{m,5\gamma/3}}{H_{m,\gamma}H_{m,2\gamma/3}} \geq \frac{H_{m,5\gamma/3}}{H_{m,\gamma}H_{m,2\gamma/3}}.$$

As $m \to \infty$, $\frac{H_{m,5\gamma/3}}{H_{m,\gamma}H_{m,2\gamma/3}}$ converges to a constant which only depends on γ. Therefore, φ is larger than a constant when $\lambda = \mu = 0$. □

C. Influence of Other Parameters

Similar to the analysis in subsection VI-A, we now analyze how the cache size, the number of unique content, and the zipf parameter affect the network capacity. Based on theorem 1, we have the following observations.

First, as cache size s grows, the network capacity increases. For simple cases like $\gamma = \frac{3}{2}$ and $m \ln m < sn$, by omitting the insignificant constant (i.e., $H_{m,3/2}$), we can see that the capacity scales like $\frac{\sqrt{s}}{(\ln m)^{3/2}}$. Accordingly, the capacity grows with s. For more complicated cases like $\gamma = \frac{3}{2}$ and $m \ln m \geq sn$, derivative of capacity over cache size can be used to prove that the capacity increases with the cache size. This observation is valid because as s grows, more requests can be served locally or by nearby nodes. Furthermore, when γ is large, increasing the cache size has a greater impact on network capacity. The capacity scales like $\sqrt{\frac{s}{m}}$ when $\gamma < 1$, and scales like $s^{\gamma-1}$ when $\gamma > \frac{3}{2}$.

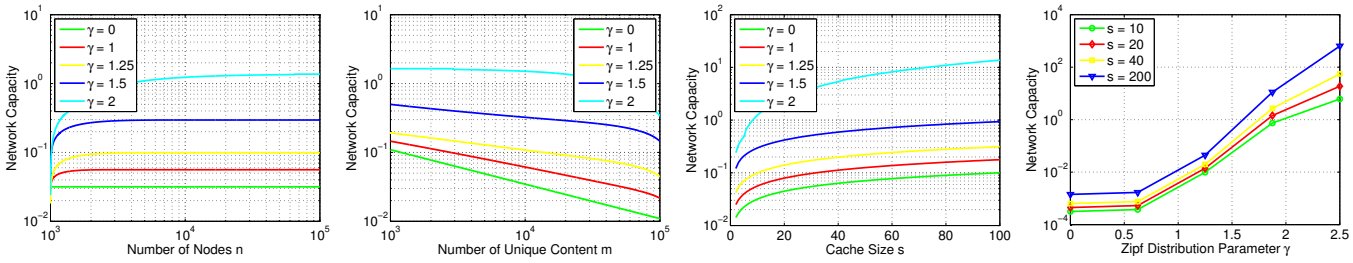
Second, with a similar approach as above, we find that the network capacity decreases when the number of unique content grows. This is because a larger m reduces the probability that the requests are served locally or by close neighbors. More specifically, as γ increases from 0 to more than $\frac{3}{2}$, m has a smaller effect on network capacity. When $\gamma < 1$, the capacity scales like $\sqrt{\frac{s}{m}}$, which means increasing m will greatly reduce the capacity. On the other hand, when $\gamma > \frac{3}{2}$, the capacity scales like $\sqrt{s}$ or $s^{\gamma-1}$, which is almost irrelevant to m.

Third, networks with larger γ have higher capacity. When $\gamma = 0$, content popularity follows uniform distribution, and the capacity scales like $\sqrt{\frac{s}{m}}$. As γ grows from 0 to more than $\frac{3}{2}$, the network capacity increases drastically, since the exponent of m grows from $-\frac{1}{2}$ to 0. This result suggests that caching is more effective when contents have skewed popularity.

VII. Numerical Results

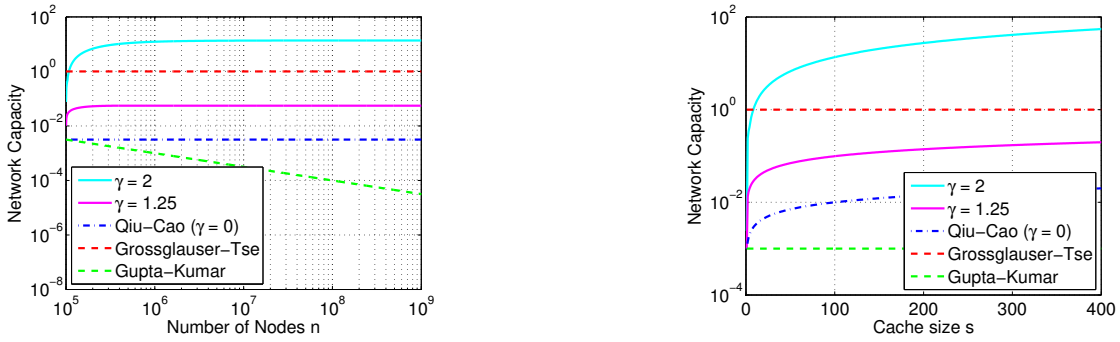
A. Effects of Various Parameters

Based on above theoretical analysis, we now present some numerical results to illustrate how various parameters affect the capacity. Fig. 1(a) shows how the network capacity varies with the number of nodes (n) given different values of γ, where $s = 10$ and $m = 10^4$. As shown in the figure, the network capacity increases as γ increases. When γ is small, the content access is more like a uniform distribution (note that $\gamma = 0$ corresponds to the uniform distribution). When γ is large, the content access is focused on some hot (frequently accessed) content which may have been cached, and then improving the network capacity. Under all five values of γ, the per node capacity first increases with n, and then remains constant. This is different from Gupta-Kumar's result where cache is not considered and the per node capacity decreases with increasing n. With caching, the per node capacity will not decrease as n increases. Moreover, it shows that when n is relatively small, it is possible that increasing n will improve the network capacity. This is because when n is relatively small, most contents only have one replica, and nodes need to traverse the whole network to obtain the contents. On the other



(a) Number of nodes vs. capacity ($s = 10$, $m = 10^4$)
(b) Number of unique content vs. capacity ($s = 12$, $n = 10^4$)
(c) Cache size vs. capacity ($m = 10^4$, $n = 10^4$)
(d) Zipf Parameter γ vs. Capacity ($m = 10^8$, $n = 10^8$)

Fig. 1: Effects of various parameters on capacity



(a) Network capacity vs. number of nodes ($s = 100$, $m = 10^7$)
(b) Network capacity vs. cache size ($n = 10^6$, $m = 10^6$)

Fig. 2: Comparisons to existing work

hand, when n is relatively large, some of the contents could be cached by nearby nodes, which makes content retrieval easier and then improves the network capacity.

Fig. 1(b) illustrates how the number of unique content m affects the network capacity. Generally, increasing m results in a reduction in network capacity. This is because with more unique content, nodes will be more likely to retrieve contents from further away nodes, which leads to more interference and less capacity. More interestingly, when γ increases, m has a diminishing effect on capacity. For example, when $\gamma = 0$, the capacity decreases quickly with m; while for $\gamma = 2$, the capacity almost remains constant for small m. This is because the unpopular contents will be less frequently requested when γ increases. When γ is relatively large, increasing m will only add a few extremely unpopular contents, and will hardly affect what contents will be cached and requested.

Fig. 1(c) shows how the cache size s affects network capacity. As can be seen, as the cache size increases, the per node capacity increases. When γ is relatively small, the contents have similar popularity, and caching a few more contents will not result in a large increase in the network capacity. When γ is relatively large, the popular contents are requested more frequently, and only caching a few more popular contents can significantly improve the network capacity.

Fig. 1(d) illustrates how the Zipf parameter γ affects the network capacity. Compared to the three parameters (s, m, n) discussed above, γ has the largest impact on network capacity. As γ grows from 0 to 2.5, the capacity dramatically increases from 10^{-3} to roughly 10^2. Increasing γ can significantly improve the capacity, since a larger γ results in more skewed content popularity where a few popular contents are more frequently requested, which makes caching more effective.

B. Comparisons to Existing Work

In Fig. 2, we compare our capacity results with previous work [1], [2], [8] based on numerical results. From [1], for wireless networks without caching, the per node capacity scales like $\frac{1}{\sqrt{n}}$, which is shown in the figure by the green dashed line. Grossglauser and Tse [2] have proved that when nodes are mobile, the per node capacity can remain constant when n grows, and their result is shown by the red dashed line in the figure. Qiu and Cao [8] have derived that the per node capacity scales like $\sqrt{\frac{s}{m}}$ when the content popularity follows the uniform distribution (i.e., $\gamma = 0$), which is shown by the blue dashed line in the figure. Since our capacity result at $\gamma = 0$ (i.e., $C = \Theta(\sqrt{\frac{s}{m}})$) conforms to their result, this blue line also represents our capacity at $\gamma = 0$. The remaining two solid lines show our capacity results at $\gamma = 1.25$ and $\gamma = 2$, respectively.

In Fig. 2(a), we show how the network capacity changes with the number of nodes, where $s = 100$ and $m = 10^7$. As shown in the figure, when the number of nodes n increases, the capacity result of Gupta-Kumar drops quickly, and the capacity of Grossglauser-Tse and Qiu-Cao remain unchanged. In our approaches, the network capacity will not decrease when n increases. This is a significant improvement compared to the Gupta-Kumar's result; that is, with caching, increasing the number of nodes will not reduce the per node capacity.

Fig. 2(b) illustrates the network capacity as a function of the cache size, where $n = m = 10^6$. Since Gupta-Kumar and Grossglauser-Tse do not consider caching, their results do not change with the cache size. When the cache size is

extremely small ($s = \frac{m}{n}$), under all three values of γ, our network capacity is comparable with Gupta-Kumar results. When s approaches $\frac{m}{n}$, each content only has one replica in the network. Then, retrieving a content is like to communicate with a random node, which is identical to the communication scenario in [1].

As the cache size increases, the network capacity of our approach with various γ increases quickly, and significantly higher than Gupta-Kumar. When compared to Grossglauser-Tse, the result depends on γ. When γ is large (i.e., $\gamma = 2$), the content access is focused on some hot (frequently accessed) content which may have been cached, and then improving the network capacity. As a result, the network capacity of our approach with $\gamma = 2$ significantly outperforms that of Grossglauser-Tse. When γ is relatively small (i.e., $\gamma = 0$), the content access is more like a uniform distribution and the caching advantage is not very high, and our network capacity is lower. However, Grossglauser-Tse has much longer delay since nodes have to wait until they move close to the destination. While compared to Qiu-Cao, our capacity conforms to theirs at $\gamma = 0$. As γ increases (e.g. $\gamma = 1.25$), caching becomes more effective due to more skewed content popularity, and our capacity grows much higher than Qiu-Cao.

VIII. Conclusion

In this paper, we have studied scaling laws of network capacity based on the skewness of content popularity. We found that as the distribution of the content popularity changes from uniform distribution to more skewed distributions, the network capacity quickly increases from $\Theta(\sqrt{\frac{s}{m}})$ to roughly $\Theta(\sqrt{s})$. Moreover, our results suggest that for wireless networks with caching, when contents have skewed popularity, increasing the number of nodes monotonically increases the per node capacity.

References

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

[2] M. Grossglauser and D. Tse, "Mobility increases the capacity of ad-hoc wireless networks," in *IEEE INFOCOM*, 2001.

[3] E. Leonardi and G. L. Torrisi, "Least recently used caches under the shot noise mode," in *IEEE INFOCOM*, 2015.

[4] E. J. Rosensweig, D. S. Menasche, and J. Kurose, "On the steady-state of cache networks," in *IEEE INFOCOM*, 2013.

[5] M. Dehghan, L. Massoulie, D. Towsley, D. Menasche, and Y. Tay, "A utility optimization approach to network cache design," in *IEEE INFOCOM*, 2016.

[6] L. Saino, I. Psaras, and G. Pavlou, "Understanding sharded caching systems," in *IEEE INFOCOM*, 2016.

[7] B. Liu, V. Firoiu, J. Kurose, M. Leung, and S. Nanda, "Capacity of cache enabled content distribution wireless ad hoc networks," in *IEEE MASS*, 2014.

[8] L. Qiu and G. Cao, "Cache increases the capacity of wireless networks," in *IEEE INFOCOM*, 2016.

[9] L. Yin and G. Cao, "Supporting cooperative caching in ad hoc networks," in *IEEE INFOCOM*, 2004.

[10] J. Zhao, P. Zhang, G. Cao, and C. R. Das, "Cooperative caching in wireless p2p networks: Design, implementation, and evaluation," *IEEE Transactions on Parallel and Distributed Systems*, vol. 21, no. 2, pp. 229–241, 2010.

[11] M. Fiore, F. Mininni, C. Casetti, and C.-F. Chiasserini, "To cache or not to cache?" in *IEEE INFOCOM*, 2009.

[12] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless d2d networks," *IEEE Transactions on Information Theory*, vol. 62, no. 2, pp. 849–869, 2016.

[13] U. Niesen, D. Shah, and G. W. Wornell, "Caching in wireless networks," *IEEE Transactions on Information Theory*, vol. 58, no. 10, pp. 6524–6540, 2012.

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

[15] S. Jin and L. Wang, "Content and service replication strategies in multi-hop wireless mesh networks," in *ACM MSWiM*, 2005.

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

[17] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: Evidence and implications," in *IEEE INFOCOM*, 1999.

[18] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in *ACM SIGCOMM conference on Internet measurement*, 2007.

[19] Y. Zhang, J. Zhao, and G. Cao, "Roadcast: A popularity aware content sharing scheme in vanets," in *IEEE ICDCS*, 2009.

[20] D. P. Bertsekas, "Nonlinear programming," 1999.

Appendix: Proof of Proposition 1

Based on Lemma 2, when $\gamma < \frac{3}{2}$, $\mu = 0$.

(i) $m < (1 - \frac{2\gamma}{3})sn$

In this case, λ is also 0. Then $p_i^* = (\frac{\rho_i}{2u^*\sqrt{sn}})^{2/3}$ for all $i = 1, ..., m$, and the optimal value of the primal problem is

$$L^* = \sum_{i=1}^{m} \left(\frac{1}{\sqrt{p_i^* sn}} - \frac{1}{\sqrt{n}} \right) \rho_i = \left(\frac{2u^*}{sn} \right)^{\frac{1}{3}} \frac{H_{m,2\gamma/3}}{(H_{m,\gamma})^{\frac{2}{3}}} - \frac{1}{\sqrt{n}}. \quad (28)$$

As $\lambda = \mu = 0$, Eq. (18) becomes

$$\left(\frac{1}{2u^*\sqrt{sn}} \right)^{\frac{2}{3}} \sum_{i=1}^{m} \left(\frac{\frac{1}{i^\gamma}}{H_{m,\gamma}} \right)^{\frac{2}{3}} = 1. \quad (29)$$

Combining Eq. (28) and Eq. (29), we can get the optimal value of the primal problem is

$$L^* = \frac{1}{\sqrt{sn}} \frac{(H_{m,2\gamma/3})^{\frac{3}{2}}}{H_{m,\gamma}} - \frac{1}{\sqrt{n}} = \frac{1}{\sqrt{sn}} \frac{(H_{m,2\gamma/3})^{\frac{3}{2}}}{H_{m,\gamma}}.$$

Here $\frac{1}{\sqrt{n}}$ can be ignored as $\frac{(H_{m,2\gamma/3})^{3/2}}{H_{m,\gamma}} \gg 1$.

(ii) $m \geq (1 - \frac{2\gamma}{3})sn$

When $m \geq (1 - \frac{2\gamma}{3})sn$, according to Lemma 3, $\lambda \neq 0$. Then, the optimal value is

$$\begin{aligned} L^* &= \sum_{i=1}^{m-\lambda} \left(\frac{1}{\sqrt{p_i^* sn}} - \frac{1}{\sqrt{n}} \right) \rho_i + \sum_{i=m-\lambda+1}^{m} \left(1 - \frac{1}{\sqrt{n}} \right) \rho_i \\ &= \left(\frac{2u^*}{sn} \right)^{\frac{1}{3}} \frac{H_{m-\lambda,2\gamma/3}}{(H_{m,\gamma})^{\frac{2}{3}}} + \frac{H_{m,\gamma} - H_{m-\lambda,\gamma}}{H_{m,\gamma}} - \frac{1}{\sqrt{n}} \quad (30) \\ &= \frac{(m-\lambda)^{1-\gamma}}{(1-\frac{2}{3}\gamma)H_{m,\gamma}} + 1 - \frac{H_{m-\lambda,\gamma}}{H_{m,\gamma}}, \end{aligned}$$

where the last step of the equality is due to $m - \lambda \approx (\frac{sn}{2u^* H_{m,\gamma}})^{\frac{1}{\gamma}}$ in Eq. (17), and $H_{m-\lambda,2\gamma/3} \approx \frac{(m-\lambda)^{1-2\gamma/3}}{1-2\gamma/3}$ in Eq. (11). The remaining task is to find the actual value of $m - \lambda$. Since $\mu = 0$ and $\lambda \neq 0$, Eq. (18) becomes

$$\left(\frac{1}{2u^*\sqrt{sn}} \right)^{\frac{2}{3}} \sum_{i=1}^{m-\lambda} \left(\frac{\frac{1}{i^\gamma}}{H_{m,\gamma}} \right)^{\frac{2}{3}} + \frac{\lambda}{sn} = 1. \quad (31)$$

As $m - \lambda \approx (\frac{sn}{2u^* H_{m,\gamma}})^{\frac{1}{\gamma}}$ and $H_{m-\lambda,2\gamma/3} \approx \frac{(m-\lambda)^{1-2\gamma/3}}{1-2\gamma/3}$, the above Eq. (31) leads to

$$m - \lambda = (sn - m) \left(\frac{3}{2\gamma} - 1 \right).$$