Fundamental Limits on Throughput Capacity in Information-Centric Networks

Bita Azimdoost, Cedric Westphal, *Senior Member, IEEE*, and Hamid R. Sadjadpour, *Senior Member, IEEE*

***Abstract*— Wireless information-centric networks consider storage as one of the network primitives, and propose to cache data within the network in order to improve latency and reduce bandwidth consumption. We study the throughput capacity and latency in an information-centric network when the data cached in each node has a limited lifetime. The results show that with some fixed request and cache expiration rates, the order of the data access time does not change with network growth, and the maximum throughput order is not changing with the network growth in grid networks and is inversely proportional to the number of nodes in one cell in random networks. Comparing these values with the corresponding throughput and latency with no cache capability (throughput inversely proportional to the network size, and latency of order $\sqrt{n}$ and the inverse of the transmission range in grid and random networks, respectively), we can actually quantify the asymptotic advantage of caching. Moreover, we compare these scaling laws for different content discovery mechanisms and illustrate that not much gain is lost when a simple path search is used.**

***Index Terms*— Next generation networking, wireless networks, performance analysis.**

I. Introduction

IN TODAY'S networking situations, users are mostly interested in accessing content regardless of which host is providing this content. They are looking for a fast and secure access to data in a whole range of situations: wired or wireless; heterogeneous technologies; in a fixed location or when moving. The dynamic characteristics of the network users makes the host-centric networking paradigm inefficient. Information-centric networking (ICN) is a new networking architecture where content is accessed based upon its name, and independently of the location of the hosts [1]–[4]. In most ICN architectures, data is allowed to be stored in the nodes and routers within the network in addition to the content publisher's servers. This reduces the burden on the servers and on the network operator, and shortens the access time to the desired content.

Manuscript received March 10, 2016; revised August 23, 2016; accepted September 27, 2016. The associate editor coordinating the review of this paper and approving it for publication was J. Widmer.

B. Azimdoost was with the Huawei Innovation Center, Santa Clara, CA 95050 USA, and also with the Department of Electrical Engineering, University of California at Santa Cruz, Santa Cruz, CA 95064 USA (e-mail: bazimdoost@soe.ucsc.edu).

C. Westphal is with the Department of Computer Engineering, University of California at Santa Cruz, Santa Cruz, CA 95064 USA, and also with the Huawei Innovation Center, Santa Clara, CA 95050 USA (e-mail: cedric.westphal@huawei.com).

H. R. Sadjadpour is with the Department of Electrical Engineering, University of California at Santa Cruz, Santa Cruz, CA 95064 USA (e-mail: hamid@soe.ucsc.edu).

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

Digital Object Identifier 10.1109/TCOMM.2016.2615624

Combining content routing with in-network-storage for the information is intuitively attractive, but there has been few works considering the impact of such architecture on the capacity of the network in a formal or analytical manner. In this work we study a wireless information-centric network where nodes can both route and cache content. We also assume that a node keeps a copy of the content only for a finite period of time, that is until it runs out of memory space in its cache and has to rotate content, or until it ceases to serve a specific content.

The nodes issue some queries for content that is not locally available. We suppose that there exists a server which permanently keeps all the contents. This means that the content is always provided at least by its publisher, in addition to the potential copies distributed throughout the network. Therefore, at least one replica of each content always exists in the network and if a node requests a piece of information, this data is provided either by its original server or by a cache containing the desired data. When the customer receives the content, it stores the content and shares it with the other nodes if needed.

The present paper thus investigates the asymptotic[1] orders of access time and throughput capacity in such content-centric networks and addresses the following questions:

1) Looking at the throughput capacity and latency, can we quantify the performance improvement brought about by a content-centric network architecture over networks with no content sharing capability?
2) How does the content discovery mechanism affect the performance? More specifically, does selecting the nearest copy of the content improve the scaling of the capacity and access time compared to selecting the nearest copy in the direction of original server?
3) How does the caching policy, and in particular, the length of time each piece of content spends in the cache's memory, affect the performance?

We state our results in three theorems; Theorem 1 formulates the capacity in a grid network which uses the shortest path to the server content discovery mechanism considering different content availability in different caches, and Theorem 2 and 3 answer the above questions studying two different network models (grid and random network) and two content discovery scenarios (shortest path to the server and shortest path to the closest copy of the content) when the

[1] Given two functions f and g, we say that $f(n) = O(g(n))$ or $f(n) \preceq g(n)$ if $sup_n(f(n)/g(n)) < \infty$, $f(n) = \Omega(g(n))$ or $f(n) \succeq g(n)$ if $g(n) = O(f(n))$, $f(n) = \Theta(g(n))$ or $f(n) \equiv g(n)$ if both $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$, $f(n) = o(g(n))$ or $f(n) \prec g(n)$ if $f(n)/g(n) \to 0$, and $f(n) = \omega(g(n))$ or $f(n) \succ g(n)$ if $g(n)/f(n) \to 0$.

information exists in all caches with the same probability. These theorems demonstrate that adding the content sharing capability to the nodes can significantly increase the capacity.

The rest of the paper is organized as follows. After a brief review of the related work in Section II, the network models, the content discovery algorithms used in the current work, and the content distribution in steady-state are introduced in Section III. The main theorems are stated and proved in Section IV. We discuss the results and study some simple examples in Section V. Finally the paper is concluded and some possible directions for the future work will be introduced in Section VI.

II. Related Work

Information Centric Networks have recently received considerable attention. While our work presents an analytical abstraction, it is based upon the principles described in some ICN architectures, such as CCN [4], NetInf [5], PURSUIT [2], or DONA [6], where nodes can cache content, and requests for content can be routed to the nearest copy. Papers surveying the landscape of ICN [3] [7] show the dearth of theoretical results underlying these architectures.

Caching, one of the main concepts in ICN networks, has been studied in prior works [3]. Reference [8] computes the performance of a Least-Recently-Used (LRU) cache taking into account the dynamical nature of the content catalog. Some performance metrics like miss ratio in the cache, or the average number of hops each request travels to locate the content have been studied in [9] and [10], and the benefit of cooperative caching has been investigated in [11].

Optimal cache locations [12], cach sizes [13], and cache replacement techniques [14] are other aspects most commonly investigated. The work in [15] considers a network of LRU caches with arbitrary topology and develops a calculus for computing bounding flows in such network. And an analytical framework for investigating properties of these networks like fairness of cache usage is proposed in [16]. Reference [17] considered information being cached for a limited amount of time at each node, as we do here, but focused on flooding mechanism to locate the content, not on the capacity of the network. Reference [18] investigates the routing in such networks in order to minimize the average access delay. Rossi and Rossini explore the impact of multi-path routing in networks with online caching [19], and also study the performance of CCN with emphasis on the size of individual caches [20].

However, to the best of our knowledge, there are just a few works focusing on the achievable data rates in such networks. Calculating the asymptotic throughput capacity of wireless networks with no cache has been solved in [21] and many subsequent works [22], [23]. Some work has studied the capacity of wireless networks with caching [24], [25], and [26]. There, caching is used to buffer data at a relay node which will physically move to deliver the content to its destination, whereas we follow the ICN assumption that caching is triggered by the node requesting the content. Reference [27] uses a network simulation model and evaluates the performance (file transfer delay) in a cache-and-forward system with no request for the data. Reference [28] proposes an analytical model for single cache miss probability and stationary throughput in cascade and binary tree topologies. Some scaling regimes for the required link capacity is computed in [29] for a static cache placement in a multihop wireless network.

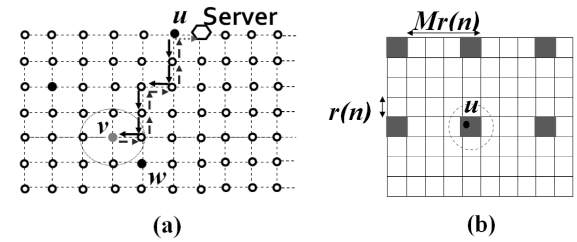

Fig. 1. a) Grid network: the transmission range of node v contains four surrounding nodes. The black vertices contain the content in their local caches. The arrow lines demonstrate a possible discovery and receive path in scenario **i**, where node v downloads the required information from u. In scenario **ii**, v will download the data from w instead. b) Random network: the grey squares are the cells that can transmit simultaneously without interference, and $r(n)$ is the transmission range of each node.

Reference [30] considers a general problem of delivering content cached in a wireless network and provides some bounds on the caching capacity region from an information-theoretic point of view, and [31] proposes a coded caching scheme to achieve the order-optimal performance. Additionally, the wireless device-to-device cache networks' performance with offline caching phase has been studied in [32]–[34]. This is important to note that our current work is different from [30]–[34] since unlike the mentioned works it considers the online caching and assumes that the cache contents are updated during the content delivery time.

A preliminary version of this paper [35] has derived the throughput capacity when all the items have exactly the same characteristics (popularity), which has been shown to be one of the important characteristics of such networks [36], [37]. In this work, we do not assume any specific popularity distribution and present the results for any arbitrary request pattern.

III. Preliminaries

A. Network Model

Two network models are studied in this work.

1) Grid Network: Assume that the network consists of n nodes $\{v_1, v_2, \dots, v_n\}$ each with a local cache of size $L_i = \Theta(1)$ located on a grid. In this work we focus on the grid shown in Figure 1(a), but conjecture the theorems could be adapted to other regular grid topologies too. Each node can transmit over a common wireless channel, with bandwidth W bits per second, shared by all nodes. The distance between two adjacent nodes equals to the transmission range of each node, so the packets sent from a node are only received by four adjacent nodes.

There are m different contents, $\{f_1, \dots, f_m\}$ with sizes $\{B_1, \dots, B_m\}$, for which each node v_j may issue a query with probabilities $\{\alpha_k,\ k = 1, \dots, m\}$, where $\sum_{k=1}^{m} \alpha_k = 1$, and m and α_k are not changing with the network size.[2] Based on the

[2] In this work we are not considering applications like YouTube where the users are content generators.

content discovery algorithms which will be explained later in this section, the query will be transmitted in the network to discover a node containing the desired content locally. v_j then downloads B_k bits of data with rate γ in a hop-by-hop manner through the path P_{xj} from either a node ($v_i, x = i$) containing it locally ($f \in v_i$) or the server ($x = s$). When the download is completed, the data is cached and shared with other nodes either by all the nodes on the delivery path, or only by the end node. In the paper we consider both options.

P_{js} denotes the nodes on the path from v_j to server. Without loss of generality, we assume that the server is attached to the node located at the middle of the network, as changing the location of the server does not affect the scaling laws. Using the protocol model and according to [38], the transport capacity in such network is upper bounded by $\Theta(W\sqrt{n})$. This is the model studied in Theorem 1 and the first two scenarios of Theorem 2.

2) Random Network: The next network studied in Theorem 2 is a more general network model where the nodes are randomly distributed over a unit square area according to a uniform distribution (Figure 1(b)). We use the same model used in [38] (section 5) and divide the network area into square cells each with side-length proportional to the transmission range $r(n)$, which is decreasing when the number of nodes increases, and is selected to be at least $\Theta\sqrt{\frac{\log n}{n}}$ to guarantee the connectivity of the network [39] and a non-zero capacity. According to the protocol model [38], if the cells are far enough they can transmit data at the same time with no interference; we assume that there are M^2 non-interfering groups which take turn to transmit at the corresponding time-slot in a round robin fashion. Again, without loss of generality the server is assumed to be located at the middle of the network. In this model the maximum number of simultaneous feasible transmissions will be in the order of $\frac{1}{r^2(n)}$ as each transmission consumes an area proportional to $r^2(n)$. All other assumptions are similar to the grid network.

B. Content Discovery Algorithm

1) Path-Wise Discovery: To discover the location of the desired content, the request is sent through the shortest path toward the server containing the requested content. If an intermediate node has the data in its local cache, it does not forward the request toward the server anymore and the requester will start downloading from the discovered cache. Otherwise, the request will go all the way toward the server and the content is obtained from the main source. In case of the random network when a node needs a piece of information, it will send a request to its neighbors toward the server, i.e. the nodes in the same cell and one adjacent cell in the path toward the server, if any copy of the data is found it will be downloaded. If not, just one node in the adjacent cell will forward the request to the next cell toward the server.

2) Expanding Ring Search: In this algorithm the request for the information is sent to all the nodes in the transmission range of the requester. If a node receiving the request contains the required data in its local cache, it notifies the requester and then downloading from the discovered cache is started.

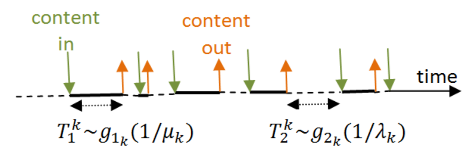

Fig. 2. Data access process time diagram in a cache for content k

Otherwise, all the nodes that receive the request will broadcast the request to their own neighbors. This process continues until the content is discovered in a cache and the downloading follows after that. This will return the nearest copy from the requester.

C. Content Distribution in Steady-State

The time diagram of data access process in a cache is illustrated in Figure 2. When a query for content f_k is initiated, the content is downloaded from a cache containing it and is received by another cache where it is going to be kept. The same cache may receive the same data after some random time (T_2^k) with distribution g_{2_k} and mean $1/\lambda_k$. Note that 1) no specific caching policy is assumed here, and 2) a node will receive the content only if it does not have it in its local cache. The solid lines in this diagram denote the portions of time that the data is available at the cache.

As the requests for different contents are assumed to be independent and holding times are set for each content independent of the others, we can do the calculations for one single content. If the total number of contents is not a function of the network size, this will not change the capacity order. Assume that content sizes B_k are much larger than the request packet size, so we ignore the overhead of the discovery phase in our calculations.

The average portion of time that each node contains a content in its local cache is

$$\rho^{(k)}(n) = \frac{1/\mu_k}{1/\mu_k + 1/\lambda_k} = \frac{\lambda_k}{\lambda_k + \mu_k}, \tag{1}$$

which is the average probability that a node contains the content k at steady-state. λ_k is the rate of requests for content k received by a cache in case of the data not being available, and μ_k is the rate of the data being expunged from the cache. Both these parameters can strongly be dependent on the total number of users, or the topology and configuration of the network or the cache characteristics like size and replacement policy.

D. Performance Indices

The performance indices studied in this work are:

1) Throughput Capacity: Throughput capacity is the maximum common content download rate which can be achieved by all users on average.

2) Average Latency: The average amount of time it takes for a customer to receive its required content from a cache or server.

3) Total Traffic: The total traffic generated by downloading item k is the number of item k bits moving across the netwrok in a second. In other words, it is the product of total request rate (the product of the number of requesting nodes and the rate at which each node is sending the request), the number of hops between source and destination, and the content size.

IV. Theorem Statements and Proofs

Consider a grid wireless network consisting of n nodes, transmitting over a common wireless channel, with shared bandwidth of $W = \Theta(1)$ bits per second. Assume that there is a server which contains all the information. Without loss of generality we assume that this server is located in the middle of the network. Each node contains some information in its local cache. Assume that according to the symmetry, the probability of each content k being in all the caches with the same distance (j hops) from the server is the same ($\rho_j^{(k)}(n)$).

Theorem 1: The maximum achievable throughput capacity order (γ_{max}) in the above network when the nodes use the nearest copy of the required content on the shortest path toward the server is given by[3]

$$\gamma_{max} \equiv \frac{n}{\sum_{k=1}^{m} \alpha_k \sum_{i=1}^{\sqrt{n}} 4i \sum_{j=0}^{i-1} (i-j)\rho_j^{(k)}(n) \prod_{l=j+1}^{i} (1-\rho_l^{(k)}(n))},$$

where $\rho_0^{(k)}(n) = 1$, which means that the server always contains all the contents.

Proof: Considering the grid topology and large number of nodes, each cache may receive requests and downloaded contents originated from different nodes. Since the users are sending requests independent of each other, the requests received by different caches can be assumed independent of each other. Thus, the information in each cache is independent of the contents in the other caches. This assumption has been made in some other works too, among which are [28] and [40]-[43] to name a few.

A request initiated by a user v_i in i-hop distance from the server (located in level $i = 1, .., \sqrt{n}$) is served by cache u_j located in level j, $1 \le j \le i$ on the shortest path from v_i to the server if no caches before u_j, including v_i, on this path contains the required information, and u_j contains it. This request is served by the server if no copy of it is available on the path. Let $P_{i,j}^{(k)}$ denote the probability of v_i's request for item k being served by u_j, this probability is given by $P_{i,j}^{(k)} =$

$$(1-\rho_i^{(k)}(n))(1-\rho_{i-1}^{(k)}(n))...(1-\rho_{j+1}^{(k)}(n))\rho_j^{(k)}(n) \quad (2)$$

where $\rho_j^{(k)}(n)$ is the probability of content k being available in a cache in level $j, 1 \le j \le \sqrt{n}$, and $j = 0$ shows the server and $\rho_0^{(k)}(n) = 1$. Thus a content k requested by v_i is traveling $i - j$ hops with probability $P_{i,j}^{(k)}$. There are $4i$ nodes in level i so the average number of hops ($E[h_k]$) traveled by item k from the serving cache (or the original server) to the requester is

$$E[h_k] = \frac{1}{n} \sum_{i=1}^{\sqrt{n}} 4i \sum_{j=0}^{i-1} (i-j) P_{i,j}^{(k)} \quad (3)$$

Therefore the average number of hops in the network is given by $E[h] = \sum_{k=1}^{m} \alpha_k E[h_k]$.

Assume that each user is receiving data with rate γ. The transport capacity in this network, which equals to $n\gamma E[h]$, is upper bounded by $\Theta(\sqrt{n})$ bits-meters/sec divided by the distance of each hop $\Theta(\frac{1}{\sqrt{n}})$, which is $\Theta(n)$ bits-hops/sec. So $\gamma_{max} = \Theta(\frac{1}{E[h]})$ and the Theorem is proved. ■

Now consider a wireless network consisting of n nodes, with each node containing information k in its local cache with common probability,[4,5] $\rho^{(k)}(n) \nrightarrow 1$ (meaning that it does not tend to 1 when n increases.), otherwise for $\rho^{(k)}(n) \to 1$, the request is served locally and no data is transferred between the nodes. Assume that the request process and cache look up time in each node is not a function of the number of nodes. Here, based on the network models and content discovery methods, we define the following different scenarios, and then study the corresponding performance of caching in Theorems 2 and 3;

- Scenario **i**- The nodes are located on a grid and search for the contents just on the shortest path toward the server,
- Scenario **ii**- The nodes are located on a grid and use ring expansion to find contents,
- Scenario **iii**- The nodes are randomly distributed over a unit square area and use path-wise content discovery algorithm. Each node has a transmission range of $r(n)$ which at least equals to $\Theta(\sqrt{\frac{\log n}{n}})$ so the network is connected.

Theorem 2: The average latency order in the three scenarios defined above is

- *Scenario* **i**- $\Theta(min(\sqrt{n}, \frac{1}{\min_k(\rho^{(k)}(n))}))$.
- *Scenario* **ii**- $\Theta(min(\sqrt{n}, \frac{1}{\sqrt{\min_k(\rho^{(k)}(n))}}))$.
- *Scenario* **iii**- $\Theta(max[1, min(\frac{1}{r(n)}, \frac{1}{\min_k(\rho^{(k)}(n))nr^2(n)})])$.

Here we prove Theorem 2 by utilizing some Lemmas. The proof of lemmas are presented in the Appendix.

Lemma 1: Consider the wireless networks described in Theorem 2. The average number of hops between the customer and the serving node (a cache or original server) for item k is

- Scenario **i**- $E[h_k]$ asymptotically equals to

$$\frac{1}{n}\sum_{i=1}^{\sqrt{n}} i^2(1-\rho^{(k)}(n))^i + \frac{\rho^{(k)}(n)}{n}\sum_{i=1}^{\sqrt{n}} i \sum_{l=1}^{i-1} l(1-\rho^{(k)}(n))^l \quad (4)$$

- Scenario **ii**- $E[h_k]$ asymptotically equals to

$$\frac{1}{n}\{\sum_{i=1}^{\sqrt{n}} i^2(1-\rho^{(k)}(n))^{2i^2-2i+1} + \sum_{i=2}^{\sqrt{n}} i \sum_{l=1}^{i-1} l(1-\rho^{(k)}(n))^{2l^2-2l+1}(1-(1-\rho^{(k)}(n))^{4l})\} \quad (5)$$

[3]Since no online caching assumption is used in this Theorem, it can be used for offline caching networks as well. However, we skip the offline results and target the networks with online caching which is the scope of this paper.

[4]The proof does not need the probabilities to be exactly the same, they just need to be of the same order in terms of n.

[5]Note that this assumption is correct for networks with online caching. In offline caching scenarios each content is present in some specific caches. However, offline caching can be considered as a special case of online caching, and we still can use this theorem by assigning the value of the fraction of caches containing the item to the probability of each item being in a cache.

- Scenario **iii**- $E[h_k]$ asymptotically equals to

$$r^2(n)\{\sum_{i=2}^{\frac{1}{r(n)}} i^2(1-\rho^{(k)}(n))^{inr^2(n)} + (1-(1-\rho^{(k)}(n))^{nr^2(n)})$$
$$\sum_{i=2}^{\frac{1}{r(n)}} i \sum_{l=1}^{i-1} l(1-\rho^{(k)}(n))^{lnr^2(n)}\} \qquad (6)$$

Lemma 2: Consider the wireless networks described in Theorem 2. For sufficiently large networks, the average number of hops between the customer and the serving node (a cache or the original server) for item k is

- Scenario **i**- $E[h_k]$ equals $\sqrt{n}$ for $\rho^{(k)}(n) \preceq \frac{1}{\sqrt{n}}$, and $\frac{1}{\rho^{(k)}(n)}$ for $\rho^{(k)}(n) \succeq \frac{1}{\sqrt{n}}$.
- Scenario **ii**- $E[h_k]$ equals $\sqrt{n}$ for $\rho^{(k)}(n) \preceq \frac{1}{n}$, and $\frac{1}{\sqrt{\rho^{(k)}(n)}}$ for $\rho^{(k)}(n) \succeq \frac{1}{n}$.
- Scenario **iii**- $E[h_k]$ equals $\frac{1}{r(n)}$ for $\rho^{(k)}(n) \preceq \frac{1}{nr(n)}$, $\frac{1}{\rho^{(k)}(n)nr^2(n)}$ for $\frac{1}{nr(n)} \preceq \rho^{(k)}(n) \preceq \frac{1}{nr^2(n)}$, and 1 for $\rho^{(k)}(n) \succeq \frac{1}{nr^2(n)}$.

Theorem 2 is now simply proved using the above Lemmas.

Proof: The average number of hops each content is traveling is $E[h] = \sum_{k=1}^{m} \alpha_k E[h_k]$.

We assume that the number of contents and also the popularity of each item is not changing with the network size (number of users). In the three scenarios mentioned above for the cases of $\rho^{(k)}(n) \preceq \frac{1}{\sqrt{n}}$, $\rho^{(k)}(n) \preceq \frac{1}{n}$, and $\rho^{(k)}(n) \preceq \frac{1}{nr(n)}$, when there is at least one node with average number of hops equal to $\sqrt{n}$, $\sqrt{n}$, and $\frac{1}{r(n)}$ respectively, then that node's $E[h_k]$ in $E[h]$ defined above becomes the dominant factor.

If $\rho^{(k)}(n) \succeq \frac{1}{\sqrt{n}}$, $\rho^{(k)}(n) \succeq \frac{1}{n}$, and $\rho^{(k)}(n) \succeq \frac{1}{nr^2(n)}$ for all the contents, in the three scenarios, respectively, then $E[h]$ in the three scenarios is given by $\sum_{k=1}^{m} \frac{\alpha_k}{\rho^{(k)}(n)} \equiv \frac{1}{\min_k(\rho^{(k)}(n))}$, $\sum_{k=1}^{m} \frac{\alpha_k}{\sqrt{\rho^{(k)}(n)}} \equiv \frac{1}{\sqrt{\min_k(\rho^{(k)}(n))}}$, and $\sum_{k=1}^{m} \alpha_k = 1$.

In the third scenario, if there is no item for which $\rho^{(k)}(n) \preceq \frac{1}{nr(n)}$, but there is at least one item such that $\rho^{(k)}(n) \preceq \frac{1}{nr^2(n)}$, then $E[h] = \sum_{k=1}^{m} \frac{\alpha_k}{\rho^{(k)}(n)nr^2(n)} \equiv \frac{1}{\min_k(\rho^{(k)}(n)nr^2(n))}$.

The total $E[h]$ can be simply written as the results shown in *Theorem 2*.

Assuming that the delay of the request process and cache look up in each node is not increasing when the network size (the number of nodes) increases, and that there is enough bandwidth to avoid congestion, then the latency of getting the data is directly proportional to the average number of hops between the serving node and the customer. Thus, the latency and the average number of hops the data is traveling to reach the customer are of the same order and *Theorem 2* is proved. ■

Theorem 3: Consider the networks of Theorem 2, and assume each node can transmit over a common wireless channel, with $W = \Theta(1)$ bits per second bandwidth, shared by all nodes. The maximum achievable throughput capacity order γ_{max} in the three discussed scenarios are

- *Scenario* **i**- $\Theta(max(\frac{1}{n}, \min_k((\rho^{(k)}(n))^2)))$.
- *Scenario* **ii**- $\Theta(max(\frac{1}{n}, \min_k(\rho^{(k)}(n))))$.
- *Scenario* **iii**- $\Theta(max[\frac{1}{n}, min(\frac{1}{nr^2(n)}, \min_k((\rho^{(k)}(n))^2)nr^2(n))])$.

To prove Theorem 3 we use Lemma 2, and the following two Lemmas.

Lemma 3: Consider the wireless networks described in Theorem 2. In order not to have interference, the maximum throughput capacity is upper limited by

- Scenario **i**- $\Theta(max(\frac{1}{\sqrt{n}}, \min_k(\rho^{(k)}(n))))$.
- Scenario **ii**- $\Theta(max(\frac{1}{\sqrt{n}}, \sqrt{\min_k(\rho^{(k)}(n))}))$.
- Scenario **iii**- $\Theta(min[\frac{1}{nr^2(n)}, max(\frac{1}{nr(n)}, \min_k(\rho^{(k)}(n)))])$.

In the previous Lemma, the maximum throughput capacity in a wireless network utilizing caches has been calculated such that no interference occurs. Now it is important to verify if this throughput can be supported by each node (cell), i.e. the traffic carried by each node (cell) is not more than what it can support ($\Theta(1)$).

Lemma 4: The maximum supportable throughput capacities in the studied scenarios are as follows.

- Scenario **i**- $\Theta(max(\frac{1}{n}, \min_k((\rho^{(k)}(n))^2)))$.
- Scenario **ii**- $\Theta(max(\frac{1}{n}, \min_k(\rho^{(k)}(n))))$.
- Scenario **iii**- $\Theta(max[\frac{1}{n}, min(\frac{1}{nr^2(n)}, \min_k((\rho^{(k)}(n))^2)nr^2(n))])$.

The maximum throughput capacity is the value which can be supported by all the nodes while no interference occurs. Thus the throughput capacity will be the minimum of the two values derived in Lemmas 3 and 4, and Theorem 3 is proved.

V. Discussion

The Theorems above express the maximum achievable data download rate in terms of the availability of the contents in the caches($\rho^{(k)}(n)$), in networks with specific topology and content discovery mechanisms. However, no assumption on the caching policy, which is an important factor in determining $\rho^{(k)}(n)$ have been made. In this section, we discuss our results based on two examples and try to study the affect of caching policy on the performance.

In these examples we consider two different cache replacement policies based on Time-To-Live (TTL). First example uses exponentially distributed TTL, and the second one considers constant TTL. According to [44] this can predict metrics of interest on networks of caches running other replacement algorithms like LRU, FIFO, or Random.

In order to use the stated theorems, the probability of each item being in each cache is first calculated, and then, the appropriate theorem is used to give the throughput capacity. In the first example, in addition to the capacity, we analyze the total request rate ($n(1-\rho^{(k)})\lambda_k$) and total generated traffic for an item k ($n(1-\rho^{(k)})\lambda_k B_k E[h_k]$) as well. This gives us an idea about how the request rates and cache holding times affect the traffic in the network and how the resources are utilized.

A. Example 1

1) Network Model: Consider a network where the received data is stored only at the receivers (edge caching [45], [46])

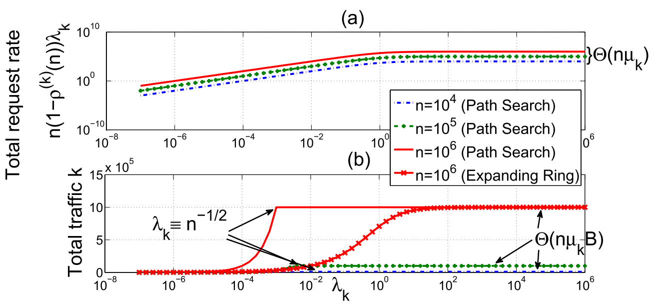

Fig. 3. Scenario **i** (a) Total request rate for an item k in the network ($\lambda_k n(1-\rho^{(k)}(n))$), (b) Total traffic in the network ($B_k \lambda_k n(1-\rho^{(k)}(n))E[h_k]$) vs. the request rate ($\lambda_k$) with fixed time-out rate ($\mu_k = 1$).

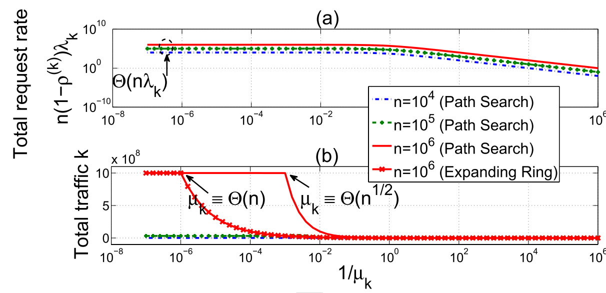

Fig. 4. Scenario **i** (a) Total request rate in the network ($\lambda_k n(1-\rho^{(k)}(n))$), (b) Total traffic in the network ($B_k \lambda_k n(1-\rho^{(k)}(n))E[h_k]$) vs. the inverse of the time-out rate ($1/\mu_k$) with fixed request ratio ($\lambda_k = 1$).

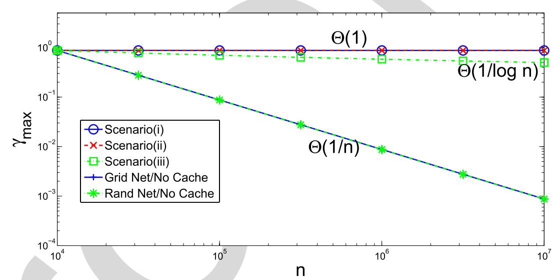

Fig. 5. Maximum download rate (γ_{max}) vs. the number of nodes (n) for $\rho = 7/8$.

and then shared with the other nodes as long as the node keeps the content. Assume that receiving a data k in the local cache of the requesting user sets a time-out timer with exponentially distributed duration with parameter η_k and no other event will change the timer until it times-out, meaning that in equation (1) $\mu_k = \eta_k$, which is not a function of n. Considering the request process for each content k from each user being a Poisson process with rate β_k not changing with n, and using the memoryless property of exponential distribution (internal request inter-arrival times), and assuming that the received data is stored only in the end user's cache (the caches on the download path do not store the downloading data), it can be proved that in equation (1) $\lambda_k = \beta_k$. Thus we can write the presence probability of each content k in each cache as $\rho^{(k)}(n) = \frac{\beta_k}{\beta_k+\eta_k}$ (equal order probability of all the caches containing an item k).

2) Results: Figures 3 (a),(b) respectively illustrate the total request rate and the total traffic generated in a fixed size network in scenario **i** for each item k for different request rates when the time-out rate is fixed. Small λ_k means that each node is sending requests for k with low rate, so fewer caches have that content, and consequently more nodes are sending requests with this low rate. In this case most of the requests are served by the server. The total request rate of item k will increase by increasing the per node request rate. High λ_k shows that each node is requesting the content with higher rate, so the number of cached content k in the network is high, thus fewer nodes are requesting it with this high rate externally. Here most of the requests are served by the caches. The total request rate then is determined by the content drop rate. So for very large λ_k, the total request rate is the total number of nodes in the network times the drop rate ($n\mu_k$) and the total traffic is $n\mu_k B_k$. As can be seen there is some request rate at which the traffic reaches its maximum; this happens when there is a balance between the requests served by the server and by the caches. For smaller request rates, most of the requests are served by the server and increasing λ_k increases the total traffic. For larger λ_k, on the other hand, most of the requests are served by the caches and increasing the request rate will not change the distance to the nearest content and the total traffic.

Figures 4 (a),(b) respectively illustrate the total request rate and the total traffic generated in a fixed size network in scenario **i** for different time-out rates when the request rate is fixed. For low $1/\mu_k$ (high time-out rates or small lifetimes), most of the item k requests are served by the server and caching is not used at all. For large time-out times, all the requests are served by the caches, and the only parameter in determining the total request rate is the time-out rate.

However, when the network grows the traffic in the network will increase and the download rate will decrease. If we assume that the new requests are not issued in the middle of the previous download, the request rate will decrease with network growth. If the holding time of the contents in a cache increases accordingly the total traffic will not change, i.e. if by increasing the network size the requests are issued not as fast as before, and the contents are kept in the caches for longer times, the network will perform similarly.

In Figure 5 we assume that the request rate is roughly 7 times the drop rate for all the contents, so $\rho^{(k)}(n) = 7/8$, and show the maximum throughput order as a function of the network size. In scenario **iii**, we set the transmission range to the minimum value needed to have a connected network ($r(n) \equiv \sqrt{\frac{\log n}{n}}$). According to Theorem 3 and as can be observed from this figure, the maximum throughput capacity of the network in a grid network with the described characteristics is not changing with the network size if the probability of each item being in each cache is fixed, while in a network with no cache this capacity will be inversely proportional to the network size. Similarly in the random network the maximum throughput is inversely proportional to $nr^2(n)$, which is the logarithm of the network size, while in a no cache network it is diminishing with the rate of network growth.

Moreover, comparing scenario **i** with **ii**, we observe that the throughput capacity in both cases are almost the same; meaning that using the path discovery scheme will lead to almost the same throughput capacity as the expanding ring discovery. Thus, we can conclude that just by knowing the address of a server containing the required data and forwarding the requests through the shortest path toward that server we can achieve the best performance, and increasing the complexity and control traffic to discover the closest copy of the required content does not add much to the capacity.

On the other hand with a fixed network size, if the probability of an item being in each cache is greater than a threshold ($\Theta(\frac{1}{\sqrt{n}})$, $\Theta(\frac{1}{n})$, and $\Theta(\frac{1}{nr^2(n)}) = \Theta(\frac{1}{\log n})$ in cases **i**, **ii** and **iii**, respectively), most of the requests will be served by the caches and not the server, so increasing the probability of an intermediate cache having the content reduces the number of hops needed to forward the content to the customer, and consequently increases the throughput. For content presence probability orders less than these thresholds ($\Theta(\frac{1}{nr(n)}) = \Theta(\frac{1}{\sqrt{n\log n}})$ in cases **iii**) most of the requests are served by the main server, so the maximum possible number of hops will be traveled by each content to reach the requester and the minimum throughput capacity ($\Theta(\frac{1}{n})$) will be achieved. Note that in these networks, the maximum throughput is limited by the maximum supportable load on each link, and more specifically on the server.

As may have been expected and according to our results, the obtained throughput is a function of the probability of each content being available in each cache, which in turn is strongly dependent on the network configuration and cache management policy.

B. Example 2

1) Network Model: Assume an n-cache grid wireless network with one server containing all the items located in the middle of the network. Each cache in level i (nodes at i hops away from the server) receives requests for a specific document k according to a Poisson distribution with rate $\beta^{(k)}$ from the local user, and with rate $\beta_i^{(k)'}(n)$ from all the other nodes. Note that rate $\beta_i^{(k)'}(n)$ is a function of the individual request rate of users for item k ($\beta^{(k)}$) and also the location of the cache in the network. The content discovery mechanism is path-wise discovery, and whenever a copy of the required data is found (in a cache or server), it will be downloaded through the reverse path, and either all the nodes on the download path or only the requester node store it in their local caches. Moreover, we assume that receiving the item k and also any request for the available cached data k by a node in level i refreshes a time-out timer with fixed duration $D_i^{(k)}(n)$. According to [47], this is a good approximation for caches with LRU replacement policy when the cache size and the total number of documents are reasonably large. Furthermore, according to the same work this value is a constant for all contents and is a function of the cache size, so we can use $D_i(n)$ for all contents in caches in level i. We will calculate the average probability of item k being in a cache in level i ($\rho_i^{(k)}(n)$) based on these assumptions and then use Theorem 1 to obtain the throughput capacity.

2) Results: Let random variable $t_{on}^{(k)i}(T)$ denote the total time of the data k being available in a cache in level i (i hop distance from the server) during constant time T. Assume that item k is received $N^{(k)i}(T)$ times during time T by each node v_i in level i (according to the symmetry all nodes in one level have similar conditions.). The data available time between any two successive receipt of item k is $D_i(n)$ if the timer set by the first receipt is expired before the second one comes, or is equal to the time between these two receipts. Let $\tau_i^{req(k)}$ denote the time between two successive receipts. This process has an exponential distribution with parameter $\beta_i^{(k)} = \beta^{(k)} + \beta_i^{(k)'}$. So the total time of data k availability in a level i cache is

$$t_{on}^{(k)i}(T) = \sum_{j=0}^{N^{(k)i}(T)} min(\tau_i^{req(k)}, D_i(n)), \tag{7}$$

and the average value of this time is ($E[t_{on}^{(k)i}(T)]$)

$$\sum_{l=0}^{\infty} E[\sum_{j=0}^{l} min(\tau_i^{req(k)}, D_i(n))]Pr(N^{(k)i}(T) = l),$$

$$= \sum_{l=0}^{\infty} lE[min(\tau_i^{req(k)}, D_i(n))]Pr(N^{(k)i}(T) = l),$$

$$= E[min(\tau_i^{req(k)}, D_i(n))]E[N^{(k)i}(T)]. \tag{8}$$

According to the Poisson arrivals of requests (data downloads) with parameter $\beta^{(k)} + \beta_i^{(k)'}$, the rightmost term in equation (8) ($E[N^{(k)i}(T)]$) equals $(\beta^{(k)}+\beta_i^{(k)'})T$. The leftmost term in this equation ($E[min(\tau_i^{req(k)}, D_i(n))]$) can also be easily calculated and equals to $\frac{1-e^{-D_i(n)(\beta^{(k)}+\beta_i^{(k)'})}}{\beta^{(k)}+\beta_i^{(k)'}}$. Therefore, $E[t_{on}^{(k)i}(T)] = (1 - e^{-D_i(n)(\beta^{(k)}+\beta_i^{(k)'})})T$. And finally the probability of an item k being available in a level i cache is $\rho_i^{(k)} = \frac{E[t_{on}^{(k)i}(T)]}{T} = 1-e^{-D_i(n)(\beta^{(k)}+\beta_i^{(k)'}(n))}$. Note that $D_0 = \infty$ so that $\rho_0^{(k)} = 1$.

Now we need to calculate the rate of item k received by each node in level i. First, assume that when an item is downloaded , only the end user (the node which has requested the content) keeps the downloaded content, and storing a new content refreshes the time-out timer with fixed duration $D_i(n)$. Thus $\beta_i^{(k)'}(n) = 0$, and $\rho_i^{(k)}(n) = 1 - e^{-D_i(n)\beta^{(k)}}$. It is obvious that in such network where all the caches have the same size and the request patterns, $D_i(n)$ will not depend on the cache location, and since the request rate and the caches sizes are not changing with n this value does not depend on the network size either. Thus, $D_i(n)$ can be replaced by fixed and constant D. Therefore, $\rho_i^{(k)}(n) = 1 - e^{-D\beta^{(k)}}$ which is similar for all the caches, and the maximum throughput capacity order (γ_{max}) is $\frac{n}{\sum_{k=1}^{m}\alpha_k \sum_{i=1}^{\sqrt{n}} i \sum_{j=0}^{i-1}(i-j)(1-e^{-D\beta^{(k)}})e^{-(i-j)D\beta^{(k)}}}$, which is

$$\frac{1}{\sum_{k=1}^{m} \frac{\alpha_k e^{-D\beta^{(k)}}}{1-e^{-D\beta^{(k)}}}} \equiv 1. \tag{9}$$

As the second case, we assume that all the nodes on the download path keep the data, and the shortest path from the requester to the server is selected such that all the nodes in level i receive the requests for item k with the same rate. There are $4i$ nodes in level i and $4(i+1)$ nodes in level $i+1$. So the request initiated or forwarded from a node in level $i+1$ will be received by a specific node in level i with probability $\frac{i}{i+1}$ if it is not locally available in that node, so $\beta_i^{(k)'}(n)$ can be expressed as

$$\beta_i^{(k)'} = \frac{(1-\rho_{i+1}^{(k)})(\beta^{(k)}+\beta_{i+1}^{(k)'})(i+1)}{i} \tag{10}$$

Combining equation (10), the relationship between $\rho_i^{(k)}$ and $\beta_i^{(k)'}$, and the fact that there is no external request coming to the nodes at the edge boundary of the network ($\beta_{\sqrt{n}}^{(k)'} = 0$), together with the result of Theorem 1 we can obtain the capacity (γ_{max}) in the grid network with path-wise content discovery and on-path storing scheme which is n divided by $\sum_{k=1}^{m} \alpha_k \sum_{i=1}^{\sqrt{n}} i \sum_{j=0}^{i-1} (i-j)(1-e^{-D_j(n)(\beta^{(k)}+\beta_j^{(k)'})}) e^{-\sum_{l=j+1}^{i} D_l(n)(\beta^{(k)}+\beta_l^{(k)'})}$.

The result of this equation cannot exceed $\Theta(1)$ since this is the maximum possible throughput order in the grid network. Thus, caching the downloaded data in all the caches on the download path does not add any asymptotic benefit in the capacity of the network, and keeping the downloaded items only in the requester caches will yield the maximum possible throughput.

VI. Conclusion And Future Work

We studied the asymptotic throughput capacity and latency of ICNs with limited lifetime cached data at each node. The grid and random networks are two network models we investigated in this work. Representing all the results in terms of the probability of the items being in the caches while not considering any specific content popularity distribution, or cache replacement policy has empowered us to have a generalized result which can be used in different scenarios. Our results show that with fixed content presence probability in each cache, the network can have the maximum throughput order of 1 and $\frac{1}{nr^2(n)}$ in cases of grid and random networks, respectively, and the number of hops traveled by each data to reach the customer (or latency of obtaining data), can be as small as one hop.

Moreover, we studied the impact of the content discovery mechanism on the performance in grid networks. It can be observed that looking for the closest cache containing the content will not have much asymptotic advantage over the simple path-wise discovery when $\min_k \rho^{(k)}(n)$ is sufficiently small ($\min_k \rho^{(k)}(n) \preceq \frac{1}{n}$) or big enough ($\min_k \rho^{(k)}(n) \nrightarrow 0$). For other values of $\min_k \rho^{(k)}(n)$, looking for the nearest copy at most decreases the throughput diminishing rate by a factor of two. Consequently, downloading the nearest available copy on the path toward the server has similar performance as downloading from the nearest copy. A practical consequence of this result is that routing may not need to be updated with knowledge of local copies, just getting to the source and finding the content opportunistically will yield the same benefit.

Another interesting finding is that whether all the caches on the download path keep the data or just the end user does it, the maximum throughput capacity scale does not change.

In this work, we represented the fundamental limits of caching in the studied networks, proposing a caching and downloading scheme that can improve the capacity order is part of our future work.

Appendix

Proof of Lemma 1: Let h_k, d_{sr}, and d_{max} denote the number of hops between the customer and the serving node (cache or original server) for content k, the number of hops between the customer and the serving node (cache or original server), and the maximum value of d_{sr}, respectively. The average number of hops between the customer and the serving node ($E[h_k]$) is given by

$$E[h_k] = \sum_{i=1}^{d_{max}} E[h_k | d_{sr} = i] Pr(d_{sr} = i). \tag{11}$$

Scenario **i**- This case can be considered as a special case of the network studied in Theorem 1, where $\rho_i^{(k)}(n)$ is the same for all i.[6] Thus, we can drop the index i and let $\rho^{(k)}(n)$ denote the common value of this probability. Using equation (2) and (3) we will have $E[h_k]$ equal to

$$\frac{4}{n} \sum_{i=1}^{\sqrt{n}} i \{ i(1-\rho^{(k)}(n))^i + \sum_{j=1}^{i-1} (i-j)(1-\rho^{(k)}(n))^{i-j} \rho^{(k)}(n) \} \tag{12}$$

The constant factor 4 does not change the scaling order and it can be dropped. By defining $l = i - j$, then the proof follows.

Scenario **ii** - d_{max} in this network is $\Theta(\sqrt{n})$, and there are $4i$ nodes at distance of i hops from the original server. Thus, $Pr(d_{sr} = i) \equiv \frac{i}{n}$. Each customer may have the required item k in its local cache with probability $\rho^{(k)}(n)$. If the requester is one hop away from the original server, it gets the required item from the server with probability $1-\rho^{(k)}(n)$. The customers at two hops distance from the server (8 such customers) download the required item from the original server (traveling $h_k = 2$ hops) if no cache in a diamond of two hops diagonals contains it (with probability $(1-\rho^{(k)}(n))^2$), and gets it from a cache at distance one hop if one of those caches has the item (with probability $(1-\rho^{(k)}(n))(1-(1-\rho^{(k)}(n))^4)$). Using similar reasoning, the customers at distance i from the server get the item from the server (distance $h_k = i$ hops) with probability $(1-\rho^{(k)}(n))^{1+4(1+2+\cdots+(i-1))} = (1-\rho^{(k)}(n))^{2i^2-2i+1}$, and from a cache at distance $h_k = l < i$ with probability $(1-\rho^{(k)}(n))^{2l^2-2l+1}(1-(1-\rho^{(k)}(n))^{4l})$ as there are $4l$ nodes at distance of l hops. Therefore, using

[6] We will give examples in Section V using this assumption.

equations (11), (2), and (3)

$$E[h_k] \equiv \frac{1}{n}\sum_{i=2}^{\sqrt{n}} i \sum_{l=1}^{i-1} l(1-(1-\rho^{(k)}(n))^{4l})(1-\rho^{(k)}(n))^{2l^2-2l+1} + \frac{1}{n}\sum_{i=1}^{\sqrt{n}} i^2(1-\rho^{(k)}(n))^{2i^2-2i+1} \quad (13)$$

Scenario **iii** - The number of caches within transmission range (one hop) is $\Theta(nr^2(n))$. d_{max} in this network is of the order of $\frac{1}{r(n)}$ and $\Pr(d_{sr}=i) \equiv ir^2(n)$.

Each customer may have the required item k in its local cache with probability $\rho^{(k)}(n)$. If the requester is one hop away from the original server ($4\Theta(nr^2(n))$ nodes), it receives the required item from the server with probability $1-\rho^{(k)}(n)$. The customers at two hops distance from the server ($8\Theta(nr^2(n))$ such customers) download the required item from the original server (traveling $h_k = 2$ hops) if no cache in the cell at one hop distance contains it (probability $(1-\rho^{(k)}(n))^{2nr^2(n)}$), and gets it from a cache at distance one hop if one of those caches has the item (probability $(1-\rho^{(k)}(n))(1-(1-\rho^{(k)}(n))^{2nr^2(n)})$). Using similar reasoning the customers at distance i from the server receive the item from the server (distance $h_k = i$ hops) with probability $(1-\rho^{(k)}(n))^{inr^2(n)}$, and from a cache at distance $h_k = l < i$ with probability $(1-\rho^{(k)}(n))^{lnr^2(n)}(1-(1-\rho^{(k)}(n))^{nr^2(n)})$. Therefore, according to equation (11) $E[h_k]$ equals to

$$r^2(n)\{(1-\rho^{(k)}(n)) + \sum_{i=2}^{\frac{1}{r(n)}} i^2(1-\rho^{(k)}(n))^{inr^2(n)} + (1-(1-\rho^{(k)}(n))^{nr^2(n)})\sum_{i=2}^{\frac{1}{r(n)}} i\sum_{l=1}^{i-1} l(1-\rho^{(k)}(n))^{lnr^2(n)}\}. \quad (14)$$

Noting that $r^2(n)(1-\rho^{(k)}(n))$ is always less than one, and tends to zero for sufficiently large n, the Lemma is proved. ■

Proof of Lemma 2: To simplify the notations, we have dropped the index k when there is no ambiguity.

To prove this Lemma we use (A): $\lim_{N\to\infty}(1-x)^N \approx e^{-xN}$ approximation, which is ≈ 1 for $x = o(\frac{1}{N})$ (region 1), $\approx e^{-1}$ for $x = \Theta(\frac{1}{N})$ (region 2), and ≈ 0 for $x = \omega(\frac{1}{N})$ (region 3).

Scenario **i** - Let us define

$$E_s^i = \frac{1}{n}\sum_{i=1}^{\sqrt{n}} i^2(1-\rho(n))^i, \quad E_c^i = \frac{\rho(n)}{n}\sum_{i=1}^{\sqrt{n}} i\sum_{l=1}^{i-1} l(1-\rho(n))^l. \quad (15)$$

Thus equation (4) is written as $E[h] = E_s^i + E_c^i$. First we investigate the value of E_s^i for different ranges of $\rho(n)$. The summation for E_s^i can be decomposed into two summations.

$$E_s^i \equiv \frac{1}{n}\{\sum_{i\prec\sqrt{n}} i^2(1-\rho(n))^i + \sum_{i\equiv\sqrt{n}} i^2(1-\rho(n))^i\} \quad (16)$$

Assume $\rho(n) \equiv \frac{1}{\sqrt{n}}$, then using first and second region of equation (VI) we have

$$E_s^i \equiv \frac{1}{n}\{\sum_{i\prec\sqrt{n}} i^2 + \sum_{i\equiv\sqrt{n}} i^2\} \equiv \frac{n^{3/2}}{n} \equiv \sqrt{n}. \quad (17)$$

Moreover it can easily be seen that E_s^i is a decreasing function of $\rho(n)$, so for $\rho(n)$ with order less than $\frac{1}{\sqrt{n}}$ it is more than $\sqrt{n}$. Since $d_{max} = \sqrt{n}$, we can say $E_s^i \equiv \sqrt{n}$ for $\rho(n) \preceq \frac{1}{\sqrt{n}}$. Now we expand the summation to obtain

$$E_s^i \equiv \frac{(1-\rho(n))(2-\rho(n))}{n\rho^3(n)} - \frac{(1-\rho(n))^{\sqrt{n}+1}}{n\rho^3(n)} \times\{n(1-\rho(n))^2 - (1-\rho(n))(2n+2\sqrt{n}-1) + (\sqrt{n}+1)^2\} \quad (18)$$

If $\rho(n) \succ \frac{1}{\sqrt{n}}$, then using third region in equation (VI), $(1-\rho(n))^{\sqrt{n}+1}$ is going to zero exponentially, so $n(1-\rho(n))^{\sqrt{n}+1} \to 0$. Thus, $E_s^i \equiv \frac{1}{n\rho^3(n)}$, and in summary

$$E_s^i \equiv \begin{cases} \sqrt{n} & \rho(n) \preceq \frac{1}{\sqrt{n}} \\ \frac{1}{n\rho^3(n)} & \rho(n) \succ \frac{1}{\sqrt{n}} \end{cases} \quad (19)$$

According to equation (19) and since $E[h] = E_s^i + E_c^i$, when $E_s^i \equiv \sqrt{n}$ (for $\rho(n) \preceq \frac{1}{\sqrt{n}}$) which is the maximum possible order for $E[h]$, then adding E_s^i to $E[h]$ cannot increase its order beyond the maximum possible value. Now to derive the order of $E[h]$ for other values of $\rho(n)$, we decompose the equation of $E_c^i = E_c^{i1} + E_c^{i2}$ to the following summations and investigate their behaviors when $\rho(n) \succ \frac{1}{\sqrt{n}}$.

$$E_c^{i1} = \frac{1}{n}\sum_{i\equiv\sqrt{n}} i\sum_{l=1}^{i-1} l\rho(n)(1-\rho(n))^l,$$

$$E_c^{i2} = \frac{1}{n}\sum_{i\prec\sqrt{n}} i\sum_{l=1}^{i-1} l\rho(n)(1-\rho(n))^l \quad (20)$$

The number of $i \equiv \sqrt{n}$ is in the order of $\Theta(1)$. Therefore using the following series $\sum_{x=1}^{n} xa^x = \frac{a^{n+1}(na-n-1)+a}{(a-1)^2}$, we have $E_c^{i1} \equiv \frac{1}{\sqrt{n}}\sum_{l=1}^{\sqrt{n}} l\rho(n)(1-\rho(n))^l \equiv \frac{1-\rho(n)}{\rho(n)\sqrt{n}}(1-(1-\rho(n))^{\sqrt{n}}(1+\rho(n)\sqrt{n}))$, which is equivalent to $\frac{1}{\rho(n)\sqrt{n}}$ when $\rho(n) \succ \frac{1}{\sqrt{n}}$.

Utilizing the same series, the first summation in E_c^{i2} is $\Theta(\sqrt{n})$. Hence we arrive at

$$E_c^{i2} \equiv \frac{1-\rho(n)}{\rho(n)n}\sum_{i\prec\sqrt{n}} i[1-\{1-\rho(n)+\rho(n)i\}(1-\rho(n))^{i-1}]$$

$$\equiv \frac{1-\rho(n)\{1-\frac{1}{n}\sum_{i\prec\sqrt{n}} i(1-\rho(n))^i - \frac{1}{n}\sum_{i\prec\sqrt{n}} i^2\rho(n)(1-\rho(n))^{i-1}\}}{\rho(n)}$$

$$\equiv \frac{1-\rho(n)}{\rho(n)} - \frac{(1-\rho(n))^2}{\rho^3(n)n} - \frac{1}{\rho^3(n)n} \equiv \frac{1}{\rho(n)} \quad (21)$$

Since $\rho(n) \succ \frac{1}{\sqrt{n}}$, E_c^{i2} is the dominant factor in E_c^i, and also it is dominant factor in $E[h]$. Thus, $E[h] \equiv E_s^i \equiv \sqrt{n}$ for $\rho(n) \preceq \frac{1}{\sqrt{n}}$, and $E[h] \equiv E_c^{i2} \equiv \frac{1}{\sqrt{\rho(n)}}$ for $\rho(n) \succ \frac{1}{\sqrt{n}}$.

Scenario **ii** - Let us define

$$E_s^{ii} = \frac{1}{n}\sum_{i=1}^{\sqrt{n}} i^2(1-\rho(n))^{2i^2-2i+1},$$
$$E_c^{ii} = \frac{1}{n}\sum_{i=2}^{\sqrt{n}} i \sum_{k=1}^{i-1} l(1-\rho(n))^{2l^2-2l+1}(1-(1-\rho(n))^{4l}) \tag{22}$$

So $E[h] = E_s^{ii} + E_c^{ii}$. Assume that $\rho(n) \equiv \frac{1}{n}$, then

$$E_s^{ii} \equiv \frac{1}{n}\sum_{i=1}^{\sqrt{n}} i^2(1-\frac{1}{n})^{2i^2-2i+1} \equiv \frac{1}{n}\sum_{i=1}^{\sqrt{n}} i^2 \equiv \sqrt{n}. \tag{23}$$

Since E_s^{ii} is increasing when $\rho(n)$ is decreasing and its maximum possible order is $\sqrt{n}$, then $E_s^{ii} \equiv \sqrt{n}$ for all $\rho(n) \preceq \frac{1}{n}$.

For $\rho(n) \succ \frac{1}{n}$, we approximate the summation with the integral.

$$\begin{aligned} E_s^{ii} &\equiv \frac{1}{n}\int_{\upsilon=1}^{\sqrt{n}} \upsilon^2(1-\rho(n))^{2\upsilon^2-2\upsilon+1} \\ &\equiv \{\frac{(1-\log(1-\rho(n)))\sqrt{2\pi(1-\rho(n))}erf(\frac{(2\upsilon-1)\sqrt{-\log(1-\rho(n))}}{\sqrt{2}})}{n\log^{3/2}(1-\rho(n))} \\ &\quad + \frac{-2\sqrt{-\log(1-\rho(n))}(2\upsilon+1)(1-\rho(n))^{2\upsilon^2-2\upsilon+1}}{n\log^{3/2}(1-\rho(n))}\}\Big|_{\upsilon=1}^{\sqrt{n}} \end{aligned} \tag{24}$$

where erf is the error function which is always limited by $[-1, 1]$ and is zero at zero. If $\rho(n) \to 1$, then it is obvious that $E_s^{ii} \to 0$. For other values of $\rho(n) \succ \frac{1}{n}$ we use the third approximation in equation (VI), and also $-\log(1-\rho(n)) \equiv \rho(n)$, which is true when $\rho(n)$ tends to zero while n approaches infinity, and $-\log(1-\rho(n)) \equiv 1$ for $\rho(n) \nrightarrow 0$ to prove that $E_s^{ii} \equiv \sqrt{n}$ for $\rho(n) \preceq \frac{1}{n}$, and $E_s^{ii} \equiv \frac{1}{n\rho^{3/2}(n)}$ for $\rho(n) \succ \frac{1}{n}$. Since for $\rho(n) \preceq \frac{1}{n}$ the E_s^{ii} reaches the maximum $E[h]$, therefore E_c^{ii} cannot increase the scaling value of $E[h]$ anymore. For $\rho \succ \frac{1}{n}$ we have $E_c^{ii} \equiv \sqrt{\frac{1}{\rho(n)}}$. Thus it can easily be verified that $E[h] \equiv E_s^{ii} \equiv \sqrt{n}$ for $\rho(n) \preceq \frac{1}{n}$, and $E[h] \equiv E_c^{ii} \equiv \sqrt{\frac{1}{\rho(n)}}$ for $\rho(n) \succ \frac{1}{n}$.

Scenario **iii** - Let us define $E[h] = E_s^{iii} + E_c^{iii}$, where

$$E_s^{iii} = r^2(n)\sum_{i=2}^{\frac{1}{r(n)}} i^2(1-\rho(n))^{inr^2(n)}$$
$$E_c^{iii} = r^2(n)(1-(1-\rho(n))^{nr^2(n)}) \times \{\sum_{i=2}^{\frac{1}{r(n)}} i \sum_{l=1}^{i-1} l(1-\rho(n))^{lnr^2(n)}\} \tag{25}$$

First we check the behavior of E_s^{iii} when $\rho(n) \equiv \frac{1}{nr(n)}$. Using the second region in equation (VI) we will have $E_s^{iii} \equiv \frac{1}{r(n)}$. E_s^{iii} is increasing when $\rho(n)$ is decreasing and the maximum possible value for the number of hops is $\frac{1}{r(n)}$, then $E_s^{iii} \equiv \frac{1}{r(n)}$ for all $\rho(n) \preceq \frac{1}{nr(n)}$.

By approximating the summation with integral, we arrive at

$$E_s^{iii} \equiv r^2(n)\int_2^{\frac{1}{r(n)}} \upsilon^2(1-\rho(n))^{\upsilon nr^2(n)}d\upsilon, \tag{26}$$

which equals to

$$\begin{aligned} &\{(\upsilon^2\log^2(1-\rho(n))^{nr^2(n)} - 2\upsilon\log(1-\rho(n))^{nr^2(n)} + 2) \\ &\times \frac{r^2(n)(1-\rho(n))^{\upsilon nr^2(n)}}{\log^3(1-\rho(n))^{nr^2(n)}}\}\Big|_{\upsilon=2}^{\frac{1}{r(n)}}. \end{aligned} \tag{27}$$

If $\frac{1}{nr(n)} \preceq \rho(n) \preceq \frac{1}{nr^2(n)}$, using the fact that $\log(1-\rho(n))^{nr^2(n)} \equiv -\rho(n)nr^2(n)$ and also equation (VI), we will have $E_s^{iii} \equiv \frac{1}{n^3\rho^3(n)r^4(n)}$.

When $\rho(n) \succeq \frac{1}{nr^2(n)}$, equation (27) tends to zero.

Using the previous approximations along with $1-(1-\rho(n))^{nr^2(n)} \equiv 1$ for $\rho(n) \succeq \frac{1}{nr^2(n)}$, and $\rho(n)nr^2(n)$ for $\rho(n) \preceq \frac{1}{nr^2(n)}$, we can approximate E_c^{iii} as its dominant terms $(E_c^{iii} \equiv \frac{1}{n\rho(n)}\sum_{i=2}^{\frac{1}{r(n)}} i \equiv \frac{1}{\rho(n)nr^2(n)})$.

When $\rho(n) \succeq \frac{1}{nr^2(n)}$, the dominant term is $\Theta(1)$. Thus,

$$E[h] \equiv \begin{cases} E_s^{iii} \equiv \frac{1}{r(n)} & \rho(n) \preceq \frac{1}{nr(n)} \\ E_c^{iii} \equiv \frac{1}{\rho(n)nr^2(n)} & \frac{1}{nr(n)} \preceq \rho(n) \preceq \frac{1}{nr^2(n)} \\ E_c^{iii} \equiv 1 & \frac{1}{nr^2(n)} \preceq \rho(n) \end{cases} \tag{28}$$

It can be seen that for large enough $\rho(n)$ the average number of hops between the nearest content location and the customer is just $\Theta(1)$ hops. This is the result of having $nr^2(n)$ caches in one hop distance for every requester. Each one of these caches can be a potential source for the content. When the network grows, this number will increase and if $\rho(n)$ is large enough $(\frac{1}{nr^2(n)} \preceq \rho(n))$ the probability that at least one of these nodes contain the required data will approach 1, i.e., $\lim_{n\to\infty}(1-(1-\rho(n))^{nr^2(n)}) = 1$. ■

Proof of Lemma 3: Assume that each content is retrieved with rate γ bits/sec. The traffic generated because of one download from a cache (or server) at average distance of $E[h]$ hops from the requester node is $\gamma E[h]$. The total number of requests for a content in the network at any given time is limited by the number of nodes n. Thus the maximum total bandwidth needed to accomplish these downloads will be $nE[h]\gamma$, which is upper limited by $(\Theta(n))$ in scenarios **i**, **ii**, and $(\Theta(\frac{1}{r^2(n)}))$ in scenario **iii**. Thus, $nE[h]\gamma \preceq n$ and $\gamma_{max} \equiv \frac{1}{E[h]}$ in scenarios **i**, **ii**, and $nE[h]\gamma \preceq \frac{1}{r^2(n)}$ and $\gamma_{max} \equiv \frac{1}{E[h]nr^2(n)}$ in scenario **iii**. Therefore the maximum download rate is easily derived using the results of Lemma 2. ■

Proof of Lemma 4: Each link between two nodes in scenarios **i** and **ii**, or two cells in scenario **iii** can carry at most $\Theta(1)$ bits per second. Here we calculate the maximum traffic passing through a link considering the throughput capacities derived in previous theorems, and check if any link can be a bottleneck.

Scenario **i**- Each one of the four links connected to the server will carry all the traffic related to the items not found in the on-path caches. Thus, the total traffic related to item k carried by each of those links is $\psi_k = \sum_{i=1}^{\sqrt{n}} \gamma\, i (1-\rho^{(k)}(n))^i$.

When $\rho^{(k)}(n) \preceq \frac{1}{\sqrt{n}}$, we have $(1-\rho^{(k)}(n))^i \equiv 1$ for all $i \leq \sqrt{n}$. So this traffic is equal to $\psi_k = \sum_{i=1}^{\sqrt{n}} \gamma\, i \equiv n\gamma$.

When $\rho^{(k)}(n) \succeq \frac{1}{\sqrt{n}}$, using equation (VI) the above summation can be written as

$$\gamma \frac{(-1+\rho^{(k)}(n))(\sqrt{n}\rho^{(k)}(n)(1-\rho^{(k)}(n))^{\sqrt{n}}+(1-\rho^{(k)}(n))^{\sqrt{n}}-1)}{(\rho^{(k)}(n))^2}$$
$$\equiv \frac{\gamma}{(\rho^{(k)}(n))^2}. \quad (29)$$

The total traffic is $\psi = \sum_{k=1}^{m} \alpha_k \psi_k$ which must be less than one. If $\rho^{(k)}(n) \succeq \frac{1}{\sqrt{n}}$ for all the items, then the item with minimum $\rho^{(k)}(n)$ will be the dominant factor in the above equation ($\psi \equiv \Theta(\frac{\gamma}{\min_k(\rho^{(k)}(n))^2)})$), and if at least one item has $\rho^{(k)}(n) \preceq \frac{1}{\sqrt{n}}$, it will put the bound on the maximum rate ($\psi \equiv n\gamma$). Thus, $\psi \equiv min(n\gamma, \frac{\gamma}{\min_k(\rho^{(k)}(n))^2)}) \preceq 1$, then $\gamma_{max} \equiv max(\frac{1}{n}, \min_k((\rho^{(k)}(n))^2))$.

Therefore, the links directly connected to the server will be a bottleneck if γ is more than the above values. On the other hand, the traffic related to item k carried by a node to cache content in level j is $\sum_{i=1}^{\sqrt{n}-j} \gamma\, i(1-\rho^{(k)}(n))^i \preceq \sum_{i=1}^{\sqrt{n}} \gamma\, i(1-\rho^{(k)}(n))^i$, so the server links carry the maximum load, and thus the derived upper limits are supportable in every link.

Scenario **ii**- Each one of the four links connected to the server will carry all the traffic related to the items not found in any caches closer to the requester. Thus, the total traffic related to item k (ψ_k) carried by each of those links is

$$\gamma\,(1-\rho^{(k)}(n)) + \sum_{i=1}^{\sqrt{n}} 4\gamma\, i(1-\rho^{(k)}(n))^{(1+4\sum_{j=1}^{i} j)}$$
$$\equiv \gamma\,(1-\rho^{(k)}(n)) + \sum_{i=1}^{\sqrt{n}} \gamma\, i(1-\rho^{(k)}(n))^{2i^2+2i+1},$$
$$\equiv \gamma\,\{(1-\rho^{(k)}(n)) + \frac{(1-\rho^{(k)}(n))^n - (1-\rho^{(k)}(n))^4}{\log(1-\rho^{(k)}(n))/(1-\rho^{(k)}(n))}$$
$$+ \frac{\sqrt{-\frac{\log(1-\rho^{(k)}(n))}{1-\rho^{(k)}(n)}}\, erf(\sqrt{-n\log(1-\rho^{(k)}(n))})}{\log(1-\rho^{(k)}(n))/(1-\rho^{(k)}(n))}$$
$$- \frac{\sqrt{-\frac{\log(1-\rho^{(k)}(n))}{1-\rho^{(k)}(n)}}\, erf(\sqrt{-\log(1-\rho^{(k)}(n))})}{\log(1-\rho^{(k)}(n))/(1-\rho^{(k)}(n))}\}. \quad (30)$$

If $\rho^{(k)}(n) \preceq \frac{1}{n}$, then $(1-\rho^{(k)}(n))^{2i^2+2i+1} \equiv 1$ for all $1 \leq i \leq \sqrt{n}$. Thus the above traffic will be $\psi_k \equiv n\gamma$. If $\rho^{(k)}(n) \succeq \frac{1}{n}$ the above equation is equivalent to $\psi_k \equiv \frac{\gamma}{\rho^{(k)}(n)}$.

The total traffic then is $\psi \equiv \sum_{k=1}^{m} \alpha_k \psi_k \preceq 1$. If $\rho^{(k)}(n) \succeq \frac{1}{n}$ for all the items, then $\psi \equiv \frac{\gamma}{\min_k(\rho^{(k)}(n))}$. If $\rho^{(k)}(n) \preceq \frac{1}{n}$ for at least one item, then $\psi \equiv n\gamma$. Thus, $\psi \equiv min(n\gamma, \frac{\gamma}{\min_k(\rho^{(k)}(n))}) \preceq 1$, then $\gamma_{max} \equiv max(\frac{1}{n}, \min_k(\rho^{(k)}(n)))$.

Using similar reasoning as in scenario **ii** other links carry less traffic, so the above capacities are supportable for all the other links.

Scenario **iii**- The traffic load for item k between the server cell and each of the four neighbor cells (ψ_k) is given by

$$\gamma\, nr^2(n)\{(1-\rho^{(k)}(n)) + \sum_{i=2}^{\frac{1}{r(n)}} i(1-\rho^{(k)}(n))^{inr^2(n)}\}$$
$$\equiv \gamma\, nr^2(n)\{(1-\rho^{(k)}(n))$$
$$+ \frac{(1-\rho^{(k)}(n))^{nr(n)}(nr(n)\log(1-\rho^{(k)}(n))-1)}{\log^2(1-\rho^{(k)}(n))^{nr^2(n)}}$$
$$- \frac{(1-\rho^{(k)}(n))^{nr^2(n)}(\log(1-\rho^{(k)}(n))^{nr^2(n)}-1)}{\log^2(1-\rho^{(k)}(n))^{nr^2(n)}}\} \quad (31)$$

If $\rho^{(k)}(n) \preceq \frac{1}{nr(n)}$, then $(1-\rho^{(k)}(n))^{inr^2(n)} \to 1$ for $2 \leq i \leq \frac{1}{r(n)}$, thus the traffic load equals to $\gamma\, nr^2(n) \sum_{i=2}^{\frac{1}{r(n)}} i \equiv n\gamma$.

If $\frac{1}{nr(n)} \preceq \rho^{(k)}(n) \preceq \frac{1}{nr^2(n)}$, then the maximum traffic load ψ_k on a link is

$$\gamma\, nr^2(n) + \gamma\, nr^2(n) \frac{1+2\rho^{(k)}(n)nr^2(n)}{(\rho^{(k)}(n))^2 n^2 r^4(n)}$$
$$\equiv \frac{\gamma}{(\rho^{(k)}(n))^2 nr^2(n)} \quad (32)$$

If $\rho^{(k)}(n) \succeq \frac{1}{nr^2(n)}$, then equation (31) is equivalent to $\gamma\, nr^2(n)$. Therefore, if $\rho^{(k)}(n) \succeq \frac{1}{nr^2(n)}$ for all the items, then the total traffic ($\psi = \sum_{k=1}^{m} \alpha_k \psi_k$) is simply $\psi \equiv \gamma\, nr^2(n)$. If $\rho^{(k)}(n) \preceq \frac{1}{nr(n)}$ for all items but there is at least one item for which $\rho^{(k)}(n) \preceq \frac{1}{nr^2(n)}$, then the total traffic is dominated by the traffic generated by the item with the least $\rho^{(k)}(n)$ ($\rho^{(k)}(n) \preceq \frac{1}{nr^2(n)}$). And finally if there is at least one item for which $\rho^{(k)}(n) \preceq \frac{1}{nr(n)}$, then it will generate the dominant traffic ($\psi \equiv n\gamma$). Thus, $\psi \equiv min[n\gamma,\ max(\gamma\, nr^2(n), \frac{\gamma}{\min_k(\rho^{(k)}(n))^2 nr^2(n)})] \preceq 1$, $\gamma_{max} \preceq max[\frac{1}{n}, min(\frac{1}{nr^2(n)}, \min_k((\rho^{(k)}(n))^2) nr^2(n))]$. Note that if there is no cache in the system, or $\rho(n)$ is very low, less than the stated threshold values, almost all the requests would be served by the server, and the maximum download rate would be $\Theta(\frac{1}{n})$. ■

REFERENCES

[1] L. Zhang *et al.*, "Named data networking (NDN) project," Relatório Técnico NDN-0001, Xerox Palo Alto Res. Center-PARC, Palto Alto, CA, USA, Oct. 2010.

[2] *PURSUIT: Pursuing a Pub/Sub Internet*. (Sep. 2010). [Online]. Available: http://www.fp7-pursuit.eu/

[3] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, and B. Ohlman, "A survey of information-centric networking," *IEEE Commun. Mag.*, vol. 50, no. 7, pp. 26–36, Jul. 2012.

[4] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard, "Networking named content," in *Proc. 5th Int Conf. Emerg. Netw. Experim. Technol.*, Dec. 2009, pp. 1–12.

[5] B. Ahlgren *et al.*, "Design considerations for a network of information," in *Proc. ACM CoNEXT Conf.*, Dec. 2008, p. 66.

[6] T. Koponen *et al.*, "A data-oriented (and beyond) network architecture," in *Proc. Comput. Commun. Rev. (SIGCOMM)*, Aug. 2007, pp. 181–192.
[7] A. Ghodsi, T. Koponen, B. Raghavan, S. Shenker, A. Singla, and J. Wilcox, "Information-centric networking: Seeing the forest for the trees," in *Proc. HotNets*, Nov. 2011, p. 1.
[8] F. Olmos, B. Kauffmann, A. Simonian, and Y. Carlinet, "Catalog dynamics: Impact of content publishing and perishing on the performance of a LRU cache," in *Proc. 26th Int. Teletraffic Congr. (ITC)*, Sep. 2014, pp. 1–9.
[9] H. Che, Z. Wang, and Y. Tung, "Analysis and design of hierarchical web caching systems," in *Proc. IEEE 20th Annu. Joint Conf. IEEE Comput. Commun. Soc. (INFOCOM)*, Apr. 2001, pp. 1416–1424.
[10] E. Rosensweig, J. Kurose, and D. Towsley, "Approximate models for general cache networks," in *Proc. IEEE INFOCOM*, Mar. 2010, pp. 1–9.
[11] A. Wolman, M. Voelker, N. Sharma, N. Cardwell, A. Karlin, and H. M. Levy, "On the scale and performance of cooperative web proxy caching," *Oper. Syst. Rev.*, vol. 33, no. 5, pp. 16–31, Dec. 1999.
[12] E. J. Rosensweig and J. Kurose, "Breadcrumbs: Efficient, best-effort content location in cache networks," in *Proc. IEEE INFOCOM*, Apr. 2009, pp. 2631–2635.
[13] B. Azimdoost, G. Farhadi, N. Abani, and A. Ito, "Optimal in-network cache allocation and content placement," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2015, pp. 263–268.
[14] L. Yin and G. Cao, "Supporting cooperative caching in ad hoc networks," *IEEE Trans. Mobile Comput.*, vol. 5, no. 1, pp. 77–89, Jan. 2005.
[15] E. J. Rosensweig and J. Kurose, "A network calculus for cache networks," in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 85–89.
[16] M. Tortelli, I. Cianci, L. A. Grieco, G. Boggia, and P. Camarda, "A fairness analysis of content centric networks," in *Proc. Int. Conf. Netw. Future (NOF)*, Nov. 2011, pp. 117–121.
[17] C. Westphal, "On maximizing the lifetime of distributed information in ad-hoc networks with individual constraints," in *Proc. ACM MobiHoc*, May 2005, pp. 26–33.
[18] M. Dehghan *et al.*, "On the complexity of optimal routing and content caching in heterogeneous networks," in *Proc. IEEE Comput. Commun. (INFOCOM)*, Apr. 2015, pp. 936–944.
[19] D. Rossi and G. Rossini, "Caching performance of content centric networks under multi-path routing," Relatório Técnico, Telecom ParisTech, Paris, France, Tech. Rep., 2011.
[20] D. Rossi and G. Rossini, "On sizing CCN content stores by exploiting topological information," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Mar. 2012, pp. 280–285.
[21] P. Gupta and P. Kumar, "The capacity of wireless networks," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 388–404, Mar. 2000.
[22] J. Li, C. Blake, D. S. de Couto, H. I. Lee, and R. Morris, "Capacity of ad hoc wireless networks," in *Proc.7th Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, Jul. 2001, pp. 61–69.
[23] U. Niesen, P. Gupta, and D. Shah, "On capacity scaling in arbitrary wireless networks," *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 3959–3982, Sep. 2009.
[24] M. Grossglauser and D. N. C. Tse, "Mobility increases the capacity of ad hoc wireless networks," *IEEE/ACM Trans. Netw.*, vol. 10, no. 4, pp. 477–486, Aug. 2002.
[25] J. D. Herdtner and E. K. Chong, "Throughput-storage tradeoff in ad hoc networks," in *Proc. IEEE 24th Annu. Joint Conf. Comput. Commun. Soc. (INFOCOM)*, Mar. 2005, pp. 2536–2542.
[26] G. Alfano, M. Garetto, and E. Leonardi, "Content-centric wireless networks with limited buffers: When mobility hurts," *IEEE/ACM Trans. Netw.*, vol. 20, no. 1, pp. 299–311, Feb. 2016.
[27] H. Liu, Y. Zhang, and D. Raychaudhuri, "Performance evaluation of the 'cache-and-forward (CNF)' network for mobile content delivery services," in *Proc. IEEE Int. Conf. Commun. Workshops*, Jun. 2009, pp. 1–5.
[28] G. Carofiglio, M. Gallo, L. Muscariello, and D. Perino, "Modeling data transfer in content-centric networking," in *Proc. 23rd Int. Teletraffic Congr.*, Sep. 2011, pp. 111–118.
[29] S. Gitzenis, G. S. Paschos, and L. Tassiulas, "Asymptotic laws for joint content replication and delivery in wireless networks," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2760–2776, May 2013.
[30] U. Niesen, D. Shah, and G. Wornell, "Caching in wireless networks," *IEEE Trans. Inf. Theory*, vol. 58, no. 10, pp. 6524–6540, Oct. 2012.
[31] M. A. M. Ali and U. Niesen, "Fundamental limits of caching," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2013, pp. 1077–1081.
[32] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 849–869, Feb. 2016.
[33] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 176–189, Jan. 2016.
[34] A. Liu and V. K. N. Lau, "Asymptotic scaling laws of wireless ad hoc network with physical layer caching," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 1657–1664, Mar. 2015.
[35] B. Azimdoost, C. Westphal, and H. R. Sadjadpour, "On the throughput capacity of information-centric networks," in *Proc. 25th Int. Teletraffic Congr. (ITC)*, Sep. 2013, pp. 1–9.
[36] B. N. Bharath, K. G. Nagananda, and H. V. Poor, "A learning-based approach to caching in heterogenous small cell networks," *IEEE Trans. Commun.*, vol. 64, no. 4, pp. 1674–1686, Apr. 2016.
[37] E. Baştuğ, M. Bennis, and M. Debbah, "A transfer learning approach for cache-enabled wireless networks," in *Proc. 13th Int. Symp. Model. Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)*, May 2015, pp. 161–166.
[38] F. Xue and P. Kumar, "Scaling laws for ad hoc wireless networks: An information theoretic approach," in *Foundations and Trends in Networking*. Norwell, MA, USA: NOW Publishers, 2006.
[39] M. D. Penrose, "The longest edge of the random minimal spanning tree," *Ann. Appl. Probab.*, May 1997, pp. 340–361.
[40] V. Martina, M. Garetto, and E. Leonardi, "A unified approach to the performance analysis of caching systems," *IEEE Trans. Model. Perform. Eval. Comput. Syst.*, vol. 1, no. 3, pp. 2040–2048, May 2014.
[41] N. C. Fofack, P. Nain, G. Neglia, and D. Towsley, "Analysis of TTL-based cache networks," in *Proc. 6th Int. Conf. Perform. Eval. Methodologies Tools (VALUETOOLS)*, Oct. 2012, pp. 1–10.
[42] A. Dabirmoghaddam, M. M. Barijough, and G. L. Aceves, "Understanding optimal caching and opportunistic caching at 'the edge' of information-centric networks," in *Proc. 1st Conf. Inf. Centric Netw.*, Sep. 2014, pp. 47–56.
[43] I. Psaras, R. Clegg, R. Landa, W. Chai, and G. Pavlou, "Modelling and evaluation of CCN-caching trees," in *Networking* (Lecture Notes in Computer Science 6640). Heidelberg, Germany: Springer, pp. 78–91.
[44] N. C. Fofack, M. Dehghan, D. Towsley, M. Badov, and D. L. Goeckel, "On the performance of general cache networks," in *Proc. 8th Int. Conf. Perf. Eval. Methodol. Tools (VALUETOOLS)*, Dec. 2014, pp. 106–113.
[45] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.
[46] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless video content delivery through distributed caching helpers," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 1107–1115.
[47] H. Che, Y. Tung, and Z. Wang, "Hierarchical web caching systems: Modeling, design and experimental results," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 7, pp. 1305–1314, Sep. 2002.

Bita Azimdoost received the B.S. and M.S. degrees from the Sharif University of Technology in 1998 and 2000, respectively. She is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, University of California at Santa Cruz, Santa Cruz, CA, USA. Her research interests span over information theory, wireless communications, ad hoc networks, social networks, cache networks, and other future network architectures.

Cedric Westphal received the M.S.E.E. degree from Ecole Centrale Paris in 1995, and the M.S. and Ph.D. degrees in electrical engineering from the University of California at Los Angeles, Los Angeles, in 1995 and 2000, respectively. He is currently a Principal Research Architect with the Huawei Innovation Center, where he is involved in future network architectures, both for wired and wireless networks. He has been an Adjunct Assistant Professor with the University of California at Santa Cruz, Santa Cruz, since 2009. He has authored or co-authored over 80 journal and conference papers, including several best paper awards, and holds over 30 patents. He has served as a reviewer for the NSF, GENI, the EU FP7, and other funding agencies. He has co-chaired the program committee of several conferences, including the IEEE ICC (NGN symposium) and the IEEE NFV-SDN. He was the General Chair for the IEEE INFOCOM 2016. He was an Area Editor of the *ACM Transactions on Networking*, the IEEE TRANSACTIONS ON NETWORKING from 2009 to 2013, an Assistant Editor of *Computer Networks Journal* (Elsevier), and a Guest Editor of the *Ad Hoc Networks Journal* and the IEEE Journal on Selected Areas in Communications.

Hamid R. Sadjadpour (S'94–M'95–SM'00) received the B.S. and M.S. degrees from the Sharif University of Technology, and the Ph.D. degree from the University of Southern California at Los Angeles, Los Angeles, CA. In 1995, he joined the AT&T Research Laboratory, Florham Park, NJ, USA, as a Technical Staff Member and a Principal Member of Technical Staff. In 2001, he joined the University of California at Santa Cruz, Santa Cruz, where he is currently a Professor. He has authored over 170 publications. He holds 17 patents. His research interests are in the general areas of wireless communications and networks. He has served as a Technical Program Committee Member and the Chair in numerous conferences. He is a co-recipient of the best paper awards at the 2007 International Symposium on Performance Evaluation of Computer and Telecommunication Systems and the 2008 Military Communications conference, and the 2010 European Wireless Conference Best Student Paper Award. He has been a Guest Editor of *EURASIP* in 2003 and 2006. He was a member of the Editorial Board of *Wireless Communications and Mobile Computing Journal* (Wiley), and the IEEE JOURNAL OF COMMUNICATIONS AND NETWORKS.