UNIVERSITY OF CALIFORNIA
SANTA CRUZ

PERFORMANCE ANALYSIS OF TODAY'S NETWORKS: NETWORK OF SOCIAL ENTITIES — NETWORK OF CACHES

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

ELECTRICAL ENGINEERING

by

Bita Azimdoost

June 2017

The Dissertation of Bita Azimdoost
is approved:

Professor Hamid R. Sadjadpour, Chair

Professor Donald Wiberg

Professor Cedric Westphal

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Performance Analysis of Today's Networks: Network of Social Entities —

Network of Caches

by

Bita Azimdoost

Extensive growth of Internet applications like communication, education, and leisure, along with the privacy concerns and significant amount of personalized contents, have clearly affected the connectivity of the networks and the users' behavior. The social nature of today's networks' users (the term 'social' is used as a general technical term for network users with specific mutual relationships, and can address a large domain of users, including but not limited to social media users) has in turn a major influence on the networks' shapes and characteristics. On the other hand, with the emergence of advanced hardware and software technologies which enable significant processing power and storage space in mobile devices, and with their ever increasing widespread use, today's Internet is moving from an infrastructure-based network towards a wireless ad-hoc network. Besides, caching data within the network is going to be inevitable to improve latency and reduce bandwidth consumption, and storage is being considered as one of the network primitives.

This research, thus, investigates the impact of the above aspects of today's networks on the performance. It first discusses and evaluates the effect of users' social behavior on the maximum achievable data rate in the wireless ad-hoc networks, and proves that social connection among nodes may actually help in scaling wireless networks. It also reveals that due to their different social status, various users have different effects on the performance, and therefore, traditional transport capacity concept for wireless networks is not appropriate for these types of networks.

Second part of this work investigates the improvements in the fundamental limits of the performance metrics in networks of caches, and evaluates the effect of different caching policies and content searching algorithms on the obtained improvements. Then a framework is presented to quantify the overhead traffic of locating contents, and is later used to define some optimal policies with respect to the contents that should be cached for an operator-driven content distribution system. This framework can generally be used in many other distributed systems contexts where a control plane has to stay aware of the state of the forwarding plane.

To my wonderful husband and amazing son,

for their unlimited love, support, encouragement, and patience.

Also to my parents,

for their unconditional love, and for having taught me to work hard for the things I aspire to achieve.

Acknowledgments

There are many individuals who helped me to make this dissertation possible. This work is not just the result of my efforts but also reflects the mentoring and support I have received.

First and foremost, I want to thank my adviser Hamid Sadjadpour, for his great ideas, wise comments, and generous support and guidance through my long graduate school journey. His enthusiasm and true knowledge influenced me to be a better researcher.

I thank my committee members, Donald Wiberg and Cedric Westphal, for their encouragement and invaluable comments and suggestions. I particularly thank Cedric who not only served on my committee, but also supervised my work when I was an intern in Futurewei Technologies. His thoughtful insights as well as his sincerity motivated me and taught me how to stay positive.

I also thank J.J. Garcia-Luna-Aceves for his insightful comments and great helps during my first years of graduate school.

I lastly thank our Lab members, past and present for helpful discussions and support. I especially want to acknowledge Mohsen Karmizadeh Kiskani for his helpful collaboration, Jose Armando Oviedo for his support and encouragement, and Lemonia Dritsoula for her constructive comments.

Part I

Network of Social Entities

Chapter 1

Introduction

1.1 Research Motivations

The extensive growth of the wireless communications market and widespread use of handheld devices that can connect to the Internet have changed the architecture of the communication networks from an infrastructure-based network towards a wireless ad-hoc network. A wireless ad-hoc network is a self-organized network system in which wireless terminals autonomously construct a multi-hop network. In this system all the nodes inside the transmission range of a transmitter contact as a relay.

At the same time, the social behaviors and interests of users are changing wireless networks into wireless social networks which are significantly influenced by users' social behaviors. In these social networks, the users do not necessarily communicate with a central server [28, 86] and instead, based on their social interests they can communicate with nodes inside a wireless ad-hoc network. Therefore, a portion of future data communication networks can be envisioned as social wireless ad-hoc networks, where the non-uniform characteristics of social contacts don't let the wireless network's nodes to behave completely autonomously, and the multi-hop characteristics of the underlying wireless network restrict the social communications.

While in today's Internet because of the use of fiber optic backbone, the throughput may not be seen a big problem, the rapid increase of video streaming applications like Youtube or Netflix which account for over half of the Internet traffic in North America can potentially be the bottleneck for communications over future wireless networks. A concrete example of such network can be the future 5G networks [28, 86] in which a portion of the data traffic and video streaming should be carried over wireless ad-hoc networks. Therefore, theoretical analysis of capacity for these networks becomes increasingly important.

Many research results have been reported on the capacity of wireless networks after the seminal work by Gupta and Kumar [55]. However, although practical networks are indeed composite networks that have characteristics of both social and communication networks, these results have focused on communication networks and ignore all the social aspects of the entities. There also has been some prior works on understanding of the propagation in composite networks, but they focus on heuristic techniques [71] and not on understanding the fundamental trade-offs or analytic studies.

1.2 Contributions

The present chapter investigates the asymptotic orders[1] of throughput capacity in networks where nodes are entities with social behavior.

Three characteristics of social networks are considered in this chapter. First,in social wireless ad-hoc networks the nodes are selecting their destinations in the context of social groups which means that the nodes are not communicating with random nodes outside their social groups. In many situations, the source may not have a prior knowledge about the members of social groups. Backstrom et al. [18] observed that, the probability that each node being selected as a member of social group decreases with its distance to the origin according to a power-law distribution.

Second, the frequency of communicating with different nodes within a social group is not the same; some nodes are contacted more frequently than others. Latane et al. [67] studied the frequency of social interactions in social networks and observed that in these networks, the probability of choosing the members of a social group is inversely proportional to distance according to a power-law distribution.

Third, the number of members of social groups is also a random number in actual wireless social networks. Studies on complex networks [6, 27, 44, 81], which are a superset of social networks, suggest that these networks are scale-free networks meaning that they have power-law degree distributions.

[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 $g(n)/f(n) \to 0$.

The current chapter tries to address the following questions in such social wireless networks ([10, 13–15, 62]):

- How does the social behavior of the network users affect the order of the maximum achievable information rate?
- How can this information rate be improved?
- Is traditional transport capacity definition [55] appropriate for such networks? Is there any better approach to demonstrate the performance of such networks in terms of throughput capacity?

The main results of this part of the research are summarized in table 1.2.

1.3 Outline

Chapter 2 goes over the previous works done in the two areas of social networks modeling and wireless networks performance analysis. Chapter 3 briefly describes the wireless network models used in the literature, and the social network models and characteristics which distinguish it from ordinary wireless networks. Chapter 4 summarizes some basics about throughput capacity derivation in networks. Chapter 5 entails the theorems and results of our research on performance analysis of social networks, and finally, chapter 6 concludes this chapter.

Ref.	**Size of Each Social Group ($q(n)$)**	**Distance Distribution Between Communicating Peers**	**Peer Selection to Communicate**	**Capacity per Node**
[55]	n	Uniform	Uniform	$\Theta(\frac{1}{nr(n)})$
5.1	Fixed, $\Theta(1)$	Power-Law (α)	Uniform	$\begin{cases} \Theta(\frac{1}{nr(n)}), & for\ 0 \leq \alpha \leq 2 \\ \Theta(\frac{1}{nr^{\alpha-1}(n)}), & for\ 2 \leq \alpha \leq 3 \\ \Theta(\frac{1}{nr^2(n)}), & for\ 3 \leq \alpha \end{cases}$
5.1	Fixed, $o(n)$, $\omega(1)$	Power-Law (α)	Uniform	$\Theta(\frac{1}{nr(n)})$
5.1	Fixed, $\Theta(n)$	Power-Law (α)	Uniform	$\Theta(\frac{1}{nr(n)})$
5.2	Power-Law (γ)	Power-Law (α)	Uniform	$\Theta(\frac{1}{nr(n)})$ in average if $2 \leq \gamma$, $\Theta(\frac{1}{nr(n)})$ for popular nodes $\begin{cases} \Theta(\frac{1}{nr(n)}), & for\ 0 \leq \alpha \leq 2 \\ \Theta(\frac{1}{nr^{\alpha-1}(n)}), & for\ 2 \leq \alpha \leq 3 \\ \Theta(\frac{1}{nr^2(n)}), & for\ 3 \leq \alpha \end{cases}$ for ordinary nodes
5.3	Fixed, $\Theta(1)$	Power-Law (α)	Distance-Based Power-Law (β)	$\begin{cases} \Theta(\frac{1}{nr^{\beta+1}(n)}), & for\ 0 \leq \alpha \leq 2, 0 \leq \beta \leq 1 \\ \Theta(\frac{1}{nr^{\alpha+\beta-1}(n)}), & for\ 0 \leq \alpha + \beta \leq 3, 2 \leq \alpha \\ \Theta(\frac{1}{nr^2(n)}), & Otherwise \end{cases}$
5.3	Fixed, $o(n)$, $\omega(1)$	Power-Law (α)	Distance-Based Power-Law (β)	$\begin{cases} \Theta(\frac{1}{nr(n)}), & for 0 \leq \beta \leq 1 \\ \Theta(\frac{1}{q(n)nr^{\beta+1}(n)}), & for\ \substack{1\leq\beta\leq3 \\ q(n)=\Omega(r^{\beta-1}(n))} \\ \Theta(\frac{1}{q(n)nr^4(n)}), & for\ \substack{3\leq\beta \\ q(n)=\Omega(r^2(n))} \\ \Theta(\frac{1}{nr^2(n)}), & Otherwise \end{cases}$
5.3	Fixed, $\Theta(n)$	Power-Law (α)	Distance-Based Power-Law (β)	$\begin{cases} \Theta(\frac{1}{nr(n)}), & for\ 0 \leq \beta \leq 2 \\ \Theta(\frac{1}{nr^{\beta-1}(n)}), & for\ 2 \leq \beta \leq 3 \\ \Theta(\frac{1}{nr^2(n)}), & for\ 3 \leq \beta \end{cases}$
5.4	Power-Law (γ)	Power-Law (α)	Distance-Based Power-Law (β)	$\begin{cases} \Theta(\frac{1}{nr^{\beta+1}(n)}), & for\ 0 \leq \beta \leq 1 \\ \Theta(\frac{1}{nr^2(n)}), & for\ 1 \leq \beta \end{cases}$

Table 1.1: Summary of the Main Social Characteristics and Results.

Chapter 2

Previous Work

An early work by Milgram on the small-world phenomenon [79] has evoked a considerable attention to the modeling of social networks which include a large family of networks. Studies have shown that the Web [7, 31], scientific collaboration on research papers [82], and general social networks [3] have small-world properties. Several models have been proposed and analyzed in this category. Watts and Strogatz [104] divided the edges of a network into local and long-range contacts and assumed that there is always an edge between a node and any of its local or long-range social contacts. Dietzfelbinger et al. [42] studied a ring-based network where each node is connected to its left and right neighbors and possibly to some further nodes, and the long-range contacts may be selected through any distribution.

Liben-Nowell et al. [72] found a strong correlation between friendship and geographic location in social networks by using data from Live Journal, and Backstrom et. al. [18] observed that in practical networks like Facebook the geography and social relationships are inseparable; the nodes that interact with each other are more likely to be geographically close.

Fraigniaud et al. [48] assumed that the probability of a node being the long-range contact of a source is proportional to the rank of their distance among the distances from the source to all the other nodes and derived the upper bound for the expected number of steps for any source-target pair. Kleinberg [64] proposed a model to explain the small-world phenomenon. His model consists of a two-dimensional extended grid with point-to-point links in which each node has four local contacts and one long-range contact. The source node s selects any other node v as its long-range contact with a probability proportional to $d^{-\alpha}(s, v)$, where $d(s, v)$ is the lattice distance between s and v, and α shows the density of the social network. Li et al. [70] computed upper bounds on the capacity of a wireless network in which source-destination pairs followed a power-law distribution as in Kleinberg's model.

Chapter 3

Network Models

3.1 Wireless Network Models

Properties of wireless ad hoc networks like their throughput capacity are so challenging, that they require, and may thus strongly depend on assumptions on the physical features of the radio channels, on the power assignments, on the node locations, on the traffic matrix, to name a few. The scalability of these properties is of primary concern, hence the results obtained are often of asymptotic nature, valid when the number of nodes is large enough. This can be achieved in two different ways: either the network is deployed on a finite area, with a sufficiently large node density (dense network model), or the node density is kept constant, but the surface is made sufficiently large (extended network model). According to the condition of successful transmission, three following main network models may be used to study the capacity scaling laws in ad hoc wireless networks.

Definition 3.1.1. Protocol Model: *This model assumes a model for successful packet reception at a receiver, by specifying either a guard zone around a receiver or an interference footprint around a transmitter. Node i at position X_i can successfully transmit to node j at position X_j if for any node k at position X_k, $k \neq i$, that transmits on the same subchannel at the same time as i, then $|X_i - X_j| \leq r(n)$ and $|X_k - X_j| \geq (1 + \Delta)r(n)$, where X_i, X_j and X_k are the cartesian positions in the unit square network for these nodes.*

Definition 3.1.2. Physical Model: *This model models successful reception in terms of the received signal-to-noise-plus-interference ratio at a receiver. Let $\{X_k; k \in \mathcal{N}\}$ be the subset of nodes simultaneously transmitting at some time instant over a certain subchannel. All nodes in this subchannel choose a common power level P for all their transmissions. For each subchannel, the noise power is N. A node can transmit over several subchannels. A transmission from a node X_i, $i \in \mathcal{N}$, is successfully received by a node $X_{i(R)}$ if*

$$\frac{\frac{P}{|Xi - X_{i(R)}|^\alpha}}{N + \sum_{k \in \mathcal{N}, k \neq i} \frac{P}{|X_k - X_{k(R)}|^\alpha}} \geq \beta. \tag{3.1}$$

for every sub-channel.

Definition 3.1.3. Information Theoretic Model: *Both the protocol and physicals model are a simplification of the successful transmission condition. The actual amount of information that can be transmitted through the network should be derived from information theory, which is referred as the information theoretical model.*

The network studied in this research is a dense network in a unit square area with n uniformly distributed nodes. We use the protocol model [108] to determine the success of communication in the presence of multiple access interference (MAI). In particular, if χ_i, χ_j and χ_k denote the Cartesian positions in the unit square area for nodes v_i, v_j and v_k, assuming that node $v_k \neq v_i$ transmits on the same sub-channel at the same time as v_i, and $r(n)$ is the common transmission range of all the nodes in the network, then node v_i can successfully transmit to node v_j if $|\chi_i - \chi_j| \leq r(n)$ and $|\chi_k - \chi_j| \geq (1 + \Delta)r(n)$, where $\Delta > 0$ is the guard zone factor. To guarantee connectivity in this network [87], the transmission range ($r(n)$) is assumed to be[1]

$$r(n) = \Omega(\sqrt{\log n/n}) \tag{3.2}$$

As Figure 3.1 illustrates, a TDMA medium access control scheme is assumed to avoid MAI. The network area is divided into square cells with side-length $C_1 r(n)$, ($C_1 < \frac{1}{4}$), and at any given time the cells separated by M-cell distance are the only cells allowed to transmit as shown with a cross sign inside the cells in figure 3.1 where $M \geq (2 + \Delta)/C_1$.

3.2 Social Network Model

3.2.1 Social Network Characteristics

Social network is a network of entities that are linked to each other through some kind of common interest like friendship. The individuals (called nodes) that are

[1] For n points placed uniformly at random on the unit square, the probability that there is no node in the $r(n)$ vicinity of any selected node tends to zero if $r(n)$ is at least $\Theta(\sqrt{\frac{\log n}{n}})$.

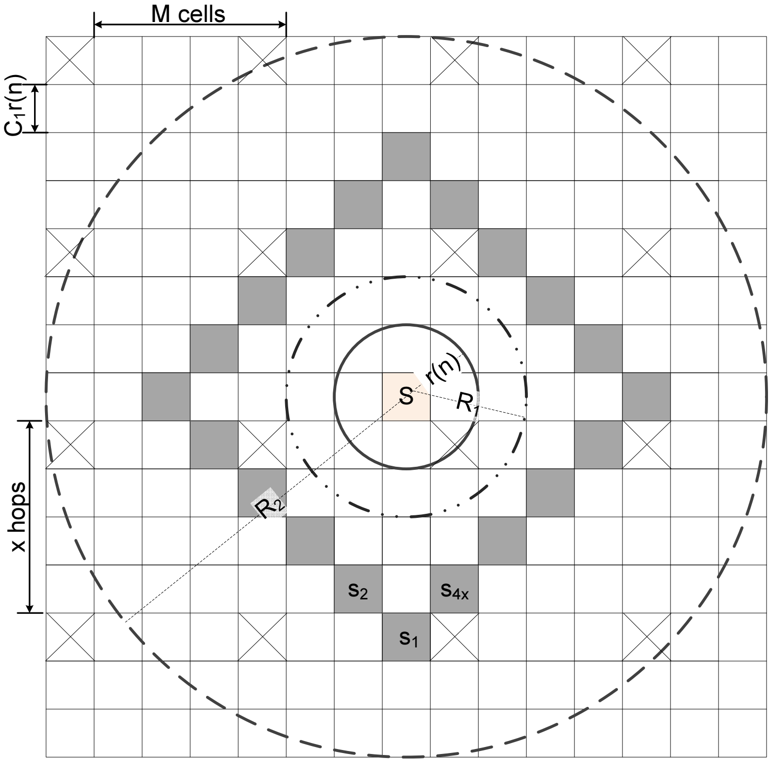

Figure 3.1: **TDMA Channel Access Scheme to Avoid MAI-**The solid-line circle shows the transmission range. Dark gray cells (s_i) contain the nodes with $X = x$. R_1 (R_2) are used as the distance of each node in this region instead of their real distances to achieve upper (lower) bounds on $P(X = x)$.

connected to one node are the social contacts of that node. Studies show that the social behavior imposes some properties on the network structure, out of which we focus on three power-law distributed metrics:

1. Power-law distributed path length - According to [18, 67, 106] the social ties are more probable to get formed between individuals that are closer to each other. In other words, it has been observed [18] that, each node selects its social group members according to a power-law distribution versus distance. In this work we use α as the social group density parameter showing the skew factor of this power-law distribution.

2. Power-law distributed communication length - The frequency of communicating with different nodes within a social group is not the same; some nodes are contacted more frequently than others. Latane et al. [67] studied the frequency of social interactions in social networks and observed that in these networks, nodes tend to communicate with their social contacts that are geographically closer to them. In other words, the probability of choosing the members of a social group to communicate with is inversely proportional to distance according to a power-law distribution. We will consider a power-law distribution with parameter β for the frequency of communications within a social group and call it social communication density.

3. Power-law degree distribution - The number of members of social groups is also a random number in actual wireless social networks. Studies on complex networks

[6, 27, 44, 81], which are a superset of social networks, suggest that these networks are scale-free networks, meaning that they have power-law degree distributions. Therefore, we assume a power law distribution with parameter γ for the size of social groups in our derivations.

3.2.2 Small-World Phenomenon

A social network exhibits the small-world phenomenon if, roughly speaking, any two individuals in the network are likely to be connected through a short sequence of intermediate acquaintances. Recent works has suggested that this phenomenon exists in networks arising in nature and technology, and a fundamental ingredient in the structural evolution of the World Wide Web.

Most of the early work on this issue was based on versions of the following explanation: random networks have low diameter. That is, if every individual were to have a small number of acquaintances selected uniformly at random from the population, and if acquaintanceship were symmetric, then two random individuals would be linked by a short chain with high probability. However, it is obvious that the uniform random model has some limitations; if A and B are two individuals with a common contact, it is much more likely that they themselves are directly connected. But at the same time, a network of acquaintanceships that is too clustered will not have the low diameter.

Watts and Strogatz [104] proposed a model for the small-world phenomenon based on a class of random networks that interpolates between these two extremes, in which the edges of the network are divided into local and long-range contacts. The paradigmatic example they studied was a re-wired ring lattice, containing a set V of n points spaced uniformly on a circle, in which each point is joined by an edge to each of its k nearest neighbors, for a small constant k. These are the local contacts in the network. There are also a small number of edges in which the endpoints are chosen uniformly at random from V, the long-range contacts. Watts and Strogatz argued that such a model captures two crucial parameters of social networks: there is a simple underlying structure that explains the presence of most edges, but a few edges are produced by a random process that does not respect this structure. Their networks thus have low diameter (like uniform random networks), but also have the property that many of the neighbors of a node u are themselves neighbors (unlike uniform random networks). They showed that a number of naturally arising networks exhibit this pair of properties; and their approach has been applied to the analysis of the hyperlink graph of the World Wide Web as well [2]. Figure 3.2 shows some examples of these networks.

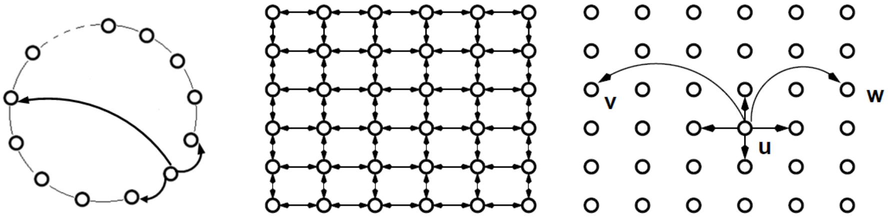

Figure 3.2: **Social Network Models-**Some networks that are formed from a superposition of a structured subgraph and a random subgraph.

3.2.3 A Power-Law Distribution for Social Groups

In Kleinberg's model [64], every node s has a directed edge to every other node v_i within lattice distance $p \geq 1$, and directed edges to $q \geq 0$ other nodes using independent random trials. Each directed edge from s has endpoint $v_i, i = 1, .., n$ with probability proportional to $d_i^{-\alpha} \triangleq d^{-\alpha}(s, v_i)$ and normalizing factor $\sum_{i=1}^{n} d_i^{-\alpha}$. Considering the same probability distribution function for *long-range social contacts* (grouped into set G), the probability that G for a source node contains exactly q independently selected members is the summation of all possible q-member subsets of nodes probabilities.

$$\begin{aligned}
P(|G| = q) &= \sum_{1 \leq i_1 < ... < i_q \leq n} P(G = \{v_{i_1}, ..., v_{i_q}\}) \\
&= \sum_{1 \leq i_1 < ... < i_q \leq n} \prod_{j=1}^{q} P(v_{i_j} \in G) \\
&= \sum_{1 \leq i_1 < ... < i_q \leq n} \frac{d_{i_1}^{-\alpha} ... d_{i_q}^{-\alpha}}{(\sum_{j=1}^{n} d_j^{-\alpha})^q}. \qquad (3.3)
\end{aligned}$$

where v_{i_j} is the i_j^{th} node in the network for $j = 1, ..., q$ and $i_j = 1, ..., n$. As can be seen, this probability is close to one for $q = \Theta(1)$, decreases by increasing q, and approaches zero when $q = \Theta(n)$. Kleinberg [64] assumed that q is a universally constant value and the above derivation proves that the original power-law distribution used in his work should be modified to consider those cases when q is a function of n. We assume that a source node has $q(n)$ long range contacts selected in independent random trials.

The long-range contacts are selected independently, while closer nodes to the source have a better chance of being selected as a G, thus, the probability that a particular q-member set is the G set is proportional to the product of the inverse of the distances of its members from the source. This probability can be written as

$$P(G = \{v_{i_1}, ..., v_{i_q}\}) = \frac{d_{i_1}^{-\alpha} ... d_{i_q}^{-\alpha}}{N_{\alpha,q}}. \tag{3.4}$$

The normalization factor $N_{\alpha,q}$ is obtained using the fact that $\sum_{1 \leq i_1 < ... < i_q \leq n} P(G = \{v_{i_1}, ..., v_{i_q}\}) = 1$.

$$N_{\alpha,q} = \sum_{1 \leq i_1 < ... < i_q \leq n} d_{i_1}^{-\alpha} ... d_{i_q}^{-\alpha} \tag{3.5}$$

The probability that a particular node v_k is selected as a G for source s (i.e., the probability that v_k is a member of the s's G set) is given by

$$\begin{aligned} P(v_k \in G) &= \sum_{1 \leq i_1 < ... < i_{q-1} \leq n, i_j \neq k} P(G = \{v_k, v_{i_1}, ..., v_{i_{q-1}}\}), \\ &= \frac{\sum_{1 \leq i_1 < ... < i_{q-1} \leq n, i_j \neq k} d_k^{-\alpha} d_{i_1}^{-\alpha} ... d_{i_{q-1}}^{-\alpha}}{\sum_{1 \leq i_1 < ... < i_q \leq n} d_{i_1}^{-\alpha} ... d_{i_q}^{-\alpha}}. \end{aligned} \tag{3.6}$$

The above probability function denotes the probability of node v_k being in s's G, and is non-decreasing in q. It also guarantees that the described process ends up with a q-member G set for source node s.

Let ϑ_t be a random variable denoting the destination node. Then, for each particular $v_k \in V$ (the set of nodes except source), we have

$$\begin{aligned} P(\vartheta_t = v_k) &= P(\vartheta_t = v_k \mid v_k \in G) \times P(v_k \in G) \\ &+ P(\vartheta_t = v_k \mid v_k \notin G) \times P(v_k \notin G). \end{aligned} \tag{3.7}$$

Given that the destination is only selected from G, $P(\vartheta_t = v_k \mid v_k \notin G) = 0$.

$$\begin{aligned} P(\vartheta_t = v_k) &= P(\vartheta_t = v_k \mid v_k \in G) P(v_k \in G) \\ &= \frac{\sum_{1 \leq i_1 < \ldots < i_{q-1} \leq n, i_j \neq k} d_k^{-\alpha} \prod_{j=1}^{q-1} d_{i_j}^{-\alpha}}{\sum_{1 \leq i_1 < \ldots < i_q \leq n} \prod_{j=1}^{q} d_{i_j}^{-\alpha}} P(\vartheta_t = v_k \mid v_k \in G) \qquad (3.8) \end{aligned}$$

We use the notation of [80] to denote the elementary symmetric polynomials of the variables $x = (x_1, ..., x_n)$ by $\sigma_{p,n}, 1 \leq p \leq n$. In other words,

$$\sigma_{p,n}(x) = \sigma_{p,n}(x_1, ..., x_n) = \sum_{1 \leq i_1 < i_2 < .. < i_p \leq n} x_{i_1} ... x_{i_p}. \qquad (3.9)$$

Moreover, we define the elementary symmetric polynomials of the same set of variables except one, x_k, as

$$\sigma^{\overline{k}}_{p,n-1}(x_1, ..., x_n) = \sigma_{p,n-1}(x_1, ..., x_{k-1}, x_{k+1}, ..., x_n). \qquad (3.10)$$

Now let $v = (v_1, ..., v_n)$ denote $(d_1^{-\alpha}, ..., d_n^{-\alpha})$, then the above equations can be written as

$$P(v_k \in G) = \frac{d_k^{-\alpha} \sigma^{\overline{k}}_{q-1,n-1}(v)}{\sigma_{q,n}(v)} \qquad (3.11)$$

$$P(\vartheta_t = v_k) = \frac{d_k^{-\alpha} \sigma^{\overline{k}}_{q-1,n-1}(v)}{\sigma_{q,n}(v)} P(\vartheta_t = v_k \mid v_k \in G). \qquad (3.12)$$

This equation is used later to analyze the performance of social networks.

3.3 Routing Model

We have assumed to have a very simple routing algorithm. Each node is assumed to know the locations of its intended destination and its immediate neighbors, and selects as its next hop to the destination the local contact that is closest to the destination. The local contacts are within the radio range since they are the one hop physical neighbors of the node. Assuming that there is at least one local contact in each of the four adjacent cells of the source guarantees that this simple routing protocol converges. If each node has more than four local contacts, i.e., all nodes within transmission range are local contacts, then the order throughput capacity computation does not change and the same results can be derived. The four local contacts assumption was first considered in [64] for grid networks.

Chapter 4

Throughput Capacity

The maximum common data transfer rate which can be achieved on average by all users in a network is called throughput capacity. Let λ denote the data rate for each node and X be the number of hops traveled by each bit from source to destination. The total number of concurrent transmissions per second in such a network is then $n\lambda E[X]$, where $E[X]$ is the average number of hops in a route for any given source-destination pair. This value is upper bounded by the total bandwidth W available, divided by the number of non-interfered groups in the TDMA scheme as shown in Figure 3.1 (*i.e.*, $\frac{W}{M^2C_1^2r^2(n)}$). Therefore, the maximum data rate for each node is [15]

$$\lambda \leq \lambda_{max} = \Theta(\frac{1}{nr^2(n)E[X]}). \tag{4.1}$$

The average number of hops can be computed as

$$E[X] = \sum_{x=1}^{x_{max}} xP(X = x) = P(X = 1) + \sum_{x=2}^{x_{max}} xP(X = x). \tag{4.2}$$

$P(X = 1)$ is the probability that the packets travel just one hop from source to destination, and its value resides between 0 and 1. Since each packet needs to travel at least one hop from the source to reach the destination, the average number of hops between the sources and destinations cannot be less than 1. Therefore, $P(x = 1)$ does not change the order of the $E[X]$ and can be ignored when deriving the order of expected number of hops.

To compute $P(X = x)$ for $x > 1$, we need to consider the long-range contacts outside the circle with radius $r(n)$ centered at the source node. Given that all the nodes inside the transmission range of a source receive the data transmitted from it in just one hop, $P(X = x) = 0$ for $1 < x < \lceil \frac{1}{C_1} + 1 \rceil$. The information between source and destination located on two opposite corners of the network area passes through the maximum number of hops which is $\lceil \frac{2}{C_1 r(n)} \rceil$. Thus, $E[X]$ can be calculated as

$$E[X] \equiv \sum_{\lceil \frac{1}{C_1}+1 \rceil}^{\lceil \frac{2}{C_1 r(n)} \rceil} x P(X = x). \tag{4.3}$$

To compute $P(X = x)$ for $x = \lceil \frac{1}{C_1} + 1 \rceil, ..., \lceil \frac{2}{C_1 r(n)} \rceil$, the number of nodes at a distance of x hops from the source and their corresponding Euclidean distances from the source are required. The geometric place of such nodes is a rhombus around the source node as shown in Figure 3.1 and explained in [15]. The probability that the number of hops between source and destination is x hops equals the probability that the destination is located in one of the cells on the boundaries of this rhombus. Hence,

$$\begin{aligned} P(X=x) &= \sum_{l=1}^{4x} P(destination\ is\ inside\ s_l) \\ &= \sum_{l=1}^{4x} \sum_{v_k\ in\ s_l} P(\vartheta_t = v_k). \end{aligned} \tag{4.4}$$

Therefore,

$$E[X] \equiv \sum_{\lceil \frac{1}{C_1}+1 \rceil}^{\lceil \frac{2}{C_1 r(n)} \rceil} x \sum_{l=1}^{4x} \sum_{v_k\ in\ s_l} P(\vartheta_t = v_k) \tag{4.5}$$

In the following chapter we derive the values of $P(\vartheta_t = v_k)$, to compute $E[X]$ and, consequently, λ_{max} for networks with users that have social characteristics. To make the whole derivation process simpler to follow we add one social property at a time and figure out how each characteristic may change the network throughput.

Chapter 5

Throughput Capacity in Social Networks

5.1 Fixed-Size Groups/Uniform Peer Selection

This section presents a modeling framework for the capacity of a wireless network in which nodes communicate in the context of social groups and successful transmissions can occur only between nodes within transmission range of each other. The model characterizes a wireless network of n nodes each with a social group size that is a function of the number of nodes n, the probability of a node being a long-range social contact of a source that is inversely proportional to their Euclidean distance with power factor α, and MAI is modeled according to the protocol model ([14, 15]).

5.1.1 Results and Discussion

Theorem 5.1.1. *Consider a wireless network consisting of n connected nodes with social behavior modeled by the following properties.*

- *Any two nodes in distance d away from each other are socially connected with a probability inversely proportional to d^{α}, where α is the social group density.*
- *All the nodes have exactly q independent social contacts where $q = 1, .., n-1$.*
- *Each source selects one of its social contacts as its destination randomly with no preference.*

Under these conditions, the maximum capacity order in this wireless social network is

$$\lambda_{max} = \begin{cases} \Theta(\frac{1}{nr(n)}), & for\ q = \Theta(n) \\ \Theta(\frac{1}{nr(n)}), & for\ (q, \frac{q}{n}) \overset{n\to\infty}{\to} (\infty, 0) \\ \Theta(\frac{n-q+1}{n^2r(n)}), & for\ q < \infty, 0 \le \alpha < 2 \\ \Theta(\frac{n-q+1}{n^2r^{\alpha-1}(n)}), & for\ q < \infty, 2 \le \alpha \le 3 \\ \Theta(\frac{n-q+1}{n^2r^2(n)}), & for\ q < \infty, 3 < \alpha \end{cases}$$

It is important to compute the traffic carried in each cell and find out if this throughput capacity can be supported for each cell.

Theorem 5.1.2. *Throughput capacity order obtained in Theorem 5.1.1 is achievable and the flow in no node may become the bottleneck.*

For the purpose of making the theorems simpler to quantify we assume that the according to equation 3.2 the transmission range is the minimum value required to have a connected network ($r(n) = \Theta(\sqrt{\frac{\log n}{n}})$). Figure 5.1 illustrates the results of Theorem 1 by plotting the network capacity as a function of n for different values of α (shown in dash-dot lines) when the number of long-range contacts is a fixed number, i.e., $q(n) = 1$. The solid lines show similar results obtained through simulations which follow closely the theoretical results. It can be observed that the capacity order decreases exponentially as the number of nodes increases. However, increasing the value of α affects the rate of this capacity decrease. Small values of α correspond to the case in which the social groups are highly distributed in the wireless network, and lead to a rate of order-capacity decrease similar to the results derived by Gupta and Kumar [55], in which no social groups exist.

In contrast, for large values of α, social groups are localized, the paths from sources to destinations involve only $\Theta(1)$ hops, and the maximum throughput capacity is achieved. Furthermore, rate of order-capacity decrease is much smaller than with small values of α.

Figure 5.2 shows the throughput capacity versus the power law exponent (α) for two values of $q(n)$. In one case, $q(n)$ is a function of n, i.e., $q(n) = f(n)$, where $f(n)$ is an increasing function of n, and in the second case $q(n)$ is a constant value, i.e., $q(n) = 100$. It can be concluded that if the number of long-range contacts is not a function of the number of nodes, the resulting capacity changes with the parameter α. If α assumes small values ($\alpha \leq 2$), the network behaves as if there were no social

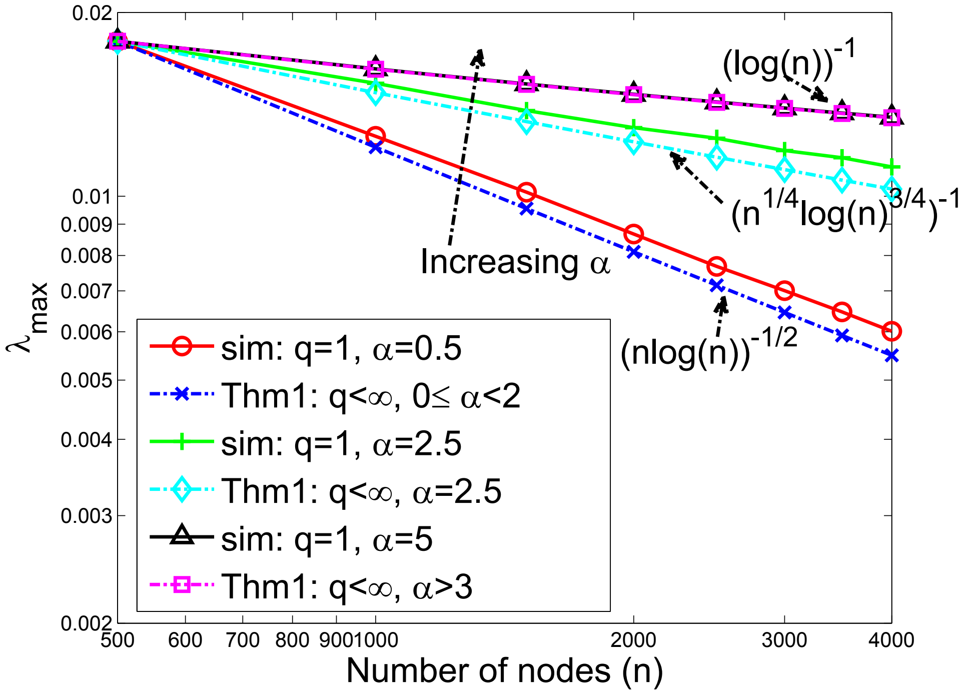

Figure 5.1: **Throughput vs. Network Size (Fixed-Size Groups/Uniform Peer Selection)-**Throughput capacity vs. the number of nodes for different social network densities α, when each source has $q = 1$ long-range contact based on Theorem 1 results (dash-dot curves), and the simulation results (solid curves).

groups. For medium values of α ($2 < \alpha < 3$), an exponential growth is observed in the throughput capacity from $\Theta(1/\sqrt{n \log n})$ to $\Theta(1/\log n)$. For large values of α ($\alpha \geq 3$), each source selects its destination along a path involving only $\Theta(1)$ hops w.h.p. and the resulting capacity is the maximum capacity that can be obtained. We also observe that the rate of capacity increase is very slow for $\alpha > 4$.

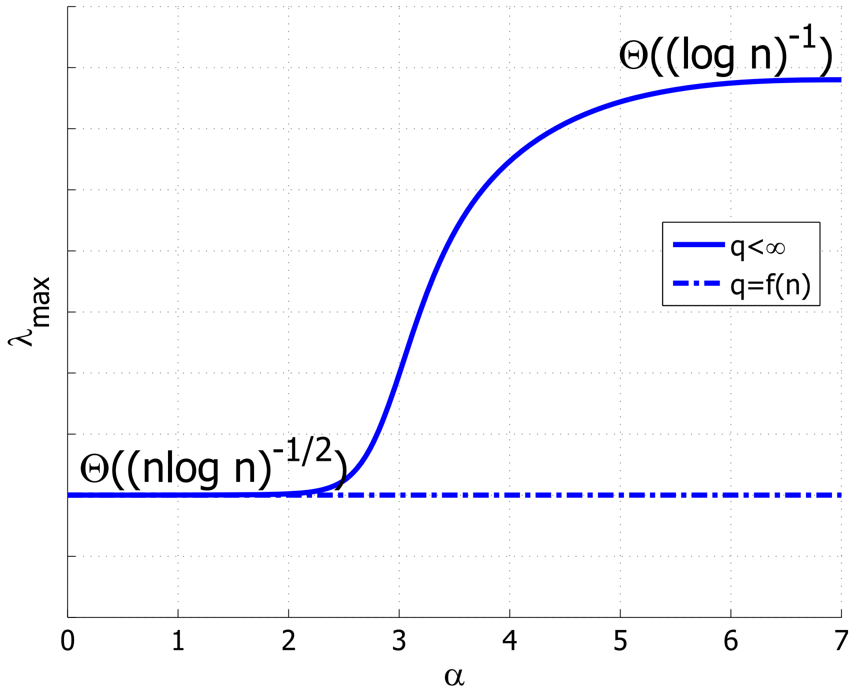

Figure 5.2: **Throughput vs. Social Group Density (Fixed-Size Groups/Uniform Peer Selection)-**Throughput capacity is constant (or changes) with respect to social network density α when each source has infinite $q = f(n)$ (or finite $q = \Theta(1)$) number of social contacts.

However, if the number of long-range social contacts $q(n)$ grows proportional to the number of nodes n, the network behaves as if the network had no social groups, independently of the rate of growth for $q(n)$, and each node selects its destination randomly from all the other network nodes. In this case, the throughput capacity does not change with parameter α, and this is true even if $q(n)$ is much smaller than n, i.e., $q(n) = \log\log(n)$, which is a small number even when n is a very large number.

This phenomenon can be described considering the probability of the source-destination distance (d_{st}) order being $\Theta(1)$. When the number of social contacts of each node is a finite number, this probability is very small, even if that finite number is very large. While in the latter case, if the number of social contacts grows with the network, it can be proved that with high probability the source-destination distance is $\Theta(1)$.

$$\begin{aligned} Pr(d_{st} = \Theta(1)) &= Pr(D_1 < d_{st} < D_2) \\ &= Pr(destination\ is\ inside\ the\ Ring(source, D_1, D_2)) \\ &\equiv \int_{D_1}^{D_2} \frac{nx^{1-\alpha}\sigma_{q-1,n-1}^{\overline{x}}}{q\sigma_{q,n}} dx \end{aligned} \tag{5.1}$$

where $D_1, D_2 < \infty$ are real finite numbers, and Ring (source, D_1, D_2) is a ring with the inner radius of D_1 and outer radius of D_2 centered on the source. Using the approximations and techniques used before in this work, the following probability is proved in Appendix.

$$Pr(d_{st} = \Theta(1)) = \begin{cases} 0 & for\ q < \infty, 2 \leq \alpha \\ \\ 1 & for\ q \to \infty\ or\ 0 \leq \alpha \leq 2 \end{cases} \tag{5.2}$$

Figure 5.3 illustrates the simulation results for a fixed large social group density $\alpha = 5$ for three social group sizes; $q = 1, n-1, n$. It can be seen that the results are very close to the analytical results even when n is not a very large value, i.e., $500 < n < 4000$.

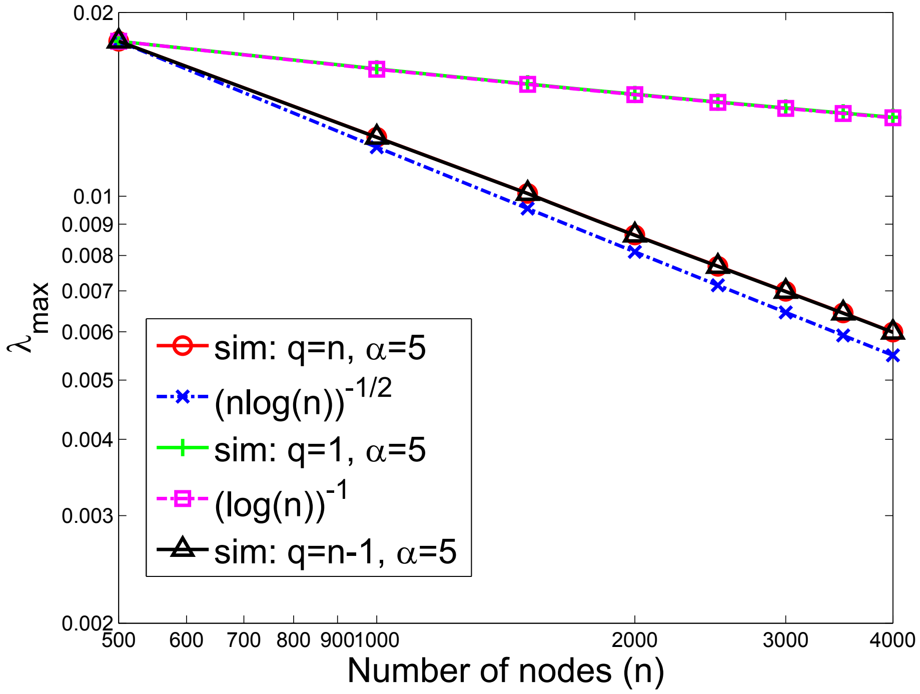

Figure 5.3: **Simulation results (Fixed-Size Groups/Uniform Peer Selection)-** Simulation results for network size $n = 500 - 4000$ and social group size $q = 1, n-1, n$ for social network density $\alpha = 5$. The dash-dot and solid lines correspond to theoretical and simulation results, respectively.

5.1.2 Proofs to Theorems

Proof to Theorem 5.1.1. Since the destination is selected randomly among the q contacts with no preference, we have

$$P(\vartheta_t = v_k \mid v_k \in G) = \frac{1}{q} \tag{5.3}$$

Combining the above equation with equations (3.12) and (4.5) the average number of hops can easily be written as the following equation,

$$E[X] \equiv \sum_{\lceil \frac{1}{C_1}+1 \rceil}^{\lceil \frac{2}{C_1 r(n)} \rceil} x \sum_{l=1}^{4x} \sum_{v_k \ in \ s_l} \frac{d_k^{-\alpha} \sigma^{\overline{k}}_{q-1,n-1}(v)}{q \sigma_{q,n}(v)} \tag{5.4}$$

We now compute the average number of hops based on different values of q as a function of n.

- Case I: q grows with n

  If $q = n$, then $E[X]$ can be rewritten as

  $$E[X] \equiv \sum_{x=\lceil \frac{1}{C_1}+1 \rceil}^{\lceil \frac{2}{C_1 r(n)} \rceil} x \sum_{l=1}^{4x} \sum_{v_k \ in \ s_l} \frac{d_k^{-\alpha} \sigma^{\overline{k}}_{n-1,n-1}(v)}{n \sigma_{n,n}(v)}. \tag{5.5}$$

  Since

  $$\begin{aligned} d_k^{-\alpha} \sigma^{\overline{k}}_{n-1,n-1}(v) &= d_k^{-\alpha} \prod_{i=1, i \neq k}^{n} d_i^{-\alpha} \\ &= \prod_{i=1}^{n} d_i^{-\alpha} = \sigma_{n,n}(v), \end{aligned} \tag{5.6}$$

  then

  $$E[X] \equiv \sum_{x=\lceil \frac{1}{C_1}+1 \rceil}^{\lceil \frac{2}{C_1 r(n)} \rceil} x \sum_{l=1}^{4x} \sum_{v_k \ in \ s_l} \frac{1}{n}. \tag{5.7}$$

Because nodes are uniformly distributed over the network area, there are $nC_1^2r^2(n)$ nodes inside each cell s_l with high probability. Thus[1]

$$\begin{aligned} E[X] &\equiv \sum_{x=\lceil \frac{1}{C_1}+1\rceil}^{\lceil \frac{2}{C_1 r(n)}\rceil} 4x^2C_1^2r^2(n) \\ &\equiv r^2(n)\int_{\lceil \frac{1}{C_1}+1\rceil}^{\lceil \frac{2}{C_1 r(n)}\rceil} u^2du \equiv \frac{1}{r(n)}. \end{aligned} \tag{5.8}$$

Hence, the per-node throughput capacity is $\frac{1}{nr(n)}$, which will lead to the same result obtained by Gupta and Kumar ($\frac{1}{\sqrt{n\log n}}$) [55], if we use the minimum transmission range necessary to guarantee connectivity (equation 3.2). This result is consistent, because the number of social contacts is equal to the total number of nodes in the network, and one of these nodes is selected randomly and uniformly as the destination, which is a similar assumption to that of the original work by Gupta and Kumar [55].

The second case is when $q = \Theta(n)$ but $q \neq n$. Define i.i.d. random variables $Y_i = d_i^{-\alpha}$ for $1 \leq i \leq n$ and define the sequence $Z_i = \log Y_i$ for all values of i. It is obvious that Z_i are i.i.d. as well. Utilizing the law of large numbers, we have $\lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} Z_i = \overline{Z}$ where $\overline{Z}$ is the expected value of random variable Z_i.

[1] Note that we are computing the order of $E[X]$ dropping constant factors.

Thus equation (3.12) can be computed as

$$\begin{aligned}
P(\vartheta_t = v_k) &\equiv \frac{\sum_{1\le i_1<..<i_q\le n, \exists h: i_h = k} \prod_{j=1}^{q} Y_{i_j}}{q \sum_{1\le i_1<..<i_q\le n} \prod_{j=1}^{q} Y_{i_j}} \\
&\equiv \frac{\sum_{1\le i_1<..<i_q\le n, \exists h: i_h = k} \exp \sum_{j=1}^{q} Z_{i_j}}{q \sum_{1\le i_1<..<i_q\le n} \exp \sum_{j=1}^{q} Z_{i_j}}, \\
&\equiv \frac{\sum_{1\le i_1<..<i_q\le n, \exists h: i_h = k} \exp q\overline{Z}}{q \sum_{1\le i_1<..<i_q\le n} \exp q\overline{Z}} \\
&\equiv \frac{\binom{n-1}{q-1}}{q\binom{n}{q}} = \frac{1}{n}. \qquad (5.9)
\end{aligned}$$

Therefore, the value of $E[X]$ is similar to the case $q = n$.

$$E[X] \equiv \sum_{x=\lceil \frac{1}{C_1}+1 \rceil}^{\lceil \frac{2}{C_1 r(n)} \rceil} x \sum_{l=1}^{4x} \sum_{v_k \ in \ s_l} \frac{1}{n} \equiv \frac{1}{r(n)}. \qquad (5.10)$$

Using equation (4.1) provides the capacity as $\lambda_{max} = \Theta(\frac{1}{nr(n)})$.

- Case II: n grows much faster than q

In this case, the expected number of hops between source and destination is obtained when $\lim_{n\to\infty} \frac{q}{n} = 0$, and two mutually exclusive situations must be considered, namely: $lim_{n\to\infty} q = \infty$ and $\lim_{n\to\infty} q < \infty$.

When $\lim_{n\to\infty} q = \infty$, we can use law of large numbers and a similar procedure as before to arrive at

$$\begin{aligned}
E[X] &= \Theta(\frac{1}{r(n)}), \\
\lambda_{max} &= \Theta(\frac{1}{nr(n)}). \qquad (5.11)
\end{aligned}$$

When each node has finite number of contacts ($\lim_{n\to\infty} q < \infty$), the numerator of $P(\vartheta_t = v_k)$ can be expanded as

$$\begin{aligned} d_k^{-\alpha}\sigma_{q-1,n-1}^{\overline{k}}(v) &= d_k^{-\alpha}\sigma_{q-1,n}(v) - d_k^{-2\alpha}\sigma_{q-2,n-1}^{\overline{k}}(v) \\ &= d_k^{-\alpha}\sigma_{q-1,n}(v) - d_k^{-2\alpha}\sigma_{q-2,n}(v) + d_k^{-3\alpha}\sigma_{q-3,n-1}^{\overline{k}}(v) \quad (5.12) \end{aligned}$$

Note that $d_k^{-\alpha}$ and $\sigma_{q-i,n-j}$ are positive values; therefore, the upper and lower bounds for $P(\vartheta_t = v_k)$ are obtained as

$$d_k^{-\alpha}\frac{\sigma_{q-1,n}(v) - d_k^{-\alpha}\sigma_{q-2,n}(v)}{q\sigma_{q,n}(v)} \leq P(\vartheta_t = v_k) \leq \frac{d_k^{-\alpha}\sigma_{q-1,n}(v)}{q\sigma_{q,n}(v)}. \tag{5.13}$$

Lemma 5.1.3. *Let $\Psi = \{\psi_1, ..., \psi_n\}$ be a set of $n \geq 2$ non-negative real numbers. Then for a finite p, i.e., $\lim_{n\to\infty} p < \infty$, we have*

$$\frac{\sigma_{1,n}(\Psi)\sigma_{p,n}(\Psi)}{(p+1)\sigma_{p+1,n}(\Psi)} = \Theta\left(\frac{n}{n-p}\right). \tag{5.14}$$

Proof. Define random variables $U_i^p = \psi_{i_1}...\psi_{i_p}$ for $i = 1, .., \binom{n}{p}$ where $1 \leq i_1 < .. < i_p \leq n$. Due to symmetry, these random variables are identically distributed. Moreover, their mean $\overline{U_p}$ is a function of p. It can be easily seen that these random variables are not independent, as they may have common factors of ψ_{i_j}. We partition the set Ψ into p-member subsets. Assume that T^p is the set of all possible such partitioning (each denoted by T_i^p) with no common member, i.e., $T_i^p \cap T_j^p = \phi$. For a finite p, the number of T^p members is $|T^p| \equiv \binom{n}{p}/\left(\frac{n}{p}\right) = \binom{n-1}{p-1}$. Now we can expand $\sigma_{p,n}(\Psi)$ to separate summations over different partitions described above. Thus,

$$\sigma_{p,n} = \sum_{1\leq i_1<..<i_p\leq n} \psi_{i_1}..\psi_{i_p} = \sum_{j=1}^{|T^p|} \sum_{\{\psi_{i_1}..\psi_{i_p}\}\in T_j^p} \psi_{i_1}..\psi_{i_p}. \tag{5.15}$$

Because each inner summation is applied over one possible partitioning of Ψ, it is performed over $\frac{n}{p}$ of independent U_i as described before. The law of large numbers can be applied here.

$$\lim_{n\to\infty} \sum_{\{\psi_{i_1}..\psi_{i_p}\}\in T_j^p} \psi_{i_1}..\psi_{i_p} = \lim_{n\to\infty} \sum_{\{\psi_{i_1}..\psi_{i_p}\}\in T_j^p} U_i^p = \frac{n}{p}\overline{U_p} \tag{5.16}$$

Thus,

$$\sigma_{p,n} = \sum_{j=1}^{|T^p|} \frac{n}{p}\overline{U_p} = \binom{n}{p}\overline{U_p}. \tag{5.17}$$

A similar formulation can be derived for $\sigma_{p+1,n}(\Psi)$.

$$\sigma_{p+1,n} = \sum_{j=1}^{|T^{p+1}|} \frac{n}{p+1}\overline{U_{p+1}} = \binom{n}{p+1}\overline{U_{p+1}} \tag{5.18}$$

Therefore,

$$\frac{\sigma_{1,n}\sigma_{p,n}}{(p+1)\sigma_{p+1,n}} = \frac{\sigma_{1,n}\binom{n}{p}\overline{U_p}}{(p+1)\binom{n}{p+1}\overline{U_{p+1}}}. \tag{5.19}$$

Note that U_i^p have identical distribution and ψ_i are i.i.d.. Therefore, the expected value $\overline{U_{p+1}}$ can be expressed in terms of $\overline{U_p}$

$$\begin{aligned}
\overline{U_{p+1}} &= E[U_i^{p+1}] = E[\psi_{i_1}...\psi_{i_{p+1}}] \\
&= \sum_{\psi_{i_{p+1}}} E[\psi_{i_1}...\psi_{i_p}\psi_{i_{p+1}}|\psi_{i_{p+1}}]p(\psi_{i_{p+1}}), \\
&= \sum_{\psi_{i_{p+1}}} \psi_{i_{p+1}}E[\psi_{i_1}...\psi_{i_p}]p(\psi_{i_{p+1}}) \\
&= \overline{U_p}\sum_{\psi_{i_{p+1}}} \psi_{i_{p+1}}p(\psi_{i_{p+1}}) \\
&= \overline{U_p}.\overline{\psi_{p+1}} = \overline{U_p}.\overline{\psi}.
\end{aligned} \tag{5.20}$$

Furthermore, by utilizing law of large numbers for $\sigma_{1,n}$ results in $\sigma_{1,n}(\Psi) \to n\overline{\psi}$. Thus

$$\frac{\sigma_{1,n}(\Psi)\sigma_{p,n}(\Psi)}{(p+1)\sigma_{p+1,n}(\Psi)} \equiv \frac{n\binom{n}{p}}{(p+1)\binom{n}{p+1}} = \frac{n}{n-p}. \tag{5.21}$$

□

Returning to the case of finite contacts, we use Lemma 5.1.3 (for $p = q - 1$) and inequality (5.13) to obtain an upper bound for $E[X]$ in equation (5.4).

$$\begin{aligned} E[X] &\leq \sum_{\lceil \frac{1}{C_1}+1 \rceil}^{\lceil \frac{2}{C_1 r(n)} \rceil} x \sum_{l=1}^{4x} \sum_{v_k \ in \ s_l} \frac{d_k^{-\alpha}\sigma_{q-1,n}(v)}{q\sigma_{q,n}(v)} \\ &\equiv \frac{n}{n-q+1} \sum_{\lceil \frac{1}{C_1}+1 \rceil}^{\lceil \frac{2}{C_1 r(n)} \rceil} x \sum_{l=1}^{4x} \sum_{v_k \ in \ s_l} \frac{d_k^{-\alpha}}{\sigma_{1,n}} \end{aligned} \tag{5.22}$$

Referring to the results presented in [15], it can be observed that the average number of hops in this case is $\frac{n}{n-q+1}$ times more than the case when there is only one long-range contact for each source. To calculate the above summation, we need to compute the distance between each node in s_i and the source. To simplify the problem, we use distances $R_1 = xC_1r(n)/A_1$ and $R_2 = A_2xC_1r(n)$ $(A_1, A_2 > 1)$ for all such nodes to reach upper and lower bounds for this summation (see figure 3.1).

$$\begin{aligned} \sum_{l=1}^{4x} \sum_{v_k \ in \ s_l} (A_2xC_1r(n))^{-\alpha} &\leq \sum_{l=1}^{4x} \sum_{v_k \ in \ s_l} d_k^{-\alpha} \\ &\leq \sum_{l=1}^{4x} \sum_{v_k \ in \ s_l} (xC_1r(n)/A_1)^{-\alpha} \end{aligned} \tag{5.23}$$

By replacing the number of nodes in each cell by $nC_1^2r^2(n)$ and ignoring the constant values in the above inequality, we can see that the order of both upper and lower bounds are the same.

$$\sum_{\lceil \frac{1}{C_1}+1 \rceil}^{\lceil \frac{2}{C_1 r(n)} \rceil} x \sum_{l=1}^{4x} \sum_{v_k \ in \ s_l} d_k^{-\alpha} \equiv nr^{2-\alpha}(n) \sum_{\lceil \frac{1}{C_1}+1 \rceil}^{\lceil \frac{2}{C_1 r(n)} \rceil} x^{2-\alpha}$$

$$\overset{a}{\equiv} nr^{2-\alpha}(n) \int_{\lceil \frac{1}{C_1}+1 \rceil}^{\lceil \frac{2}{C_1 r(n)} \rceil +1} u^{2-\alpha} du \qquad (5.24)$$

The last equality (a) is obtained by replacing the sum by its integral approximation. After computing that integral for a sufficiently large value of n which leads to sufficiently small transmission range, we arrive at

$$\sum_{\lceil \frac{1}{C_1}+1 \rceil}^{\lceil \frac{2}{C_1 r(n)} \rceil} x \sum_{l=1}^{4x} \sum_{v_k \ in \ s_l} d_k^{-\alpha} \equiv \begin{cases} \Theta(\frac{n}{r(n)}) & , for \ 0 \leq \alpha \leq 3 \\ \Theta(\frac{n}{r^{\alpha-2}(n)}) & , for \ 3 \leq \alpha \end{cases} \qquad (5.25)$$

Moreover, $\sigma_{1,n}$ can be written as

$$\sigma_{1,n} = \sum_{v_k} d_k^{-\alpha} \equiv \int_{r(n)}^{\gamma d_{max}} n u^{1-\alpha} du, \qquad (5.26)$$

where d_{max} is the maximum distance between any two nodes in the network, and $\gamma \leq 1$. Calculating the integral for a sufficiently large value of n leads to

$$\sigma_{1,n} \equiv \begin{cases} \Theta(n) & for \ 0 \leq \alpha \leq 2 \\ \Theta(\frac{n}{r^{\alpha-2}(n)}) & for \ 2 \leq \alpha \end{cases} \qquad (5.27)$$

The derivations of equations (5.25) and (5.27) are described in the Appendix.

Now we can use these results in equation (5.22) to obtain the following upper bound for $E[X]$. Note that $E[X] \geq 1$; therefore, if the computation ends up with

$E[X] < 1$, we replace it with 1.

$$E[X] = \begin{cases} O(\frac{n}{n-q+1}\frac{1}{r(n)}) & for\ 0 \leq \alpha < 2 \\ O(\frac{n}{n-q+1}\frac{1}{r^{3-\alpha}(n)}) & for\ 2 \leq \alpha \leq 3 \\ O(\frac{n}{n-q+1}) & for\ 3 < \alpha \end{cases}$$

The lower bound capacity follows immediately.

$$\lambda_{max} = \begin{cases} \Omega(\frac{n-q+1}{n^2 r(n)}) & for\ 0 \leq \alpha < 2 \\ \Omega(\frac{n-q+1}{n^2 r^{\alpha-1}(n)}) & for\ 2 \leq \alpha \leq 3 \\ \Omega(\frac{n-q+1}{n^2 r^2(n)}) & for\ 3 < \alpha \end{cases}$$

Thus, these are the upper bounds of $E[X]$ and the lower bounds on the capacity if the number of long-range contacts is a finite number greater than one.

To compute the lower bound for $E[X]$, we will study the lower bound of $P(\vartheta_t = v_k)$ in equation (5.13). First, we calculate the order of $\frac{\sigma_{q-2,n}(v)}{q\sigma_{q,n}(v)}$. This value is obtained by replacing $p = q - 1$ and $p = q - 2$ in equation (5.14).

$$\begin{aligned} \frac{\sigma_{1,n}\sigma_{q-1,n}}{q\sigma_{q,n}} &= \Theta\left(\frac{n}{n-q+1}\right) \\ \frac{\sigma_{1,n}\sigma_{q-2,n}}{(q-1)\sigma_{q-1,n}} &= \Theta\left(\frac{n}{n-q+2}\right) \end{aligned} \tag{5.28}$$

By multiplying these two equations and combining with equation (5.27), we arrive at

$$\begin{aligned}
\frac{\sigma_{q-2,n}}{q\sigma_{q,n}} &= \Theta(\frac{(q-1)n^2}{(n-q+1)(n-q+2)\sigma_{1,n}^2}) \\
&= \begin{cases} \Theta(\frac{(q-1)}{(n-q+1)(n-q+2)}) & for\ 0 \le \alpha \le 2 \\ \\ \Theta(\frac{(q-1)r^{2\alpha-4}(n)}{(n-q+1)(n-q+2)}) & for\ 2 \le \alpha \end{cases} \qquad (5.29)
\end{aligned}$$

The lower bound for $E[X]$ is derived by combining equations (5.4) and (5.13).

$$\begin{aligned}
E[X] &\ge \sum_{\lceil \frac{1}{C_1}+1 \rceil}^{\lceil \frac{2}{C_1 r(n)} \rceil} x \sum_{l=1}^{4x} \sum_{v_k\ in\ s_l} \frac{d_k^{-\alpha}\sigma_{q-1,n}(v) - d_k^{-2\alpha}\sigma_{q-2,n}(v)}{q\sigma_{q,n}(v)} \\
&= \frac{\sigma_{q-1,n}(v)}{q\sigma_{q,n}(v)} \sum_{\lceil \frac{1}{C_1}+1 \rceil}^{\lceil \frac{2}{C_1 r(n)} \rceil} x \sum_{l=1}^{4x} \sum_{v_k\ in\ s_l} d_k^{-\alpha} \\
&- \frac{\sigma_{q-2,n}(v)}{q\sigma_{q,n}(v)} \sum_{\lceil \frac{1}{C_1}+1 \rceil}^{\lceil \frac{2}{C_1 r(n)} \rceil} x \sum_{l=1}^{4x} \sum_{v_k\ in\ s_l} d_k^{-2\alpha}. \qquad (5.30)
\end{aligned}$$

Following similar steps for deriving equation (5.25), we have

$$\begin{aligned}
&\sum_{\lceil \frac{1}{C_1}+1 \rceil}^{\lceil \frac{2}{C_1 r(n)} \rceil} x \sum_{l=1}^{4x} \sum_{v_k\ in\ s_l} d_k^{-2\alpha} \\
\equiv &\begin{cases} \Theta(\frac{n}{r(n)}) & ,for\ 0 \le \alpha \le 3/2 \\ \\ \Theta(\frac{n}{r^{2\alpha-2}(n)}) & ,for\ 3/2 \le \alpha \end{cases} \qquad (5.31)
\end{aligned}$$

If the terms in the negative part of equation (5.30) are replaced with their equivalents from equations (5.29) and (5.31), it is easy to show that for connected networks (minimum transmission range given by equation 3.2), these negative parts will be of an order less than one.

$$\frac{\sigma_{q-2,n}(v)}{q\sigma_{q,n}(v)} \sum_{\lceil \frac{1}{C_1}+1 \rceil}^{\lceil \frac{2}{C_1 r(n)} \rceil} x \sum_{l=1}^{4x} \sum_{v_k \ in \ s_l} d_k^{-2\alpha}$$

$$= \begin{cases} \Theta(\frac{1}{nr(n)}) & , for\ 0 \leq \alpha \leq 3/2 \\ \Theta(\frac{1}{r^{2\alpha-2}(n)}) & , for\ 3/2 \leq \alpha \leq 2 \\ \Theta(\frac{1}{r^2(n)}) & , for\ 2 \leq \alpha \end{cases}$$

$$= \begin{cases} \Theta(\frac{1}{\sqrt{n \log n}}) & , for\ 0 \leq \alpha \leq 3/2 \\ \Theta(\frac{1}{n^{2-\alpha} \log^{\alpha-1} n}) & , for\ 3/2 \leq \alpha \leq 2 \\ \Theta(\frac{1}{\log n}) & , for\ 2 \leq \alpha \end{cases} = o(1) \tag{5.32}$$

Thus, these terms can be ignored compared to the positive part of $E[X]$ and the lower bound for $E[X]$ is the same as its upper bound. Therefore, the obtained bounds on capacity are indeed tight bounds.

□

Proof to Theorem 5.1.2. In order to prove that the throughput capacity is achievable, we just need to compute the total traffic load that a cell is required to accommodate and check if it is not greater than the maximum rate a cell can support.

The traffic load of a node may appear in different situations of being source, relay or destination, the maximum of this value multiplied by the number of nodes in a cell (traffic load of a cell) should not exceed the maximum rate that each cell can support which is $\Theta(1)$.

- Traffic load of a source node

Each source is assumed to transmit data at rate λ, so the maximum load created by each source will be $\Theta(\lambda_{max})$.

- Traffic load of a relay node

We need to compute the maximum number of paths passing through each relay node. To compute this value, we calculate the maximum number of source-destination paths passing through each cell which is ([55, 103])

$$E[X]Pr(Path_i\ intersects\ cell_j) = \\ O(E[X]r^2(n)) < O(nr^2(n)). \tag{5.33}$$

As they are $\Theta(nr^2(n))$ nodes in each cell, using a routing protocol that randomly and uniformly selects one node in a cell to forward the packets will result in the maximum traffic load of a relay node to be $\Theta(\lambda_{max}nr^2(n)/nr^2(n)) = \Theta(\lambda_{max})$.

- Traffic load of a destination node

The power-law distribution of the social contacts leads to a non-uniform distribution for destinations. However, we prove that for large n this distribution is asymptotically uniform. The probability that a node v_k is selected as destination is

$$Pr(v_k\ is\ destination) = \sum_{v_i} Pr(v_k\ is\ destination|v_i\ is\ source)Pr(v_i\ is\ source), \tag{5.34}$$

and as the source nodes are uniformly distributed, this probability is equal to

$$\frac{1}{n}\sum_{v_i} Pr(v_k \textit{ is destination}|v_i \textit{ is source}). \tag{5.35}$$

As we have shown in equation (3.12), the probability inside the summation is equal to $\frac{d_{k_i}^{-\alpha}\sigma_{q-1,n-1,i}^{\overline{v_k}}(v)}{q\sigma_{q,n,i}(v)}$, where index i shows that all the distances in this equation are measured toward source v_i. Replacing this value, which has been shown by $P(\vartheta_t = v_k)$ throughout this work, with the equivalent values obtained for different values of q, it can be easily seen that the probability that v_k is destination will be $\Theta(\frac{1}{n})$.

If q goes to infinity for sufficiently large n,

$$Pr(v_k \textit{ is destination}|v_i \textit{ is source}) = \frac{1}{n}. \tag{5.36}$$

Thus,

$$Pr(v_k \textit{ is destination}) = \frac{1}{n}\sum_{v_i}\frac{1}{n} = \frac{1}{n}. \tag{5.37}$$

If q does not grow with n,

$$Pr(v_k \textit{ is destination}|v_i \textit{ is source}) = \frac{n}{n-q+1}\frac{d_{k_i}^{-\alpha}}{\sigma_{1,n,i}}. \tag{5.38}$$

Thus,

$$Pr(v_k \textit{ is destination}) = \frac{1}{n-q+1}\sum_{v_i}\frac{d_{k_i}^{-\alpha}}{\sigma_{1,n,i}} \tag{5.39}$$

Since $\sigma_{1,n,i}$ has the same order for all i, and by definition is equal to $\sum_{v_i} d_{k_i}^{-\alpha}$, the above equation is equivalent to

$$\frac{1}{n-q+1}\frac{\sum_{v_i} d_{k_i}^{-\alpha}}{\sigma_{1,n}} = \frac{1}{n}. \tag{5.40}$$

The maximum number of routes passing through each cell is $\Theta(nr^2(n))$, thus, it can be concluded that the maximum number of paths destined to each cell has the same order. Thus, similar to the relay traffic load, each destination traffic load will be $\Theta(\lambda_{max})$.

The total traffic load of a node is $\lambda_{max}(\Theta(1) + \Theta(1) + \Theta(1)$, which results in a total traffic of $\lambda_{max}\Theta(nr^2(n))$ for each cell. If the transmission range is greater than $\Theta(\sqrt{\frac{\log n}{n}})$, the traffic load will be $\lambda_{max}\Theta(\log n)$ which is less than $\Theta(1)$ for all values of α and q.

Therefore, the maximum throughput capacity is upper bounded by the inverse of this traffic [74], i.e., $\lambda_{max} \leq \Theta(\frac{1}{\log n})$, which does not violate the throughput capacity bounds we derived earlier. □

5.2 Power-Law Group Size/Uniform Peer Selection

Section 5.1 focused on the homogeneous social networks where all nodes have the same number of social group members. However, study of social networks reveals that the number of social group members for each node is different, and a large number of networks like WWW [7] can be characterized as scale-free networks. More specifically, a small portion of the nodes have a very large social group size while majority of nodes have social contact with a few other nodes in the network. In other words, the number of nodes with q contacts in these networks is inversely proportional to q^γ, where the exponent γ illustrates the clustering property of the network. According to [27], it has been shown that for each social network, this exponent is a constant number which does not change over time. Therefore, different nodes in the network have different social status, i.e., few nodes are highly popular while most of the nodes are less popular in the network. Such networks have been named scale-free networks [6,7]. The present section thus addresses these networks.

5.2.1 Results and Discussion

Theorem 5.2.1. *Consider a social wireless network consisting of n connected nodes with the following properties.*

- *Any two nodes in distance d away from each other are socially connected with a probability inversely proportional to d^{α}, where α is the social group density.*
- *Each node has $q = 1, 2, .., $ or $n - 1$ social contacts and the number of nodes with q social contacts is inversely proportional to q^{γ}, where γ is the social degree distribution exponent.*
- *Each source communicates with one of its social contacts randomly with no preference.*

Under these conditions the throughput capacity will be $\lambda_{max} = \Theta(\frac{1}{nr(n)})$ for sufficiently large n.

Theorem 5.2.1 shows that these networks are not scalable. However, Theorem 5.2.2 demonstrates th at by separating few popular nodes (nodes with large number of social contacts) from the rest of nodes that have few social contacts using different sections of available bandwidth, the majority of nodes can scale.

Theorem 5.2.2. *Consider the social network characterized in theorem 5.2.1 with large social degree distribution exponent ($2 < \gamma$), and assume that social connectivity between nodes is highly concentrated ($\alpha > 2$). Let's divide the total bandwidth (W) into two distinct parts, $W/2$ each; one part to be used to transfer the information generated from the highly connected source nodes ($G_{>q_0}$) and the other part to be used for communication by the source nodes with small social group size ($G_{\leq q_0}$) where q_0 is a constant value independent of n.*

The maximum data rate for the first group ($G_{>q_0}$) is

$$\lambda_{max} = \Theta(\frac{1}{nr(n)}), \ for \ 2 < \alpha. \tag{5.41}$$

The maximum data rate for the second group ($G_{\leq q_0}$) is

$$\lambda_{max} = \begin{cases} \Theta(\frac{1}{nr^{\alpha-1}(n)}), & for \ 2 < \alpha < 3 \\ \Theta(\frac{1}{nr^{2}(n)}), & for \ 3 < \alpha \end{cases} \tag{5.42}$$

Theorem 5.2.3. *The obtained capacity results in theorem 5.2.2 are achievable. In other words, no cell is a bottleneck and the traffic passing through each cell can be routed through.*

According to Theorem 5.2.1 the order of the throughput capacity in scale free networks with uniform peer selection has been derived and proved to be $\Theta(\frac{1}{nr(n)})$, which is $\Theta(\frac{1}{\sqrt{n \log n}})$ for connected network with minimum transmission range (3.2). Further investigation in Theorem 5.2.2 reveals that nodes with different social status, i.e., different number of social contacts, have different effect on throughput capacity. Therefore, traditional transport capacity concept for wireless networks is not appropriate for these types of networks. However, if we divide the nodes into two groups based on their social status and assign to each group half of the available bandwidth, then nodes with small number of social group members can easily scale. On the other hand, the limiting factor in scaling the capacity is the existence of few nodes with high social status that consume majority of the network resources in terms of relaying requirements. More specifically, it is shown that the nodes that limit the capacity consist of a small portion of the network under the condition that the social groups are geographically highly concentrated ($\alpha > 2$) and the degree distribution exponent is large ($\gamma > 2$). Figures 5.5(a) and (b) demonstrate data rates for these two groups of nodes, when $\gamma > 2$.

There exist many other features of social groups that we add in next sections and study the throughput capacity performance of such networks. However, these preliminary results show that the results of Gupta and Kumar and many other papers followed that work are overly pessimistic and social connection among nodes may actually help in scaling wireless networks.

5.2.2 Proofs to Theorems

Proof to Theorem 5.2.1. The number of social contacts of a node, or its degree, is a random variable (Q) which can take the values $q = 1, .., n-1$ with the probability

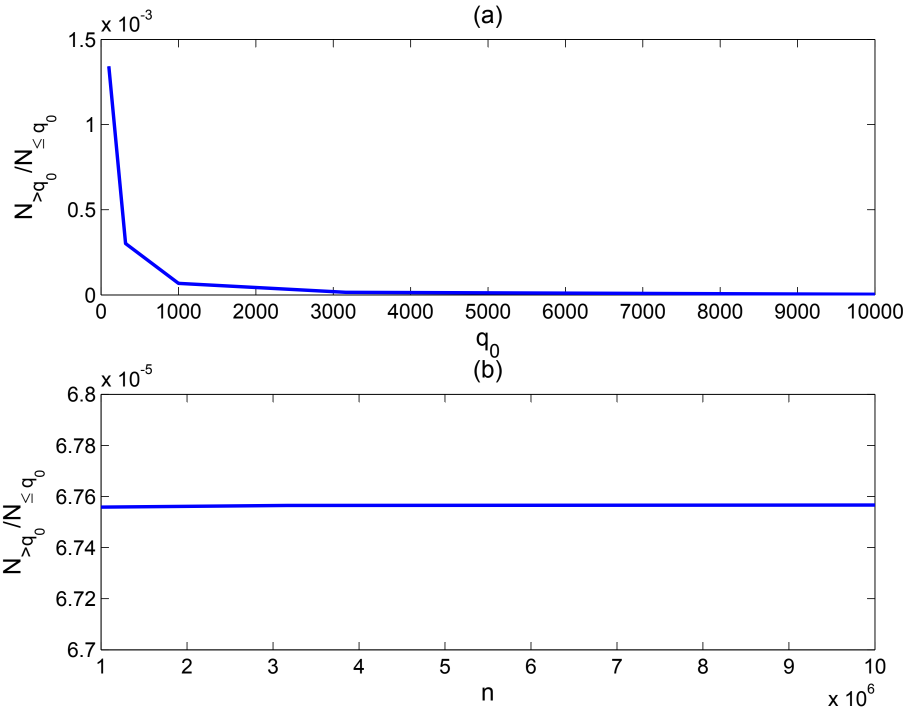

Figure 5.4: **Popular Nodes to Ordinary Nodes Ratio (Power-Law-Distributed Group Size/Uniform Peer Selection)-**The ratio between the number of very popular nodes and less popular ones ($\frac{N_{>q_0}}{N_{\leq q_0}}$) for large degree distribution exponent ($\gamma = 2.3$) and (a) diffrent values of social group size threshold q_0 for a fixed number of nodes ($n = 10^7$), (b) different network sizes and a fixed value for q_0 (10^3).

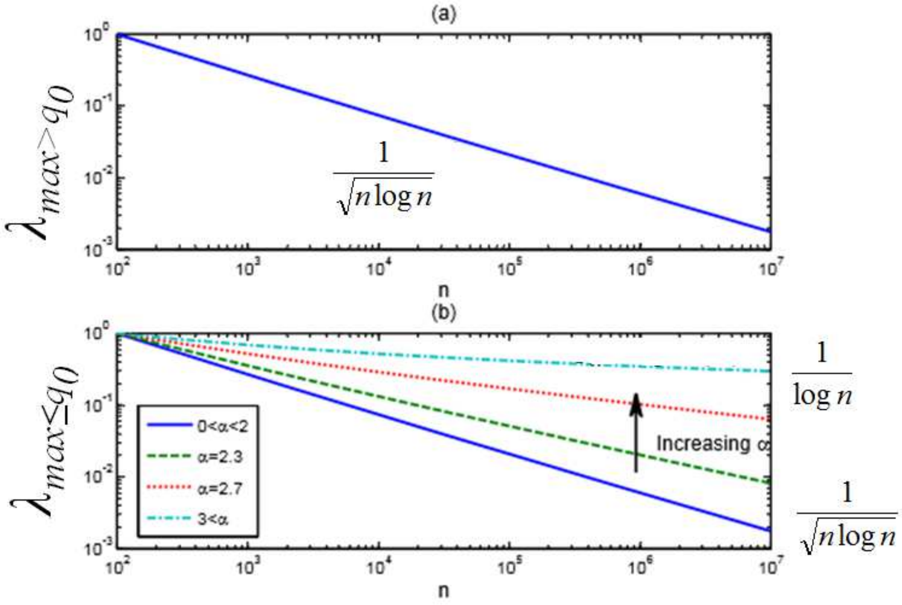

Figure 5.5: **Maximum Achievable Data Rate Order (Power-Law-Distributed Group Size/Uniform Peer Selection)-** for (a) nodes with small social group (Group $G_{\leq q_0}$), (b) highly connected source nodes (Group $G_{>q_0}$), in networks with large degree distribution exponent ($\gamma > 2$).

distribution

$$Pr(Q = q) = \frac{q^{-\gamma}}{N_{q,\gamma,n}}, \tag{5.43}$$

where the normalization factor for this probability ($N_{q,\gamma,n}$) is

$$N_{q,\gamma,n} = \sum_{q=1}^{n-1} q^{-\gamma}. \tag{5.44}$$

Let's assume that each node selects its destination in random from its social contacts, so that the average number of hops (X) passed by the information from source v_i to its destination is

$$E[X|Source = v_i] = \sum_{q=1}^{n-1} Pr(Q = q) E[X|Source = v_i, Q = q]. \tag{5.45}$$

Replacing the value of $Pr(Q = q)$ from (5.43) and $E[X|Source = v_i, Q = q]$ from [14] in the above equation results in

$$E[X|Source = v_i] = \sum_{1}^{\frac{1}{r(n)}} x \sum_{l=1}^{4x} \sum_{v_k \ in \ s_l} \sum_{q=1}^{n-1} \frac{q^{-\gamma}}{\sum_{b=1}^{n-1} b^{-\gamma}} \frac{d_k^{-\alpha} \sigma^{\overline{k}}_{q-1,n-1}(v)}{q \sigma_{q,n}(v)}, \tag{5.46}$$

where d_k is the distance between source and any other node v_k in the network. s_l represents a square cell at distance of x hops from the source node v_i. $\sigma_{q,n}(v)$ is the polynomial symmetric function described in [80] and is equal to $\sum_{1 \le i_1 < .. < i_q \le n} \prod_{j=1}^{q} d_{i_j}^{-\alpha}$. $\sigma^{\overline{k}}_{q-1,n-1}(v)$ is defined in [14] as $\sum_{1 \le i_1 < .. < i_{q-1} \le n, i_h \neq k} \prod_{j=1}^{q-1} d_{i_j}^{-\alpha}$. Expanding the elementary symmetric polynomials, we have

$$\frac{d_k^{-\alpha} \sigma^{\overline{k}}_{q-1,n-1}(v)}{\sigma_{q,n}(v)} = \frac{\sum_{1 \le i_1 < .. < i_q \le n, \exists h : i_h = k} \prod_{j=1}^{q} d_{i_j}^{-\alpha}}{\sum_{1 \le i_1 < .. < i_q \le n} \prod_{j=1}^{q} d_{i_j}^{-\alpha}}. \tag{5.47}$$

Since each d_{i_j} is an independent sample of a random variable (distance between the source and any other random node), we can define i.i.d. random variables $Y_{i_j} = d_{i_j}^{-\alpha}$ for $1 \leq i_j \leq n$ and the random variable sequence $Z_{i_j} = \log Y_{i_j}$ for all values of i_j, which will obviously be i.i.d. as well.

$$\begin{aligned}\frac{d_k^{-\alpha}\sigma_{q-1,n-1}^{\overline{k}}(v)}{\sigma_{q,n}(v)} &= \frac{\sum_{1\leq i_1<..<i_q\leq n, \exists h: i_h=k} \prod_{j=1}^{q} Y_{i_j}}{\sum_{1\leq i_1<..<i_q\leq n} \prod_{j=1}^{q} Y_{i_j}} \\ &= \frac{\sum_{1\leq i_1<..<i_q\leq n, \exists h: i_h=k} \exp(\sum_{j=1}^{q} Z_{i_j})}{\sum_{1\leq i_1<..<i_q\leq n} \exp(\sum_{j=1}^{q} Z_{i_j})}, \qquad (5.48)\end{aligned}$$

For sufficiently large value of q_0 that is independent of n, we can apply the Law of Large Numbers (LLN) for $q > q_0$ and q random variables of type Z_i. Thus for $q > q_0$, for any small $\epsilon > 0$ we can find small $\delta(\epsilon)$ such that $\lim_{Large\ q} \frac{1}{q}\sum_{i=1}^{q} Z_i = \overline{Z} + \epsilon$, with probability $1 - \delta(\epsilon) \to 1$, where $\overline{Z}$ is the expected value of random variable Z_i. Therefore the order of the above equation equals to[2]

$$\frac{\sum_{1\leq i_1<..<i_q\leq n, \exists h: i_h=k} \exp(q(\overline{Z}+\epsilon))}{\sum_{1\leq i_1<..<i_q\leq n} \exp(q(\overline{Z}+\epsilon))} = \frac{\binom{n-1}{q-1}}{\binom{n}{q}} = \frac{q}{n}. \qquad (5.49)$$

Let's define E_1 as

$$E_1 = \sum_{x=1}^{\frac{1}{r(n)}} x \sum_{l=1}^{4x} \sum_{v_k\ in\ s_l} \sum_{q=q_0+1}^{n-1} \frac{q^{-\gamma-1}}{\sum_{b=1}^{n-1} b^{-\gamma}} \frac{d_k^{-\alpha}\sigma_{q-1,n-1}^{\overline{k}}(v)}{\sigma_{q,n}(v)}, \qquad (5.50)$$

which is equal to

$$E_1 \equiv \sum_{x=1}^{\frac{1}{r(n)}} x \sum_{l=1}^{4x} \sum_{v_k\ in\ s_l} \sum_{q=q_0+1}^{n-1} \frac{q^{-\gamma-1}}{\sum_{b=1}^{n-1} b^{-\gamma}} \frac{q}{n} \qquad (5.51)$$

[2] We use the notation $\equiv$ to show the order equality.

The term $\sum_{q=q_0+1}^{n-1} \frac{q^{-\gamma-1}}{\sum_{b=1}^{n-1} b^{-\gamma}} \frac{q}{n}$ is not a function of k or l, so it can be taken out of the summation, and the number of terms of the two summations over k and l is in the order of $x(nr^2(n))$. Thus,

$$\begin{aligned} E_1 &\equiv \frac{r^2(n)}{\sum_{b=1}^{n-1} b^{-\gamma}} \sum_{x=1}^{\frac{1}{r(n)}} x^2 \sum_{q=q_0+1}^{n-1} q^{-\gamma} \\ &\equiv \frac{1}{r(n) \sum_{b=1}^{n-1} b^{-\gamma}} \sum_{q=q_0+1}^{n-1} q^{-\gamma}. \end{aligned} \tag{5.52}$$

The last equality is computed by approximating the sum as integral, i.e., $\sum_{1}^{\frac{1}{r(n)}} x^2 \equiv \frac{1}{r^3(n)}$.

Now define

$$E_2 = \sum_{x=1}^{\frac{1}{r(n)}} x \sum_{l=1}^{4x} \sum_{v_k \ in \ s_l} \sum_{q=1}^{q_0} \frac{q^{-\gamma-1}}{\sum_{b=1}^{n-1} b^{-\gamma}} \frac{d_k^{-\alpha} \sigma_{q-1,n-1}^{\overline{k}}(v)}{\sigma_{q,n}(v)}.$$

Note that $E[X] = E_1 + E_2$. We will compute E_1 and E_2 for different values of α and γ and investigate which one of them will be the dominant factor in computation of $E[X]$.

Lemma 5.2.4. *E_1 has higher order value than E_2 for all values of α and γ.*

Proof. We observe that E_1 is not a function of α.

$$E_1 \equiv \frac{1}{r(n) \sum_{q=1}^{n-1} q^{-\gamma}} \sum_{q=q_0+1}^{n-1} q^{-\gamma} \tag{5.53}$$

If $0 \leq \gamma \leq 1$, it can be easily shown that the order of $\sum_{q=q_0+1}^{n-1} q^{-\gamma}$ and $\sum_{q=1}^{n-1} q^{-\gamma}$ are both equal to $\frac{n^{1-\gamma}}{1-\gamma}$. Therefore, $E_1 \equiv \frac{1}{r(n)}$, and as this is the maximum number of hops that $E[X]$ can have in a unit square area, therefore E_2 does not have any effect on the order of $E[X]$.

If $1 \leq \gamma$, $\sum_{q=1}^{n-1} q^{-\gamma} \equiv \sum_{q=q_0+1}^{n-1} q^{-\gamma} \equiv 1$. Therefore $E_1 \equiv \frac{1}{r(n)}$ and again E_1 will be the dominant factor in computation of $E[X]$. □

Lemma 5.2.4 implies that regardless of the density and clustering degree of the social network, each piece of information needs to travel $\Theta(\frac{1}{r(n)})$ hops on average to reach the destination. Finally combining this result and by utilizing the minimum transmission range to assure connectivity in the network (equation 3.2), the maximum data rate is equal to $\lambda \leq \lambda_{max} = \Theta(\frac{1}{\mathrm{E}[X]nr^2(n)})$. Then, Theorem 5.2.1 is readily proved.

$$\lambda_{max} = \Theta(\frac{W}{E[X]nr^2(n)}) \tag{5.54}$$

$$= \Theta(\frac{1}{nr(n)}) \tag{5.55}$$

W is the available bandwidth in the network.

□

Proof to Theorem 5.2.2. We observe from 5.1 and computation of E_1 and E_2 that when nodes have large social contact size, they require significant network resources to transport unicast data to destinations. On the other hand, when nodes have small social contact size, they require much less network resources in order to transport packets from sources to destinations. Such significant disparity in capacity behavior among nodes suggests that the conventional definition of transport capacity for wireless communication networks is not appropriate for scale free wireless social networks. In these networks, a more accurate analysis should be based on the fact that nodes with different social status in terms of popularity (i.e., number of social contacts) should be grouped separately. For this reason, we divide the bandwidth W into two equal parts and allow communication for each group of nodes within its allocated bandwidth. Note that as we are talking about the order of the throughput capacity, any bandwidth allocation which does not depend on the number of nodes will not change the order of throughput capacity result. Clearly, in order to preserve the connectivity in the network, we still allow nodes in different social status to relay messages for the other group of nodes. However, this condition requires that each node is equipped with two radios, each one operating in different frequency. Further discussion on the details of this approach is beyond the scope of this work.

Lemma 5.2.5. *Let q_0 be a large constant number. For small social degree distribution exponent ($0 < \gamma < 1$), the number of nodes with more than q_0 social contacts ($N_{>q_0}$) is $\Theta(n)$ and the number of nodes with less than q_0 social contacts ($N_{\leq q_0}$) is $\Theta(n^\gamma)$. Further, for large social degree distribution exponent ($2 < \gamma$), the ratio of the number of nodes with more than q_0 social contacts to the number of nodes with less than q_0 social contacts ($\frac{N_{>q_0}}{N_{\leq q_0}}$) is $\Theta(q_0^{1-\gamma})$ which is a very small number for sufficiently large q_0.*

Proof. According to [27], the number of nodes with q connections is proportional to $q^{-\gamma}$ and based on (5.43), the number of nodes with q connections is on average equal to $n\frac{q^{-\gamma}}{\sum_{q=1}^{n-1} q^{-\gamma}}$. Therefore the number of nodes with more than q_0 social connections ($N_{>q_0}$)is

$$N_{>q_0} = n\frac{\sum_{q=q_0+1}^{n-1} q^{-\gamma}}{\sum_{q=1}^{n-1} q^{-\gamma}}, \tag{5.56}$$

and the number of nodes with less than $q_0 + 1$ social connections ($N_{\leq q_0}$)is

$$N_{\leq q_0} = n\frac{\sum_{q=1}^{q_0} q^{-\gamma}}{\sum_{q=1}^{n-1} q^{-\gamma}}. \tag{5.57}$$

These summations can be approximated by integrals.

$$\begin{aligned} 1+\int_2^n \frac{dq}{q^\gamma} \leq \sum_{q=1}^{n-1} q^{-\gamma} \leq 1+\int_2^n \frac{dq}{(q-1)^\gamma} \\ 1+\int_2^{q_0+1} \frac{dq}{q^\gamma} \leq \sum_{q=1}^{q_0} q^{-\gamma} \leq 1+\int_2^{q_0+1} \frac{dq}{(q-1)^\gamma} \\ \int_{q_0+1}^n \frac{dq}{q^\gamma} \leq \sum_{q=q_0+1}^{n-1} q^{-\gamma} \leq \int_{q_0+1}^n \frac{dq}{(q-1)^\gamma} \end{aligned}$$

Therefore the upper and lower bounds for $N_{>q_0}$ and $N_{\leq q_0}$ are

$$n\frac{\frac{1}{1-\gamma}(n^{1-\gamma}-(q_0+1)^{1-\gamma})}{1+\frac{1}{1-\gamma}((n-1)^{1-\gamma}-1)} \leq N_{>q_0} \leq n\frac{\frac{1}{1-\gamma}((n-1)^{1-\gamma}-q_0^{1-\gamma})}{1+\frac{1}{1-\gamma}((n^{1-\gamma}-2^{1-\gamma})}, \tag{5.58}$$

and

$$n\frac{1+\frac{1}{1-\gamma}((q_0+1)^{1-\gamma}-2^{1-\gamma})}{1+\frac{1}{1-\gamma}((n-1)^{1-\gamma}-1)} \leq N_{\leq q_0} \leq n\frac{1+\frac{1}{1-\gamma}(q_0^{1-\gamma}-1)}{1+\frac{1}{1-\gamma}((n^{1-\gamma}-2^{1-\gamma})}. \tag{5.59}$$

Based on these inequalities, it can be easily seen that for small γ (less than 1), the number of nodes with large number of social connections is a tight bound, i.e., $N_{>q_0} = \Theta(n)$. For larger γ, the number of such nodes decreases significantly and will be negligible compared to the number of nodes with very small number of social contacts. When γ is larger than 2, both $N_{>q_0}$ and $N_{\leq q_0}$ are $\Theta(n)$ but their ratio is inversely proportional to $q_0^{\gamma-1}$. Figure 5.4(a) shows the ratio of the number of nodes with more than q_0 social contacts to the number of nodes with less social connections for $\gamma = 2.3$ and $n = 10^7$, and Figure 5.4(b) illustrates the same ratio for similar γ and $q_0 = 1000$. It can be seen that the network size (n) does not considerably affect this ratio as long as it is much more than q_0, and the value of q_0 changes the ratio exponentially. □

In other words, for large γ, the number of nodes involving in E_1 is much less than the nodes which generate the E_2 part of the total average number of hops. E_2 is calculated in appendix and it is shown that for large values of γ and α this term is much smaller than E_1. The following Lemma describes the size of $E1$ and E_2 for large values of α and γ.

Lemma 5.2.6. *In highly concentrated social networks (large α) with large social degree distribution exponent (large γ) , a very small group of nodes ($N_{>q_0}$) use the majority of the resources (due to the large average number of hops traveled by each packet to reach the destination), while a large group of nodes ($N_{\leq q_0}$) use a small portion of the resources.*

This Lemma implies that conventional definition of transport capacity may not be appropriate for scale-free networks. In these networks, transportation of a single packet requires different amount of network resources in terms of relaying and average number of hops to reach destination. Based on this observation, it makes sense that we divide the nodes into two categories. One group of nodes are less popular and their social group size is small, i.e., $N_{\leq q_0}$ and the other group of nodes are those nodes that are more popular with higher social status with many social contacts, i.e., $N_{>q_0}$. We divide the available bandwidth W into two equal parts and allow communication for each group inside their own bandwidth. By doing so, there is more fairness in each group in terms of utilizing the network resources for transmission of packets to destinations which will ultimately allow us to better understand the performance of the network.

For example if $q_0 = 100$ and $\gamma = 2.5$, then it is easy to show that 99.9% of nodes can scale while only 0.1% of nodes with larger than 100 social contacts will not scale.

We have computed E_1 before and by utilizing (5.54), the maximum data rate for sources with $N_{>q_0}$ is given by

$$\lambda_{max_{>q_0}} = \Theta(\frac{W/2}{E_1 nr^2(n)}) = \Theta(\frac{1}{nr(n)}). \tag{5.60}$$

We use the results of appendix and particularly equation (A.20) for sources in the second category, i.e., $N_{\leq q_0}$, to compute the throughput capacity.

$$\begin{aligned} \lambda_{max_{\leq q_0}} &= \Theta(\frac{W/2}{E_2 nr^2(n)}) && (5.61) \\ &= \begin{cases} \Theta(\frac{1}{nr(n)}) & for\ 0 < \alpha < 2 \\ \Theta(\frac{1}{nr^{\alpha-1}(n)}) & for\ 2 < \alpha < 3 \\ \Theta(\frac{1}{nr^2(n)}) & for\ 3 < \alpha \end{cases} && (5.62) \end{aligned}$$

These two capacity results prove Theorem 5.2.2. □

Proof of theorem 5.2.3. The proof of this theorem is very similar to the proof of theorem 5.1.2 For relay and transmit modes we can readily use the same proof as in theorem 5.1.2. For receive mode, we only need to prove that the destinations have a uniform distribution.

The source nodes are uniformly distributed in the network. Thus the probability that a specific node v_k is the destination can be written as

$$\begin{aligned}\Pr(\vartheta_t = v_k) &= \sum_{i=1}^{n} \Pr(\vartheta_t = v_k | v_i \textit{ is source}) \Pr(v_i \ \textit{ is source}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \Pr(\vartheta_t = v_k | v_i \textit{ is source}). \qquad (5.63)\end{aligned}$$

Let G_i be the set of social contacts if node v_i is the source, and Q_i be the number of social contacts of source node v_i. Using equations (3.11) and (3.12) which has been written for one specific source node, we have

$$\begin{aligned}&\Pr(\vartheta_t = v_k | v_i \textit{ is source}) \\ =& \sum_{q=1}^{n} \Pr(\vartheta_t = v_k | v_i \textit{ is source}, v_k \in G_i, Q_i = q) \\ &\quad \times \Pr(v_k \in G_i, Q_i = q) \qquad (5.64)\end{aligned}$$

$$\begin{aligned}=& \sum_{q=1}^{n} \Pr(\vartheta_t = v_k | v_i \textit{ is source}, v_k \in G_i, Q_i = q) \\ &\quad \times \Pr(v_k \in G_i | Q_i = q) \Pr(Q_i = q) \\ =& \sum_{q=1}^{n} \frac{q^{-\gamma}}{q\sigma_1(\mathbf{b})} \frac{{d_k^{-}}^{\alpha}\sigma_{q-1}(\mathbf{d}_{\mathbf{n}}^{\bar{\mathbf{k}}})}{\sigma_q(\mathbf{d}_{\mathbf{n}})} \qquad (5.65)\end{aligned}$$

Now let P_1 and P_2 represent $\sum_{q=q_0+1}^{n} \frac{q^{-\gamma}}{q\sigma_1(\mathbf{b})} \frac{{d_k^{-}}^{\alpha}\sigma_{q-1}(\mathbf{d}_{\mathbf{n}}^{\bar{\mathbf{k}}})}{\sigma_q(\mathbf{d}_{\mathbf{n}})}$ and $\sum_{q=1}^{q_0} \frac{q^{-\gamma}}{q\sigma_1(\mathbf{b})} \frac{{d_k^{-}}^{\alpha}\sigma_{q-1}(\mathbf{d}_{\mathbf{n}}^{\bar{\mathbf{k}}})}{\sigma_q(\mathbf{d}_{\mathbf{n}})}$, respectively. Using the results from [63], we have $\frac{{d_k^{-}}^{\alpha}\sigma_{q-1}(\mathbf{d}_{\mathbf{n}}^{\bar{\mathbf{k}}})}{\sigma_q(\mathbf{d}_{\mathbf{n}})} \equiv \frac{q}{n}$. Also using results from [13] we have $\frac{d_k^{-\alpha}\sigma_{q-1}(\mathbf{d}_\mathbf{n}^{\bar{\mathbf{k}}})}{\sigma_q(\mathbf{d_n})} < \frac{d_k^{-\alpha}q}{\sigma_1(\mathbf{d_n})}$. Therefore,

$$\begin{aligned}
P_1 &= \Theta(\frac{1}{n\sigma_1(\mathbf{b})}\sum_{q=q_0+1}^{n} q^{-\gamma}) \\
P_2 &= O(\frac{d_k^{-\alpha}}{\sigma_1(\mathbf{b})\sigma_1(\mathbf{d_n})}\sum_{q=1}^{q_0} q^{-\gamma})
\end{aligned} \tag{5.66}$$

For large values of γ, $\sigma_1(\mathbf{b})$, $\sum_{q=q_0+1}^{n} q^{-\gamma}$, and $\sum_{q=1}^{q_0} q^{-\gamma}$ are all $\Theta(1)$. Hence, we have $P_1 \equiv \frac{1}{n}$ and $P_2 = O(\frac{d_k^{-\alpha}}{\sigma_1(\mathbf{d_n})})$. Then,

$$\begin{aligned}
\Pr(\vartheta_t = v_k) &= \frac{1}{n}\sum_{i=1}^{n}(P_1 + P_2) \\
&= \frac{1}{n}\sum_{i=1}^{n}\Theta(\frac{1}{n}) + O(\frac{d_k^{-\alpha}}{\sigma_1(\mathbf{d_n})}),
\end{aligned} \tag{5.67}$$

where d_k in the above formulation is the distance from v_k to the source node v_i which can be shown as d_{k_i}. Thus using similar notation for $\mathbf{d_{n_i}}$ we have

$$\sum_{i=1}^{n} O(\frac{d_k^{-\alpha}}{\sigma_1(\mathbf{d_{n_i}})}) = O(\frac{\sigma_1(\mathbf{d_{n_k}})}{\sigma_1(\mathbf{d_{n_i}})}) = O(1), \tag{5.68}$$

that results in $\Pr(\vartheta_t = v_k) \equiv \frac{1}{n}$. Therefore, the destinations are distributed uniformly similar to the relay nodes, and no node in receive mode will be a bottleneck. □

5.3 Fixed-Size Groups/Power-Law Peer Selection

This section focuses on the case when nodes select their destinations inside the social groups based on distance, according to a power law distribution. This assumption is based on a highly cited paper [67] on frequency of communication inside social groups. We assume that within the social group G, the source selects its destination according to a power law distribution with parameter β. Further, the number of long range social contacts for all nodes in this section is assumed to be a same number $q(n)$.

5.3.1 Results and Discussion

Theorem 5.3.1. *Consider a social wireless network consisting of n connected nodes with the following properties.*

- *Any two nodes in distance d away from each other are socially connected with a probability inversely proportional to d^{α}, where α is the social group density.*
- *All the nodes have exactly q independent social contacts where $q = 1, .., n-1$.*
- *Each source communicates with one of its social contacts randomly with a probability inversely proportional to d^{β}, where β is the social communication density.*

Under these conditions the throughput capacity will be

- *If $q = \Theta(n)$:*

$$\lambda_{max} = \begin{cases} \Theta(\frac{1}{nr(n)}), & for\ 0 \leq \beta \leq 2 \\ \Theta(\frac{1}{nr^{\beta-1}(n)}), & for\ 2 \leq \beta \leq 3 \\ \Theta(\frac{1}{nr^{2}(n)}), & for\ 3 \leq \beta \end{cases}$$

- *If* $q \overset{n\to\infty}{\to} \infty$ *and* , $\frac{q}{n} \overset{n\to\infty}{\to} 0$*:*

$$\lambda_{max} = \begin{cases} \Theta(\frac{1}{nr(n)}), & for\ 0 \leq \beta \leq 1 \\ \Theta(\frac{1}{q(n)nr^{\beta+1}(n)}), & for\ {}^{1\leq\beta\leq3,}_{q(n)=\Omega(r^{\beta-1}(n))} \\ \Theta(\frac{1}{q(n)nr^{4}(n)}), & for\ {}^{3\leq\beta,}_{q(n)=\Omega(r^{2}(n))} \\ \Theta(\frac{1}{nr^{2}(n)}), & Otherwise \end{cases}$$

- *If* $q = \Theta(1)$*:*

$$\lambda_{max} = \begin{cases} \Theta(\frac{1}{nr^{\beta+1}(n)}), & for\ {}^{0\leq\alpha\leq2,}_{0\leq\beta\leq1} \\ \Theta(\frac{1}{nr^{\alpha+\beta-1}(n)}), & for\ {}^{0\leq\alpha+\beta\leq3,}_{2\leq\alpha} \\ \Theta(\frac{1}{nr^{2}(n)}), & Otherwise \end{cases}$$

Theorem 5.3.2. *These capacity results are achievable. In other words, no cell is a bottleneck and the traffic passing through each cell can be routed through.*

Here we take a closer look at each of the regions stated in Theorem 5.3.1. In case of $q = \Theta(n)$, by replacing $r(n)$ with its minimum value for a connected network, the maximum achievable throughput is given by

$$\lambda_{max} = \begin{cases} \Theta(\frac{1}{\sqrt{n\log n}}), & 0 \leq \beta < 2 \\ \Theta(\frac{1}{\log n}\sqrt{\frac{\log n}{n}}^{3-\beta}), & 2 \leq \beta \leq 3 \\ \Theta(\frac{1}{\log n}). & 3 < \beta \end{cases} \tag{5.69}$$

This result demonstrates that for $q = \Theta(n)$ and when the destination is selected based on distance, then the throughput capacity is independent of α. Further, we can achieve highest possible capacity ($\lambda_{max} = \Theta(\frac{1}{\log n})$) even for small values of α when $\beta > 3$. Based on this observation, it can be concluded that for this case, selecting destination based on distance is the dominant factor. This result can be justified by observing that since the total number of social contacts is proportional to n, then selecting them based on a power law distribution (with parameter α) does not make much difference since most of the nodes belong to all social groups. As we see in the next region ($q = \Theta(1)$), the effect of α will appear as the nodes become more selective in choosing the members of their social groups.

If $q = \Theta(1)$, then replacing $r(n)$ with its minimum value, gives $\lambda_{\max}$ as

$$\lambda_{\max} \equiv \begin{cases} \Theta(\frac{1}{\log n}\sqrt{\frac{\log n}{n}}^{1-\beta}), & 0 \leq \beta \leq 1, 0 \leq \alpha \leq 2 \\ \Theta(\frac{1}{\log n}\sqrt{\frac{\log n}{n}}^{3-\alpha-\beta}), & 0 \leq \alpha + \beta \leq 3, 2 \leq \alpha \\ \Theta(\frac{1}{\log n}). & \text{Otherwise} \end{cases} \tag{5.70}$$

The results indicate that when both α and β are small, then social characteristics of the network has little effect on the throughput capacity which is the first capacity region for this case. However, by increasing the value of α beyond the threshold of 2, social characteristics start influencing and increasing the throughput capacity while the effect of communication network decreases (second capacity region). When we move beyond these values, social characteristics become dominant factor and the communication network does not have any effect on the capacity of the network. In this capacity region, average hop count is proportional to $\Theta(1)$ which is the direct result of strong social aspects of the network.

Similarly, in case of $q \overset{n\to\infty}{\to} \infty$ and , $\frac{q}{n} \overset{n\to\infty}{\to} 0$, the achievable throughput for minimum transmission range is derived as

$$\lambda_{\max} \equiv \begin{cases} \Theta(\frac{1}{f(n)\log n}\sqrt{\frac{\log n}{n}}^{1-\beta}), & 1 \leq \beta \leq 3, f(n) = \Omega(\sqrt{\frac{\log n}{n}}^{\beta-1}) \\ \Theta(\frac{1}{f(n)\log n}\frac{n}{\log n}), & 3 \leq \beta, f(n) = \Omega(\frac{\log n}{n}) \\ \Theta(\frac{1}{\log n}). & \text{Otherwise} \end{cases}$$

The result in this region provides insight on the behavior of throughput capacity as a function of the number of social contacts for each node. This part explains how different social characteristics of the network that are represented by two parameters of β and α (α in these equations is indirectly reflected in $f(n)$) influences the throughput capacity in the most general case.

To summarize, in general when the social characteristics of the network become a dominant factor, then the throughput capacity of the network improves. On the other hand, when the wireless communication characteristics of the network is dominant, the throughput capacity will decrease up to the point that in the extreme case, it will be the same as Gupta-Kumar result (small β in case that the social group size increases with the network size, and when $\beta = 0$ in case of fix social groups).

The results in this research, which are obtained through mathematical proofs are expressed in terms of scaling laws. In order to validate our theoretical results with simulations, we need to use very large values for n. However, using very large values for n is not practical due to the non-polynomial number of computations. For instance, we have $\binom{n}{q}$ different possibilities to choose q members of the social group from the total number of nodes n. Each one of these choices has an associated probability expressed as

$$\Pr(G = \{v_{g_1}, ..., v_{g_q}\}) = \frac{d_{g_1}^{-\alpha} ... d_{g_q}^{-\alpha}}{\sum_{1 \leq i_1 < ... < i_q \leq n} d_{i_1}^{-\alpha} ... d_{i_q}^{-\alpha}} \tag{5.71}$$

This means that for any numerical simulation, we need to compute the associated probabilities. Now, if $q = \Theta(f(n))$, then we should compute these probabilities for at least $\binom{n}{q} = \binom{n}{f(n)} \geq \left(\frac{n}{f(n)}\right)^{f(n)}$ different choices. This value grows faster than exponential for many choices of $f(n)$. Therefore, conducting a comprehensive numerical analysis for the theoretical results in this work is almost impossible except for special cases of $q = \Theta(1)$ and $q = \Theta(n)$. We simulated our results and compared them against the theoretical results. Figure 5.6 shows the average hop count in theory and by simulation. The results clearly demonstrate that our theoretical derivations are very close to simulation results as the number of nodes in the network increases. For the case of $\beta = 3.5$, we only show the simulation results which is consistent with theory, i.e., $E[X] = \Theta(1)$.

Figure 5.7 demonstrates the maximum throughput as a function of n when $q = \Theta(n)$. We can see from this figure, that for different values of α and β, the simulation results are very close to theoretical results which verifies the accuracy of the analytic work.

Figures 5.6 and 5.7 compare the simulation results with theory for the case of $q = \Theta(1)$. Both the analytical results for the average number of hops and throughput capacity for different values of α and β are close to simulation results. From all these

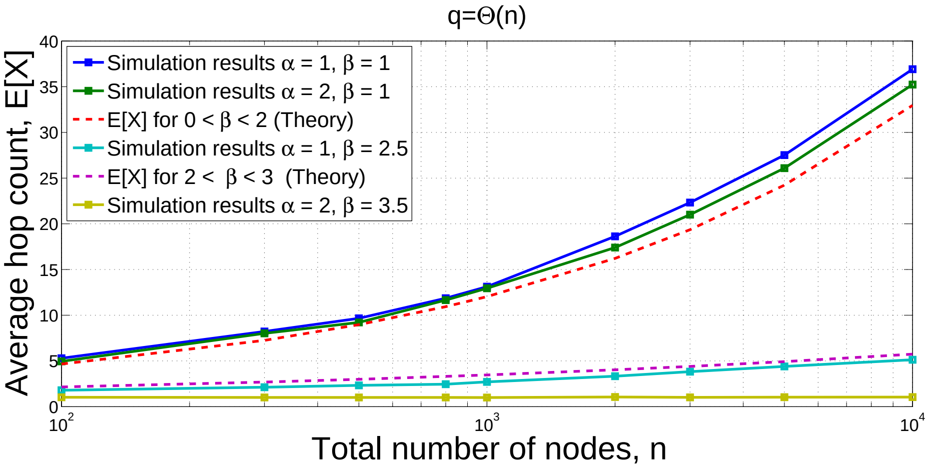

Figure 5.6: **Hop Count vs. Network Size When $q = \Theta(n)$ (Fixed-Size Groups/Power-Law-Distributed Peer Selection).**

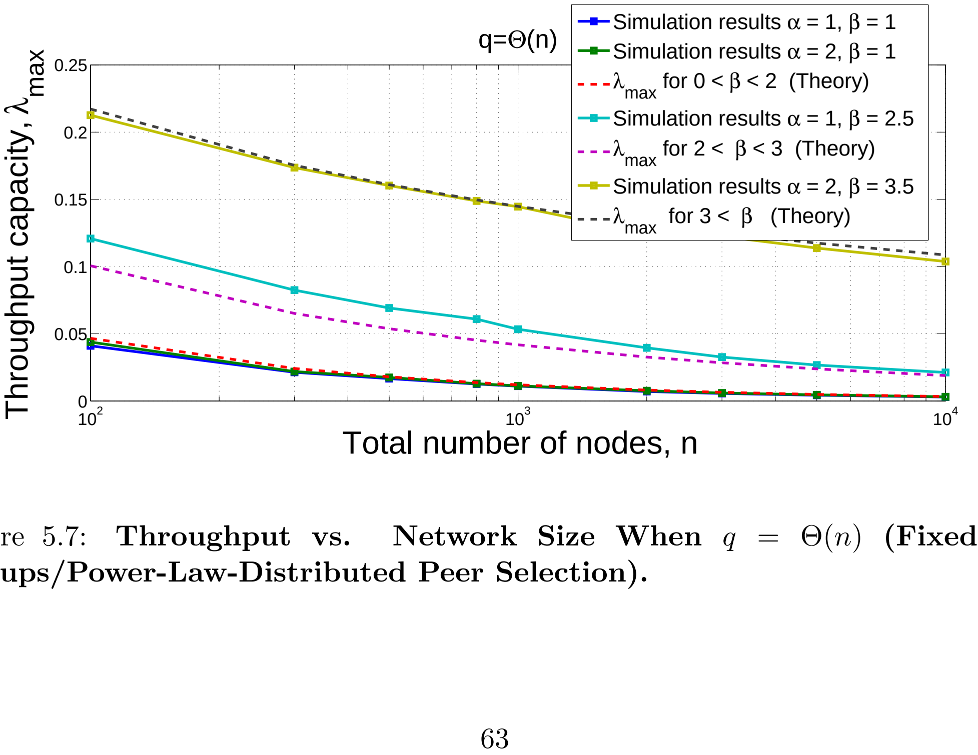

Figure 5.7: **Throughput vs. Network Size When $q = \Theta(n)$ (Fixed-Size Groups/Power-Law-Distributed Peer Selection).**

results, we can conclude that the analytical results accurately predict the behavior of the network when social characteristics are considered.

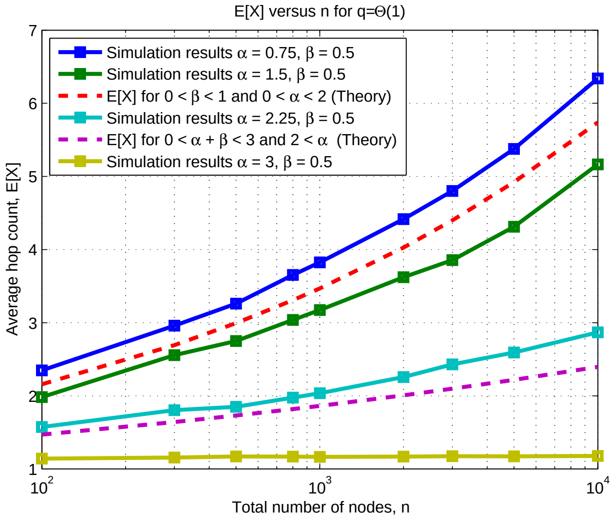

Figure 5.8: **Latency vs. Network Size When $q = \Theta(1)$ (Fixed-Size Groups/Power-Law-Distributed Peer Selection).**

5.3.2 Proofs to Theorems

Proof to Theorem 5.3.1. By defining $\mathbf{d_q} = (d_{g_1}^{-\beta}, ..., d_{g_q}^{-\beta})$, we have

$$\Pr(\vartheta_t = v_k \mid v_k \in G) = \frac{d_k^{-\beta}}{\sum_{j=1}^{q} d_j^{-\beta}} = \frac{d_k^{-\beta}}{\sigma_1(\mathbf{d_q})}, \tag{5.72}$$

which reduces equation (4.5) to

$$E[X] = \sum_{x=1}^{\frac{1}{r(n)}} x \sum_{l=1}^{4x} \sum_{v_k \in s_l} \frac{d_k^{-\alpha-\beta} \sigma_{q-1}(\mathbf{d_n^{\bar{k}}})}{\sigma_1(\mathbf{d_q}) \sigma_q(\mathbf{d_n})}. \tag{5.73}$$

Next, we state some lemmas (with proofs in Appendix) to compute the value of $E[X]$ based on the size of social group.

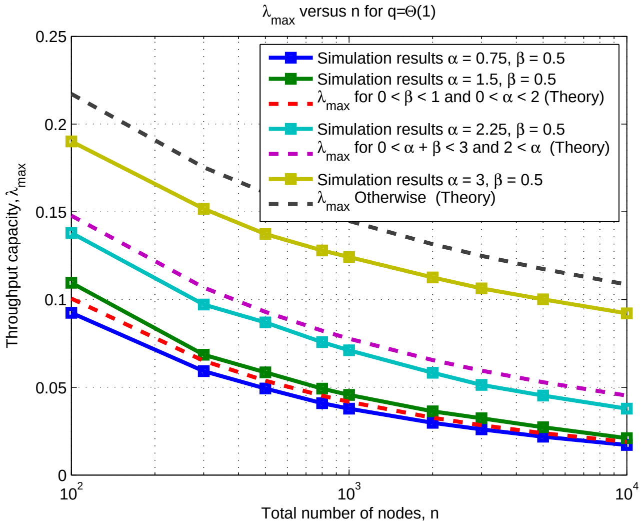

Figure 5.9: **Throughput vs. Network Size When $q = \Theta(1)$ (Fixed-Size Groups/Power-Law-Distributed Peer Selection).**

Lemma 5.3.3. *When $\lim_{n\to\infty} q = \infty$, we have $\frac{d_k^{-\alpha}\sigma_{q-1}(\mathbf{d}_\mathbf{n}^{\bar{\mathbf{k}}})}{\sigma_q(\mathbf{d}_\mathbf{n})} \equiv \frac{q}{n}$.*

specifically, when $q = \Theta(n)$, we have $\frac{d_k^{-\alpha}\sigma_{q-1}(\mathbf{d}_\mathbf{n}^{\bar{\mathbf{k}}})}{\sigma_q(\mathbf{d}_\mathbf{n})} \equiv \Theta(1)$.

The following two lemmas stated and proved in [63]. We will restate them here and use them to prove Theorem 5.3.1.

Lemma 5.3.4.

$$\sum_{x=1}^{\frac{1}{r(n)}} x \sum_{l=1}^{4x} \sum_{v_k \in s_l} d_k^{-\beta} \equiv \begin{cases} \Theta\left(nr^{-1}(n)\right), & 0 \le \beta \le 3 \\ \Theta\left(nr^{2-\beta}(n)\right), & 3 \le \beta \end{cases} \tag{5.74}$$

Lemma 5.3.5.

$$\sigma_1(\mathbf{d_n}) \equiv \begin{cases} \Theta\left(n\right), & 0 \le \alpha \le 2 \\ \Theta\left(nr^{2-\alpha}(n)\right), & 2 \le \alpha \end{cases} \tag{5.75}$$

Also notice that when $q = \Theta(n)$, we have

$$\sigma_1(\mathbf{d_q}) \equiv \begin{cases} \Theta\left(n\right), & 0 \leq \beta \leq 2 \\ \\ \Theta\left(nr^{2-\beta}(n)\right). & 2 \leq \beta \end{cases} \tag{5.76}$$

$E[X]$ in case of $q = \Theta(n)$ can be derived as a direct result of lemmas 5.3.3, 5.3.4, and 5.3.5 used in equation (5.73). Corresponding throughput capacity for this region in Theorem 5.3.1 is obtained using equation (4.1).

Lemma 5.3.6. *When $q = \Theta(1)$ or $q = \Theta(f(n))$ where $\lim_{n\to\infty} \frac{f(n)}{n} = 0$, then $\sigma_1(\mathbf{d_q})$ has the order of $\Theta\left(r(n)^{-\beta}\right)$.*

Lemma 5.3.7. *The following inequalities hold.*

$$\sigma_{q-1}(\mathbf{d_n}) - d_k^{-\alpha}\sigma_{q-2}(\mathbf{d_n}) \leq \sigma_{q-1}(\mathbf{d_n^{\overline{k}}}) \leq \sigma_{q-1}(\mathbf{d_n}) \tag{5.77}$$

We can now use lemma 5.3.3 to simplify equation (5.73) as

$$E[X] \equiv \frac{q}{n\sigma_1(\mathbf{d_q})} \sum_{x=1}^{\frac{1}{r(n)}} x \sum_{l=1}^{4x} \sum_{v_k \in s_l} d_k^{-\beta}. \tag{5.78}$$

and, using lemmas 5.3.4 and 5.3.6, and replacing q with $\Theta(f(n))$ proves the second region of Theorem 5.3.1.

Finally, the proof to the third region of Theorem 5.3.6 uses similar reasoning and can be found in [63]. □

Proof to Theorem 5.3.2. Since each node can receive or transmit just one flow at a time, the maximum rate a node (and a cell) can support is $\Theta(1)$. Each node carries traffic during transmission, reception, or relaying of the data. The maximum value of this traffic should not exceed the maximum supportable traffic of $\Theta(1)$. We will consider three different scenarios:

- Traffic load of a source node

Each node transmit at maximum rate of λ_{max} which is much less than one for all the obtained capacity regions. It has been shown [66] that there are $\Theta(nr^2(n))$ nodes in each cell which results in maximum generated traffic by each cell as $\Theta(\lambda_{max} nr^2(n))$. Since λ_{max} does not exceed $\Theta(\frac{1}{nr^2(n)})$, then the maximum traffic generated by each cell cannot exceed $\Theta(1)$. Therefore, the traffic generate in transmission mode does not create any bottleneck.

- Traffic load of a relay node

A path of length x-hops consists of exactly x cells in our model. Since we have a total of $\frac{1}{r^2(n)}$ cells, the probability that a cell is selected from a group of x specific cells is equal to $xr^2(n)$. The probability that a source-destination path of length x-hops passes through a specific cell is always less than $xr^2(n)$. Thus, the probability of a source-destination path L_i passing through a specific cell S_0 is

$$\begin{aligned} \Pr(L_i \text{ intersects } S_0) &= \sum_x \Pr(L_i \text{ intersects } S_0 | X_i = x)\Pr(X_i = x) \\ &\leq \sum_x xr^2(n)\Pr(X_i = x), \end{aligned} \tag{5.79}$$

where X_i is the number of hops the path L_i is passing through. Therefore,

$$\Pr(L_i \text{ intersects } S_0) \leq E[X]r^2(n). \tag{5.80}$$

Since we only consider unicast communications, there are at most a total of $\Theta(n)$ source-destination pairs. Therefore, using the union bound, the maximum number of paths intersecting a specific cell is $\Theta(nE[X]r^2(n))$. Consequently, the maximum traffic load of a relay cell is $\Theta(nE[X]r^2(n)\lambda_{max})$ which is $\Theta(1)$ in all regions of the throughput capacity obtained in this work. Therefore no cell will carry more than what it can support when it is in relay mode.

A relay node in a cell consisting of $\Theta(nr^2(n))$ nodes is selected with a uniform distribution. Hence, the probability that a specific node is a relay equals the probability that the corresponding cell is a relay, divided by the number of nodes in that cell. This probability is smaller than $\Theta(E[X]\lambda_{max})$ which is less than $\Theta(1)$. It is concluded that the relay nodes will never cause bottleneck in the network.

- Traffic load of a destination node

Similar to previous section argument, we conclude that receiver cells do not cause bottleneck in the network. Since the selection of friends for each node follows power-law distribution that may make the distribution of the destination nodes non-uniform. In case of $q = \Theta(1)$, each node has only $q = \Theta(1)$ social contacts and it consumes a constant bandwidth and does not cause bottleneck. For $q = \Theta(n)$, we prove that this distribution is still uniform for large n and similar to the relay nodes, the destination nodes does not create any bottleneck.

The source nodes are uniformly distributed in the network. Thus the probability that a specific node v_k is the destination can be written as

$$\begin{aligned}\Pr(\vartheta_t = v_k) &= \sum_{i=1}^{n} \Pr(\vartheta_t = v_k | v_i \text{ is source}) \Pr(v_i \ \text{ is source}) \\ &= \frac{1}{n} \sum_{i=1}^{n} \Pr(\vartheta_t = v_k | v_i \text{ is source}). \end{aligned} \tag{5.81}$$

Let d_{ki} be the distance between v_k and v_i and G_i be the set of social contacts if node v_i is the source. Let's define $\mathbf{d_{q_i}} = (d_{g_1 i}^{-\beta}, ..., d_{g_q i}^{-\beta})$ and $\mathbf{d_{n_i}} = (d_{g_1 i}^{-\alpha}, ..., d_{g_n i}^{-\alpha})$. Now, similar to equation (3.12) which has been written for one specific source node, we have

$$\begin{aligned}\Pr(\vartheta_t = v_k | v_i \text{ is source}) &= \Pr(\vartheta_t = v_k | v_i \text{ is source}, v_k \in G_i) \Pr(v_k \in G_i) \\ &= \frac{d_{ki}^{-\beta}}{\sigma_1(\mathbf{d_{q_i}})} \frac{d_{ki}^{-\alpha} \sigma_{q-1}(\mathbf{d_{n_i}^{\bar{k}}})}{\sigma_q(\mathbf{d_{n_i}})}. \end{aligned} \tag{5.82}$$

By using lemma 5.3.3 we arrive at

$$\Pr(\vartheta_t = v_k | v_i \text{ is source}) \equiv \frac{d_{ki}^{-\beta}}{\sigma_1(\mathbf{d_{q_i}})}. \tag{5.83}$$

Therefore,

$$\Pr(\vartheta_t = v_k) = \frac{1}{n} \sum_{i=1}^{n} \Pr(\vartheta_t = v_k | v_i \text{ is source}) = \frac{1}{n} \sum_{i=1}^{n} \frac{d_{ki}^{-\beta}}{\sigma_1(\mathbf{d_{q_i}})} = \frac{1}{n}. \tag{5.84}$$

So the destinations are distributed uniformly similar to the relay nodes, and no node in receive mode will be a bottleneck. Notice that since for the case of $q = \Theta(n)$ no node will become bottleneck, for the case of $q = \Theta(f(n))$ also no node will become bottleneck when $f(n) = O(n)$ as in our case.

□

5.4 Power-Law Group Size and Peer Selection

In this section we study the impact of the combination of all three power law distributions on the network performance; the social network formation with parameter α for selecting the long range contacts, parameter γ for the number of long range contacts, and the communication among the members of the social group with parameter β.

5.4.1 Results and Discussion

Theorem 5.4.1. *Consider a social wireless network consisting of n connected nodes with the following properties.*

- *Any two nodes in distance d away from each other are socially connected with a probability inversely proportional to d^{α}, where α is the social group density.*
- *Each node has $q = 1, 2, .., $ or $n - 1$ social contacts and the number of nodes with q social contacts is inversely proportional to q^{γ}, where γ is the social degree distribution exponent.*
- *Each source communicates with one of its social contacts randomly with a probability inversely proportional to d^{β}, where β is the social communication density.*

Under these conditions the throughput capacity will be

$$\lambda_{max} = \begin{cases} \Theta(\frac{1}{nr^{\beta+1}(n)}), & for\ 0 \leq \beta \leq 1 \\ \Theta(\frac{1}{nr^{2}(n)}), & for\ 1 \leq \beta \end{cases}$$

Theorem 5.4.1 implies that in a wireless network where each node has a power-law distributed number of contacts picking based on a power-law distributed distance, and where the closer contacts have more opportunity to be contacted to, the social communication density (β) is the dominant social concept, and for large enough β the network can support the most possible throughput capacity which is $\Theta(\frac{1}{\log n})$ for a connected network.

5.4.2 Proofs to Theorems

Proof to Theorem 5.4.1. The analysis in the proof of Theorem 5.2.1 and 5.3.1 can be easily modified to get the results. Using equations (4.5), (3.12), and (5.46) we have

$$\begin{aligned} E[X] &= \sum_{q=1}^{n} \Pr(\mathrm{Q}=\mathrm{q}) E[X|Q=q] \\ &\equiv \sum_{x=1}^{\frac{1}{r(n)}} x \sum_{l=1}^{4x} \sum_{v_k \in s_l} \sum_{q=1}^{n} \frac{q^{-\gamma} d_k^{-\alpha-\beta} \sigma_{q-1}(\mathbf{d}_{\mathbf{n}}^{\bar{\mathbf{k}}})}{\sigma_1(\mathbf{b})\sigma_1(\mathbf{d_q})\sigma_q(\mathbf{d_n})} \end{aligned} \tag{5.85}$$

To simplify the equation (5.85), like the process in the proof of theorem 5.2.1, we break $E[X]$ into the following two parts,

$$E_1 \equiv \sum_{x=1}^{\frac{1}{r(n)}} x \sum_{l=1}^{4x} \sum_{v_k \in s_l} \sum_{q=q_0}^{n-1} \frac{q^{-\gamma} d_k^{-\alpha-\beta} \sigma_{q-1}(\mathbf{d}_{\mathbf{n}}^{\bar{\mathbf{k}}})}{\sigma_1(\mathbf{b})\sigma_1(\mathbf{d_q})\sigma_q(\mathbf{d_n})} \tag{5.86}$$

and

$$E_2 \equiv \sum_{x=1}^{\frac{1}{r(n)}} x \sum_{l=1}^{4x} \sum_{v_k \in s_l} \sum_{q=1}^{q_0-1} \frac{q^{-\gamma} d_k^{-\alpha-\beta} \sigma_{q-1}(\mathbf{d}_{\mathbf{n}}^{\bar{\mathbf{k}}})}{\sigma_1(\mathbf{b})\sigma_1(\mathbf{d_q})\sigma_q(\mathbf{d_n})} \tag{5.87}$$

We can use the argument in the proof of theorem 5.2.1 to simplify E_1 as

$$\begin{aligned} E_1 &\equiv \sum_{x=1}^{\frac{1}{r(n)}} x \sum_{l=1}^{4x} \sum_{v_k \in s_l} \sum_{q=q_0}^{n} \frac{q^{-\gamma} d_k^{-\beta}}{\sigma_1(\mathbf{b})\sigma_1(\mathbf{d_q})} \frac{q}{n} \\ &\equiv \sum_{x=1}^{\frac{1}{r(n)}} x \sum_{l=1}^{4x} \sum_{v_k \in s_l} \frac{d_k^{-\beta}}{n\sigma_1(\mathbf{b})} \sum_{q=q_0}^{n} \frac{q^{-\gamma+1}}{\sigma_1(\mathbf{d_q})} \end{aligned} \tag{5.88}$$

Since q_0 is a very large number, law of large numbers ensures that $\frac{1}{q}\sigma_1(\mathbf{d_q})$ lies in the interval $(E[\mathbf{d_q}] - \epsilon, E[\mathbf{d_q}] + \epsilon)$ with probability one thus it can be replaced by $E[\mathbf{d_q}]$ in our work.

$$\sum_{q=q_0}^{n} \frac{q^{-\gamma+1}}{\sigma_1(\mathbf{d_q})} = \frac{1}{E[\mathbf{d_q}]} \sum_{q=q_0}^{n} q^{-\gamma} \tag{5.89}$$

To find $E[\mathbf{d_q}]$ notice that according to the proof of lemma 5.3.6 we know that with probability close to one when n approaches infinity, there exists a long-range social contact within the lattice distance of $\Theta(1)$ from the source thus

$$E[\mathbf{d_q}] \equiv \sum_{x=1}^{\frac{1}{r(n)}} \Pr(X = x)(xr(n))^{-\beta} \equiv (r(n))^{-\beta} \tag{5.90}$$

Now if $\gamma > 1$, we have $\sum_{q=q_0}^{n} q^{-\gamma} \leq \sigma_{1,n}(\mathbf{b}^{-\gamma}) \leq \sum_{q=1}^{\infty} q^{-\gamma} = \zeta(\gamma) \equiv \Theta(1)$. Therefore (5.89) can be simplified to

$$\sum_{q=q_0}^{n} \frac{q^{-\gamma+1}}{\sigma_1(\mathbf{d_q})} = (r(n))^{\beta} \tag{5.91}$$

and if $0 \leq \gamma \leq 1$ we have $\sum_{q=q_0}^{n} q^{-\gamma} \equiv \sigma_{1,n}(\mathbf{b}^{-\gamma}) \equiv \dfrac{n^{-\gamma+1}}{-\gamma+1} \equiv n^{-\gamma+1}$. Thus in this case (5.89) simplifies to

$$\sum_{q=q_0}^{n} \frac{q^{-\gamma+1}}{\sigma_1(\mathbf{d_q})} \equiv n^{-\gamma+1}(r(n))^{\beta} \tag{5.92}$$

Using the previous equations of (5.74) and (5.75)

$$\frac{1}{n\sigma_1(\mathbf{b})}\sum_{x=1}^{\frac{1}{r(n)}} x\sum_{l=1}^{4x}\sum_{v_k\in s_l} d_k^{-\beta}$$

$$\equiv \begin{cases} \Theta\left(n^{\gamma-1}r(n)^{-1}\right), & 0\leq\beta\leq 3,\ 0\leq\gamma\leq 1 \\ \Theta\left(n^{\gamma-1}r(n)^{2-\beta}\right), & 3\leq\beta, \quad\quad 0\leq\gamma\leq 1 \\ \Theta\left(r(n)^{-1}\right), & 0\leq\beta\leq 3, \quad \gamma>1 \\ \Theta\left(r(n)^{2-\beta}\right), & 3\leq\beta, \quad\quad \gamma>1 \end{cases} \tag{5.93}$$

Therefore using (5.91), (5.92) and (5.93) we have

$$\begin{aligned} E_1 &\equiv \begin{cases} \Theta\left(r(n)^{-1+\beta}\right), & 0\leq\beta\leq 3 \\ \Theta\left(r(n)^{2}\right), & 3\leq\beta \end{cases} \\ &\equiv \begin{cases} \Theta\left(r(n)^{-1+\beta}\right), & 0\leq\beta\leq 1 \\ \Theta\left(1\right), & 1\leq\beta \end{cases} \end{aligned} \tag{5.94}$$

Notice that since $E[X]$ cannot be smaller than one, thus we can replace $r(n)^{-1+\beta}$ for $1\leq\beta\leq 3$, and $r(n)^2$ with 1, thus, the second equality holds. Now we use lemma 5.3.7 and equation (5.77) to prove that the order of E_1 is dominant in the summation $E[X]=E_1+E_2$. Using the right hand side of (5.77) we have

$$E_2 \ \leq \ \sum_{x=1}^{\frac{1}{r(n)}} x\sum_{l=1}^{4x}\sum_{v_k\in s_l}\sum_{q=1}^{q_0-1}\frac{q^{-\gamma}d_k^{-\alpha-\beta}\sigma_{q-1}(\mathbf{d_n})}{\sigma_1(\mathbf{b})\sigma_1(\mathbf{d_q})\sigma_q(\mathbf{d_n})} \tag{5.95}$$

Since $q\leq q_0$, it is a finite number and we can use lemma A.4.1 to get

$$\frac{\sigma_{q-1}(\mathbf{d_n})}{\sigma_q(\mathbf{d_n})}\equiv\frac{1}{\sigma_1(\mathbf{d_n})}\Theta(\frac{nq}{n-q+1})\equiv\frac{1}{\sigma_1(\mathbf{d_n})} \tag{5.96}$$

Thus

$$E_2 \leq \sum_{x=1}^{\frac{1}{r(n)}} x \sum_{l=1}^{4x} \sum_{v_k \in s_l} \frac{d_k^{-\alpha-\beta}}{\sigma_1(\mathbf{d_n})\sigma_1(\mathbf{b})} \sum_{q=1}^{q_0-1} \frac{q^{-\gamma}}{\sigma_1(\mathbf{d_q})}. \tag{5.97}$$

Notice that using the argument in the proof of lemma 5.3.6 for very large n, there exists a long-range contact in the lattice distance of $\Theta(1)$ to the source, with high probability, which will be the dominant term in the summation $\sigma_1(\mathbf{d_q})$ thus $\sigma_1(\mathbf{d_q})$ scales as $r(n)^{-\beta}$ and hence,

$$\sum_{q=1}^{q_0-1} \frac{q^{-\gamma}}{\sigma_1(\mathbf{d_q})} \equiv \frac{1}{r(n)^{-\beta}} \sum_{q=1}^{q_0-1} q^{-\gamma} \equiv r(n)^{\beta} \tag{5.98}$$

Therefore, $E_2 \leq \dfrac{r(n)^{\beta}}{\sigma_1(\mathbf{d_n})\sigma_1(\mathbf{b})} \sum_{x=1}^{\frac{1}{r(n)}} x \sum_{l=1}^{4x} \sum_{v_k \in s_l} d_k^{-\alpha-\beta}$. Thus for $\gamma > 1$ we have

$$E_2 \equiv \begin{cases} O\left(r(n)^{-1+\beta}\right), & 0 \leq \alpha+\beta \leq 3, 0 \leq \alpha \leq 2 \\ O\left(r(n)^{\alpha+\beta-3}\right), & 0 \leq \alpha+\beta \leq 3, \alpha \geq 2 \\ O\left(r(n)^{2-\alpha}\right), & 3 \leq \alpha+\beta, 0 \leq \alpha \leq 2 \\ O\left(1\right), & 3 \leq \alpha+\beta, \alpha \geq 2 \end{cases} \tag{5.99}$$

and for $0 \leq \gamma \leq 1$, E_2 will have a scaling factor of $n^{1-\gamma}$ multiplied by the above equation. It can be verified that E_1 is the dominant term compared to E_2 and therefore using equation (4.1) theorem 5.4.1 is proved. □

Chapter 6

Conclusion and Future View

In this part of our research, we comprehensively studied the effects of social interactions among nodes on the capacity of wireless networks. We considered three power-law distributed characteristics of social networks and added them one by one to study their impact on the network performance. The three considered social characteristics are distance between members of each social group, distance between communication pairs inside each group, and the size of social groups.

Through further investigation we revealed that traditional transport capacity definition provides misleading conclusions for such network models. We showed that nodes with different social status impact the capacity differently. By categorizing nodes based on their different social status and allocating separate bandwidth to them, it was shown that majority of nodes scale in this network.

Our simulation results corroborated the analytical results. Further, we observed consistently that social interaction improves the capacity of wireless networks which implies that the Gupta-Kumar results were pessimistic for real networks.

In this work we have made many assumptions to simplify our analytical framework. For example, we have assumed that each source unicasts with a single destination in its social group, that the protocol model is used to model MAI, and that all radios are similar. In addition, we have not addressed the role of content popularity or common interest in content within social groups. Relaxing these assumptions can be one line of the works that can be done in future.

It is also worth emphasizing that the effects of social group evolution is not considered in our network model and a more comprehensive work, should consider such effects in the study of wireless networks with social considerations. For future work, proper protocols for these wireless social networks can be studied, different resource allocations based on social status can also be investigated to name a few.

Quality of Service is another important point that can be considered for future research. We studied the data transferred in networks with social users, and to improve capacity we categorized the users and assigned the bandwidth based on their social status. However, the data which is going to be transferred in the network may have different levels of importance. Assigning bandwidth just based on the social status may make barrier for more critical data. Working on some ways that prioritizes internet traffic for vital applications can be considered in future as well.

Part II

Network of Caches

Chapter 7

Introduction

7.1 Research Motivations

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, 59, 112].

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. 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.

Further, a higher level look can abstract a communication network into two (logical) layers, namely, a control plane carrying signaling and administrative traffic, and a data forwarding plane carrying the user data traffic. In many applications, for the network to function properly, the control plane must have some knowledge about the forwarding plane in order to create a view of the underlying network. The underlying network will be in an operating state which is reported by a protocol to the control/management layer. For example, in a network of caches described above, the data plane contains caches keeping the data traffic, e.g. video or audio files, which are requested and used by the users, and the information regarding the items kept in each cache reported to the control layer forms the control traffic.

However, as the networks have grown in size and complexity, as end nodes, content and virtual machines move about, it will become more difficult for the control layer to have an accurate view of the forwarding plane. Consider the example of finding a service or a piece of content. Current protocols attempt to resolve a content request to the nearest copy of the object by using DNS or redirecting HTTP requests. Further proposals suggest to share content location information in between content delivery networks (CDNs), or even to build content routing within the architecture. In all cases, this implicitly entails that the mechanism responsible to route to the content has to be dynamically updated with the content location. Meta-information from the forwarding plane needs to be delivered to the control plane. This raises the question: *how much?*

In other words, depending on the size of the domain being controlled, of the underlying state space, of the dynamics of the evolution of the state in the forwarding plane, what stream of data is required to keep the control plane up to date?

7.2 Contributions

The present research investigates the asymptotic orders of access time and throughput capacity in such networks of caches. 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 ([16]).

We also consider the issue of maintaining a consistent view of the underlying state at the control layer, and develop an abstracted mechanism, which can be applied to a wide range of scenarios. We assume the underlying state as an evolving random process, and calculate the rate that this process would create to keep the representation of this state up-to-date in the control plane. This provides a lower bound on the overhead bandwidth required for the control plane to have an accurate view of the forwarding plane[1].

We then illustrate the power of our model by focusing on the specific case of locating content in a resolution-based content-oriented network. Enabling content routing has attracted a lot of attention recently, and thus we are able to shed some light on its feasibility. In this case, the underlying state depends on the size and number of caches, on the request for content process and on the caching policy. We apply our framework to derive the bandwidth needed to accurately locate a specific piece of content. We observe that there is a trade-off for keeping an up-to-date view of the network at the cost of significant bandwidth utilization, versus the gain achieved by fetching the nearest copy of the content. We consider a simple scenario to illustrate this trade-off.

Our contribution is as follows ([11, 16, 17]):

- We first state our results on fundamental limits of throughput capacity and latency in network of caches and quantify the performance improvement brought about by a content-centric network architecture over networks with no content sharing

[1] There exists other overhead that we have not discussed in this work. We believe that addressing all the overhead of data/control plane interface in one work may not be possible, since there might be several sources for them. However, if one thinks of certain sources of overhead, like the overhead of setting up a secure connection between the forward and control planes, then the actual protocol overhead would be proportional to the information theoretic overhead, at least if the rate of update is high enough. In which case we provide a good idea of how the whole protocol overhead will trend.

capability.

- We investigate content discovery mechanism effect on the performance. More specifically, we compare the performance improvement by selecting the nearest copy of the content and selecting the nearest copy in the direction of original server.
- We study the impact of the caching policy, and in particular, the length of time each piece of content spends in the cache's memory, on the performance.
- We then present a framework to quantify the minimal amount of information required to keep a (logical) control plane aware of the state of the forwarding plane. We believe this framework to be useful in many distributed systems contexts.
- We apply our framework to the specific case of locating content, and see how content location is affected by the availability of caches, the caching policy and the content popularity. We can thus apply our results to some of the content-oriented architectures and observe that cached copies would go ignored for a large swath of the content set.
- We see how our framework allows to define some optimal policies with respect to the contents that should be cached for an operator-driven content distribution system. While it is not surprising that very unpopular contents should not be cached, we can actually compute a penalty for doing so under our model.

We quickly note that our framework does not debate the merit of centralized vs distributed, as the control layer we consider could be either. For a routing example, our model would provide a lower-bound estimate of the bandwidth for, say OpenFlow to update a centralized SDN controller, or for a BGP-like mechanism to update distributed routing instances.

Our results are theoretic in nature, and provide a lower bound on the overhead. We hope they will provide a practical guideline for protocol designers to optimize the protocols which synchronize the network state and the control plane ([17]).

7.3 Outline

Part II of this research is organized as follows. After a brief review of the related work in Chapter 8, the network models, the content discovery algorithms used in the current work, and the content distribution in steady-state are introduced in Chapter 9. The main theorems on fundamental limits of performance metrics are stated and proved in Chapter 10. We discuss the results and study some simple examples in Chapter 11. We then introduce our framework to model the protocol overhead and study the content location in the network of caches in Chapter 12. The derived model is used to study a simple caching network as well. We show the power of the model in the protocol design by computing the cost of content routing and suggesting a cache management policy. Finally this part is concluded and some possible directions for the future work will be introduced in Chapter 13.

Chapter 8

Previous 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 [59], NetInf [5], PURSUIT [1], or DONA [65], where nodes can cache content, and requests for content can be routed to the nearest copy. Papers surveying the landscape of ICN [4] [50] show the dearth of theoretical results underlying these architectures.

Caching, one of the main concepts in ICN networks, has been studied in prior works [4]. [85] 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 [34, 91], and the benefit of cooperative caching has been investigated in [107].

Optimal cache locations [90], cach sizes [12], and cache replacement techniques [110] are other aspects most commonly investigated. The work in [92] 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 [98]. [105] 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. [41] 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 [93], and also study the performance of CCN with emphasis on the size of individual caches [94].

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 [55] and many subsequent works [70] [84]. Some work has studied the capacity of wireless networks with caching [54] [57] [8] . 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. [75] uses a network simulation model and evaluates the performance (file transfer delay) in a cache-and-forward system with no request for the data. [32] 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 [51] for a static cache placement in a multihop wireless network.

[83] 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 [76] 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 [60,61,73]. This is important to note that our current work is different from [60,61,73,76,83] 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 our work [11] 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 [19,25]. In this work, we do not assume any specific popularity distribution and present the results for any arbitrary request pattern.

On the other hand, as SDN makes the separation explicit between the control and forwarding layers, it begs the question of how these layers interact. This interaction has been pointed out as one of the bottlenecks of OpenFlow [78], and several papers have been trying to optimize the performance of the traffic going from one layer to the other. For instance, [97] optimizes the controller to support more traffic, while [37] or [111] attempt to make the control layer more distributed and thus reduce the amount of interaction between the switches and the control layer. There has been no attempt to model the interaction between the control and forwarding layers to our knowledge.

Studying the gap between the state of the system and the view of the controller,

[69] focuses on the relationship between performance and state consistency, and [22] studies similar relationship in multiple controller systems. This underlines the need for the view at the control layer to be representing the network state with as little distortion as possible.

The forwarding plane in a network usually consists of a state machine which is changing because of different network characteristics. The control plane needs to obtain adequate information about the underlying states so that the network can perform within a satisfactory range of distortion. The first theoretical study of this information was conducted by Gallager in [49]. This work utilizes the rate distortion theory to calculate the bounds on the information required to show some characteristics such as the start time and the length of the messages.

The link states (validity of a link) and the geographic location and velocity of each node in a mobile wireless network are some examples of such state, which have been studied in [100] and [101], respectively. An information-theoretic framework to model the relationship between network information and network performance, and the minimum quantity of information required for a given network performance was derived in [58].

One impetus to study the relationship between the control layer and the network layer comes from the increased network state complexity from trying to route directly to content. Request-routing mechanisms have been in place for a while [21] and proposals [88] have been suggested to share information between different CDNs, in essence enabling the control planes of two domains to interact (our framework applies to this situation). And many architectures have been proposed that are oriented around content [4,53,59,65,99,112] and some have raised concerns about the scalability of properly identifying the location of up to 10^{15} pieces of content [50]. Our model presents a mathematical foundation to study the pros and cons of such architectures.

The cache management problem in the networks has been studied in several contexts. [96] presents a centralized approximation algorithm to solve the cache placement problem for minimizing the total data access cost in ad hoc networks. [26] proposes a replication algorithm that lets nodes autonomously decide on caching the information, and [12] determines whether/where to keep a copy of a content such that the overall cost of content delivery is minimized and show that such optimized content delivery significantly reduces the cost of content distribution and improves quality of service.

Some cooperative cache management algorithms are developed in [29] which tries to maximize the traffic volume served from cache and minimize the bandwidth cost in content distribution networks. [95] proposes some online cache management algorithms for Information Centric Networks (ICNs) where all the contents are available by caching in the network instead of a server or original publisher. [33] investigates if caching only in a subset of node(s) along the content delivery path in ICNs can achieve better performance in terms of cache and server hit rates. These works define a specific cost in the network and try to determine the locations and the number of copies of the contents in the network such that the defined cost is minimized. Finally [11] and [16] analytically prove that on-path content discovery has the same asymptotic capacity as finding the nearest copy in these networks.

To the best of our knowledge, there is no work considering the protocol overhead in such systems. In this work, we model the protocol overhead, then use that model to compute a general cost for data retrieval (including the protocol overhead). We also investigate whether allowing more copies of the contents cached in the network reduces the total cost. One related work on this topic is [36] which proposes a content caching scheme, in which the number of chunks (fragments) to be cached in each storage is adjusted based on the popularity of the content. In this work, each upstream node recommends the number of chunks to be cached in the downstream node according to the number of requests.

Chapter 9

Network of Caches

9.1 Network Model

Two network models are studied in this part.

9.1.1 Grid Network

Assume that the network consists of n nodes $\{v_1, v_2, ..., v_n\}$ 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 9.1(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, $\{f_1, ..., f_m\}$ with sizes $\{B_1, ..., B_m\}$, for which

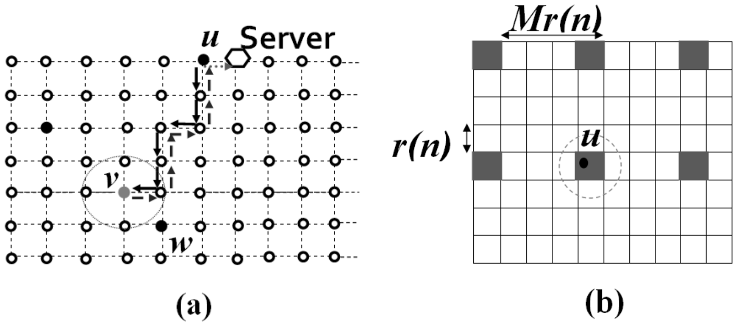

Figure 9.1: **Network models** a) Grid network: the transmission range of node v contains four surrounding nodes. The black vertices contain the content in their local caches. The arrow lines demonstrate a possible discovery and receive path in grid network with path search, where node v downloads the required information from u. In grid network with ring search, 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.

each node v_j may issue a query with probabilities $\{\alpha_k,\ k = 1, ..., m\}$, where $\sum_{k=1}^{m} \alpha_k = 1$, and m and α_k are not changing with the network size[1]. Based on the 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 this work 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

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

model and according to [108], the transport capacity in such network is upper bounded by $\Theta(W\sqrt{n})$. This is the model studied in Theorem 10.1.1 and the first two scenarios of Theorem 10.1.2.

9.1.2 Random Network

The next network studied in Theorem 10.1.2 is a more general network model where the nodes are randomly distributed over a unit square area according to a uniform distribution (Figure 9.1(b)). We use the same model used in [108] (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 [87] and a non-zero capacity. According to the protocol model [108], 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.

9.2 Content Discovery Algorithm

9.2.1 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.

9.2.2 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. 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.

9.3 Content Distribution in Steady-State

The time diagram of data access process in a cache is illustrated in Figure 9.2. 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.

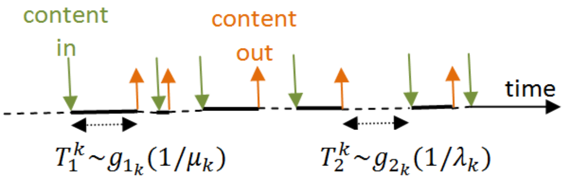
content in
content out
time
$T_1^k \sim g_{1_k}(1/\mu_k)$
$T_2^k \sim g_{2_k}(1/\lambda_k)$

Figure 9.2: **Data access process time diagram** in a cache for content k

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

$$\rho^{(k)}(n) = \frac{1/\mu_k}{1/\mu_k + 1/\lambda_k} = \frac{\lambda_k}{\lambda_k + \mu_k}, \tag{9.1}$$

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.

9.4 Performance Indices

The performance indices studied in this part are:

9.4.1 Throughput Capacity

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

9.4.2 Average Latency

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

9.4.3 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.

Chapter 10

Throughput Capacity in Networks of Caches

Our results on the asymptotic orders are stated in three sections; Section 10.1 formulates the capacity in a grid network which uses the shortest path to the server content discovery mechanism, Section 10.2 derives the capacity results for a grid network with expanding ring content discovery method. Section 10.3 formulates the capacity in a random network which uses the shortest path discovery mechanism. The theorems stated in these sections demonstrate that adding the content sharing capability to the nodes can significantly increase the capacity and gives us some ideas how the content search mechanism can affect the performance improvement.

10.1 Grid Network/Path-Wise Content Discovery

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

$$\gamma_{max} \equiv \frac{n}{\sum_{k=1}^{m} \alpha_k \sum_{i=1}^{\sqrt{n}} 4i \sum_{j=0}^{i-1} (i-j) \rho_j^{(k)}(n) \prod_{l=j+1}^{i} (1-\rho_l^{(k)}(n))}, \tag{10.1}$$

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 [32, 38, 47, 77, 89] to name a few.

[1] 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 work.

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, $1 \leq j \leq i$ 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 $P_{i,j}^{(k)} =$

$$(1 - \rho_i^{(k)}(n))(1 - \rho_{i-1}^{(k)}(n))...(1 - \rho_{j+1}^{(k)}(n))\rho_j^{(k)}(n) \tag{10.2}$$

where $\rho_j^{(k)}(n)$ is the probability of content k being available in a cache in level $j, 1 \leq j \leq \sqrt{n}$, 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

$$E[h_k] = \tfrac{1}{n} \textstyle\sum_{i=1}^{\sqrt{n}} 4i \sum_{j=0}^{i-1} (i - j) P_{i,j}^{(k)} \tag{10.3}$$

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[2,3], $\rho^{(k)}(n) \nrightarrow 1$ (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. We restate Theorem 10.1.1 as a new theorem here and derive the average latency of getting a content in such networks.

Theorem 10.1.2. *The average latency order in the above grid network when the nodes use the nearest copy of the required content on the shortest path toward the server is given by*

$$E[h] = \Theta(min(\sqrt{n}, \frac{1}{\underset{k}{min}(\rho^{(k)}(n))})) \tag{10.4}$$

Here we prove Theorem 10.1.2 by utilizing some Lemmas. The proof of lemmas are presented in the Appendix.

Lemma 10.1.3. *Consider the wireless networks described in Theorem 10.1.2. The average number of hops between the customer and the serving node (a cache or original*

[2] 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.

[3] 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.

server) for item k asymptotically equals to

$$\begin{aligned} E[h_k] &\equiv \frac{1}{n}\sum_{i=1}^{\sqrt{n}} i^2(1-\rho^{(k)}(n))^i \\ &+ \frac{\rho^{(k)}(n)}{n}\sum_{i=1}^{\sqrt{n}} i \sum_{l=1}^{i-1} l(1-\rho^{(k)}(n))^l \end{aligned} \tag{10.5}$$

Lemma 10.1.4. *Consider the wireless networks described in Theorem 10.1.2. 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*

$$E[h_k] \equiv \begin{cases} \sqrt{n}, & for\ \rho^{(k)}(n) \preceq \frac{1}{\sqrt{n}} \\ \frac{1}{\rho^{(k)}(n)}, & for\ \rho^{(k)}(n) \succeq \frac{1}{\sqrt{n}}. \end{cases} \tag{10.6}$$

Theorem 10.1.2 is now simply proved using the above Lemmas.

Proof to Theorem 10.1.2. 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 above scenario if $\rho^{(k)}(n) \preceq \frac{1}{\sqrt{n}}$, when there is at least one node with average number of hops equal to $\sqrt{n}$, then that node's $E[h_k]$ in $E[h]$ defined above becomes the dominant factor.

If for all the contents $\rho^{(k)}(n) \succeq \frac{1}{\sqrt{n}}$, then $E[h]$ is given by $\sum_{k=1}^{m} \frac{\alpha_k}{\rho^{(k)}(n)} \equiv \frac{1}{\underset{k}{min}(\rho^{(k)}(n))}$.

The total $E[h]$ can be simply written as the results shown in Theorem 10.1.2.

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 10.1.2 is proved. □

Theorem 10.1.5. *Consider the network of Theorem 10.1.2, 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} is*

$$\gamma_{max} = \Theta(max(\frac{1}{n}, \min_k((\rho^{(k)}(n))^2))). \tag{10.7}$$

To prove Theorem 10.1.5 we use Lemma 10.1.4, and the following two Lemmas.

Lemma 10.1.6. *Consider the wireless networks described in Theorem 10.1.2. In order not to have interference, the maximum throughput capacity is upper limited by $\Theta(max(\frac{1}{\sqrt{n}}, \min_k(\rho^{(k)}(n))))$.*

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 10.1.7. *The maximum supportable throughput capacities in the studied scenario is $\Theta(max(\frac{1}{n}, \min_k((\rho^{(k)}(n))^2)))$.*

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 10.1.6 and 10.1.7, and Theorem 10.1.5 is proved.

10.2 Grid Network/Expanding Ring Content Search

Here we consider the same network studied in Section 10.2 when nodes find the contents by expanding ring search as a content discovery mechanism instead of path search. Theorem 10.2.1 derives the average latency of getting a content in such networks.

Theorem 10.2.1. *The average latency order in the above grid network when the nodes use the nearest copy of the required content through expanding ring search is given by*

$$E[h] = \Theta(min(\sqrt{n}, \frac{1}{\sqrt{\underset{k}{min}(\rho^{(k)}(n))}})). \tag{10.8}$$

Here we prove Theorem 10.2.1 by utilizing some Lemmas. The proof of lemmas are presented in the Appendix.

Lemma 10.2.2. *Consider the wireless networks described in Theorem 10.2.1. The average number of hops between the customer and the serving node (a cache or original server) for item k asymptotically equals to*

$$\begin{aligned} E[h_k] &\equiv \frac{1}{n}\{\sum_{i=1}^{\sqrt{n}} i^2(1-\rho^{(k)}(n))^{2i^2-2i+1} \\ &+ \sum_{i=2}^{\sqrt{n}} i \sum_{l=1}^{i-1} l(1-\rho^{(k)}(n))^{2l^2-2l+1}(1-(1-\rho^{(k)}(n))^{4l})\} \end{aligned} \tag{10.9}$$

Lemma 10.2.3. *Consider the wireless networks described in Theorem 10.2.1. 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*

$$E[h_k] \equiv \begin{cases} \sqrt{n}, & for\ \rho^{(k)}(n) \preceq \frac{1}{n} \\ \frac{1}{\sqrt{\rho^{(k)}(n)}}, & for\ \rho^{(k)}(n) \succeq \frac{1}{n}. \end{cases} \tag{10.10}$$

Theorem 10.2.1 is now simply proved using the above Lemmas.

Proof to Theorem 10.2.1. In the above scenario when $\rho^{(k)}(n) \preceq \frac{1}{n}$, when there is at least one node with average number of hops equal to $\sqrt{n}$, then that node's $E[h_k]$ in $E[h]$ defined above becomes the dominant factor.

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

The total $E[h]$ can be simply written as the results shown in Theorem 10.2.1.

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 10.2.1 is proved. □

Theorem 10.2.4. *Consider the network of Theorem 10.2.1, 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} is*

$$\gamma_{max} = \Theta(max(\frac{1}{n}, \underset{k}{min}(\rho^{(k)}(n)))). \tag{10.11}$$

To prove Theorem 10.2.4 we use Lemma 10.2.3, and the following two Lemmas.

Lemma 10.2.5. *Consider the wireless networks described in Theorem 10.2.1. In order not to have interference, the maximum throughput capacity is upper limited by $\Theta(max(\frac{1}{\sqrt{n}}, \sqrt{\underset{k}{min}(\rho^{(k)}(n))}))$.*

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 10.2.6. *The maximum supportable throughput capacities in the studied scenario is $\Theta(max(\frac{1}{n}, \underset{k}{min}(\rho^{(k)}(n))))$.*

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 10.2.5 and 10.2.6, and Theorem 10.2.4 is proved.

10.3 Random Network/Path-Wise Content Discovery

In this section we consider a random network with nodes using the path search to find the contents. 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 10.3.1 derives the average latency of getting a content in such networks.

Theorem 10.3.1. *The average latency order in the above random network when the nodes use the nearest copy of the required content through path search is given by*

$$E[h] = \Theta(max[1, min(\frac{1}{r(n)}, \frac{1}{\underset{k}{min}(\rho^{(k)}(n))nr^2(n)})]). \tag{10.12}$$

Here we prove Theorem 10.3.1 by utilizing some Lemmas. The proof of lemmas are presented in the Appendix.

Lemma 10.3.2. *Consider the wireless networks described in Theorem 10.3.1. The average number of hops between the customer and the serving node (a cache or original server) for item k asymptotically equals to*

$$\begin{aligned} E[h_k] &\equiv r^2(n)\{\sum_{i=2}^{\frac{1}{r(n)}} i^2(1-\rho^{(k)}(n))^{inr^2(n)} \\ &+ (1-(1-\rho^{(k)}(n))^{nr^2(n)})\sum_{i=2}^{\frac{1}{r(n)}} i \sum_{l=1}^{i-1} l(1-\rho^{(k)}(n))^{lnr^2(n)}\} \end{aligned} \tag{10.13}$$

Lemma 10.3.3. *Consider the wireless networks described in Theorem 10.3.1. 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*

$$E[h_k] \equiv \begin{cases} \frac{1}{r(n)}, & for\ \rho^{(k)}(n) \preceq \frac{1}{nr(n)} \\ \frac{1}{\rho^{(k)}(n)nr^2(n)}, & for\ \frac{1}{nr(n)} \preceq \rho^{(k)}(n) \preceq \frac{1}{nr^2(n)} \\ 1, & for\ \rho^{(k)}(n) \succeq \frac{1}{nr^2(n)}. \end{cases} \tag{10.14}$$

Theorem 10.3.1 is now simply proved using the above Lemmas.

Proof to Theorem 10.3.1. In the above scenario for the case of $\rho^{(k)}(n) \preceq \frac{1}{nr(n)}$, when there is at least one node with average number of hops equal to $\frac{1}{r(n)}$, then that node's $E[h_k]$ in $E[h]$ defined above becomes the dominant factor.

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

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 $E[h] = \sum_{k=1}^{m} \frac{\alpha_k}{\rho^{(k)}(n)nr^2(n)} \equiv \frac{1}{\underset{k}{min}(\rho^{(k)}(n)nr^2(n))}$.

The total $E[h]$ can be simply written as the results shown in Theorem 10.3.1.

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 10.3.1 is proved. □

Theorem 10.3.4. *Consider the network of Theorem 10.3.1, 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} is*

$$\gamma_{max} = \Theta(max[\frac{1}{n}, min(\frac{1}{nr^2(n)}, \underset{k}{min}((\rho^{(k)}(n))^2)nr^2(n))]). \tag{10.15}$$

To prove Theorem 10.3.4 we use Lemma 10.3.3, and the following two Lemmas.

Lemma 10.3.5. *Consider the wireless network described in Theorem 10.3.4. In order not to have interference, the maximum throughput capacity is upper limited by* $\Theta(min[\frac{1}{nr^2(n)}, max(\frac{1}{nr(n)}, \underset{k}{min}(\rho^{(k)}(n)))])$.

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 10.3.6. *The maximum supportable throughput capacities in the studied scenario is* $\Theta(max[\frac{1}{n}, min(\frac{1}{nr^2(n)}, \underset{k}{min}((\rho^{(k)}(n))^2)nr^2(n))])$.

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 10.3.5 and 10.3.6, and Theorem 10.3.4 is proved.

Chapter 11

Sample Case Results and 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 [46] 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.

11.1 Sample Case 1

11.1.1 Network Model

Consider a network where the received data is stored only at the receivers (edge caching [23, 52]) 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 (9.1) $\mu_k = \eta_k$, 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 (9.1) $\lambda_k = \beta_k$. 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).

11.1.2 Results

Figures 11.1 (a),(b) respectively illustrate the total request rate and the total traffic generated in a fixed size network in Section 10.1 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 11.2 (a),(b) respectively illustrate the total request rate and the total traffic generated in a fixed size network in Section 10.1 for different time-out rates when

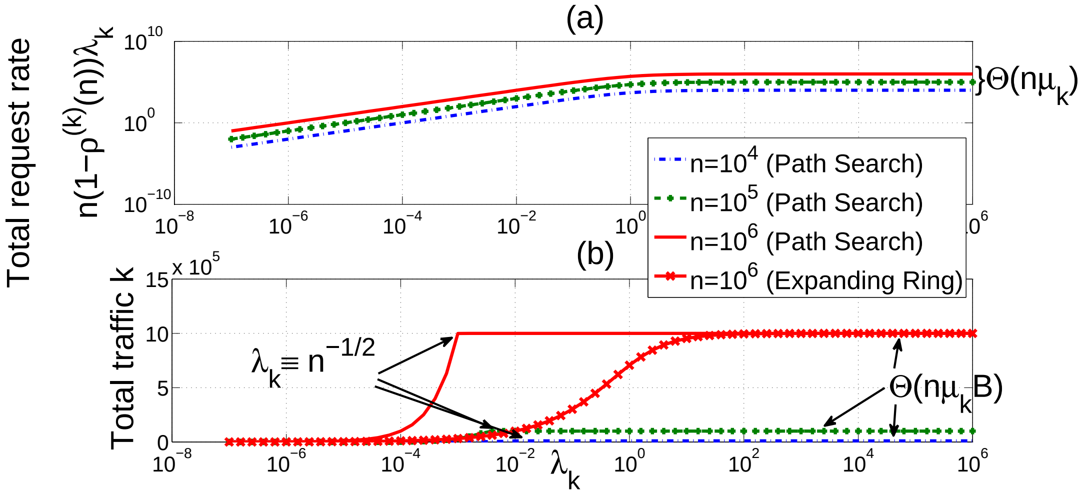

Figure 11.1: **Grid Network/Path Search results** (a) Total request rate for an item k in the network ($\lambda_k n(1-\rho^{(k)}(n))$), (b) Total traffic in the network ($B_k\lambda_k n(1-\rho^{(k)}(n))E[h_k]$) vs. the request rate ($\lambda_k$) with fixed time-out rate ($\mu_k = 1$).

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 11.3 we assume that the request rate is roughly 7 times the drop

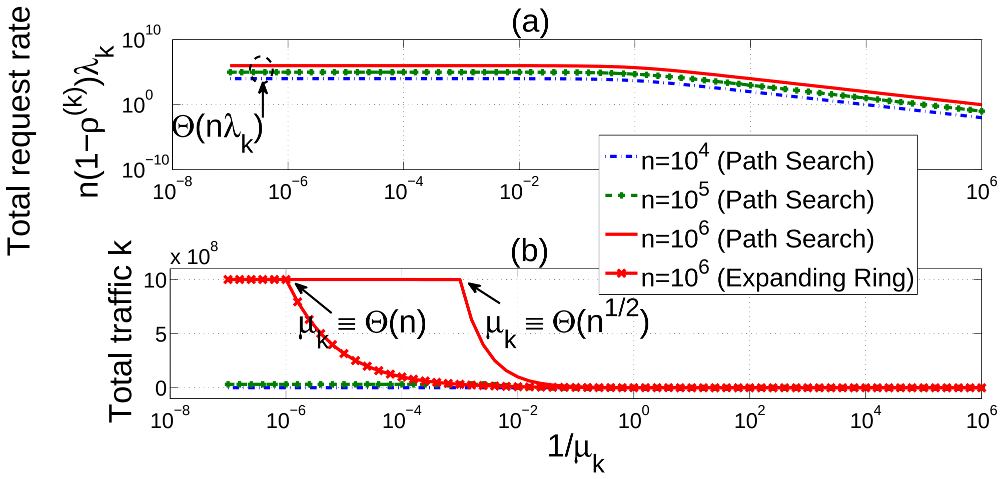

Figure 11.2: **Grid Network/Path Search results** (a) Total request rate in the network ($\lambda_k n(1-\rho^{(k)}(n))$), (b) Total traffic in the network ($B_k\lambda_k n(1-\rho^{(k)}(n))E[h_k]$) vs. the inverse of the time-out rate ($1/\mu_k$) with fixed request ratio ($\lambda_k = 1$).

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 Section 10.3, 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 10.2.1 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 results of sections 10.1 and 10.2, 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.

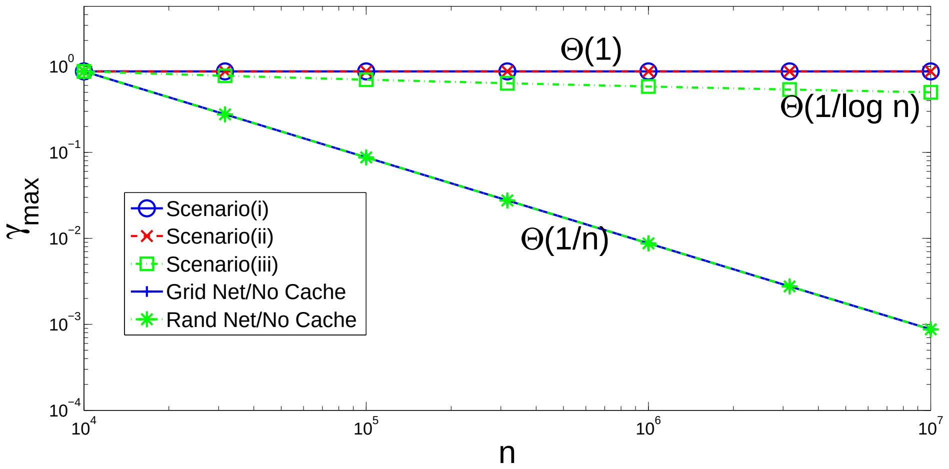

Figure 11.3: **Maximum download rate (γ_{max}) vs. number of nodes (n)** for $\rho = 7/8$.

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 of grid network with path search, grid network with ring search, and random network with path search, 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 case of random network with path search) 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.

11.2 Sample Case 2

11.2.1 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 [35], 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 10.1.1 to obtain the throughput capacity.

11.2.2 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

$$t_{on}^{(k)i}(T) = \sum_{j=0}^{N^{(k)i}(T)} min(\tau_i^{req(k)}, D_i(n)), \tag{11.1}$$

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

$$\begin{aligned} & \sum_{l=0}^{\infty} E[\sum_{j=0}^{l} min(\tau_i^{req(k)}, D_i(n))]Pr(N^{(k)i}(T) = l), \\ = & \sum_{l=0}^{\infty} lE[min(\tau_i^{req(k)}, D_i(n))]Pr(N^{(k)i}(T) = l), \\ = & E[min(\tau_i^{req(k)}, D_i(n))]E[N^{(k)i}(T)]. \end{aligned} \tag{11.2}$$

According to the Poisson arrivals of requests (data downloads) with parameter $\beta^{(k)} + \beta_i^{(k)'}$, the rightmost term in equation (11.2) $(E[N^{(k)i}(T)])$ 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

$$\frac{1}{\sum_{k=1}^{m} \frac{\alpha_k e^{-D\beta^{(k)}}}{1-e^{-D\beta^{(k)}}}} \equiv 1. \tag{11.3}$$

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

$$\beta_i^{(k)'} = \frac{(1-\rho_{i+1}^{(k)})(\beta^{(k)} + \beta_{i+1}^{(k)'})(i+1)}{i} \tag{11.4}$$

Combining equation (11.4), 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 10.1.1 we can obtain the capacity (γ_{max}) in the grid network with path-wise content discovery and on-path storing scheme which is

$$\frac{n}{\sum_{k=1}^{m} \alpha_k \sum_{i=1}^{\sqrt{n}} i \sum_{j=0}^{i-1} (i-j)(1-e^{-D_j(n)(\beta^{(k)}+\beta_j^{(k)'})})e^{-\sum_{l=j+1}^{i} D_l(n)(\beta^{(k)}+\beta_l^{(k)'})}}. \tag{11.5}$$

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.

Chapter 12

Content Discovery Control Traffic

12.1 Protocol Overhead Model

In this section we turn our attention to the mechanism to synchronize the view at the control layer with the underlying network state, and introduce a framework to quantify the minimal amount of required transferred information.

Assume that $S_X(t)$ describes the state of random process X in a network at time t. In order to update the control plane's information about the states of X in the network, the forwarding plane must send update packets regarding those states to the control plane whenever some change occurs. Let $\hat{S}_X(t)$ denote the control plane's perceived state of X at time t. It is obvious that no change in $\hat{S}_X$ will happen before S_X changes, and if S_X changes, the control plane may or may not be notified of that change. Therefore, there are some instances of time where $\hat{S}_X \neq S_X$.

In this work, we consider, systems and applications in which the state can have two values $'0'$ and $'1'$.[1] For instance, a link can be up or down; or a piece of content can be present at a node, or not. Figure 12.1 illustrates the time diagram of state changes of such binary random process which is the state of the forwarding plane in the network being announced to the control plane.

Let $\{Y_m\}_{m=1}^{\infty}$ and $\{Z_m\}_{m=1}^{\infty}$ denote the sequences of $'0'$s and $'1'$s time durations of $S_X(t)$ respectively, and $\{T_m\}_{m=1}^{\infty}$ denote the times of changes. We consider large distributed systems, where the input is driven by a large population of users (smaller systems offer no difficulty in tracking in the control plane what is happening in the data plane). It is a well known result that the aggregated process resulting from a large population of uncoordinated users will converge to a Poisson process (chapter 3.6 [43]), and therefore the events in the future are independent of the events in the past and depend only on the current state. Thus we assume with no loss of generality that Y_m is an independent and identically distributed (i.i.d) sequence with probability density function (pdf) $f_Y(y)$ and mean θ_X, and Z_m is another i.i.d. sequence with pdf $f_Z(z)$ and mean τ_X. We also assume that any two Y_m and Z_m are mutually independent.

S_X and $\hat{S}_X$ may differ in two cases resulting in two types of distortion; first, when the state of X is changed from $'0'$ to $'1'$ (change type I) but the control plane is not

[1] Note that this Boolean case is just an example to illustrate the method, and can be generalized to other possible values. For instance, to measure the congestion on a link, one could quantize the link congestion into bins (say bins b_1 to b_{10} for normalized link utilization between 0 to 0.1, 0.1 to 0.2,..., 0.9 to 1) and map the link utilization to a 0/1 variable such that $b_h = 1$ if the current link utilization is in $((h-1)/10,\ h/10)$ and 0 otherwise. Obviously using this quantization method the b_h variables would not be independent and only one of them can be $'1'$ at each instant of time. As other way to solve such problem, one can model the changes in the quantized levels as a binary variable. Since the values of the congestion levels change smoothly and there is not any kind of discontinuity in the congestion levels, one can expect going one level up or down in case of any changes. Using this method, one needs to have new distortion definitions. Due to the lack of space we leave it as future work to study other state distributions, where other distortion functions would apply.

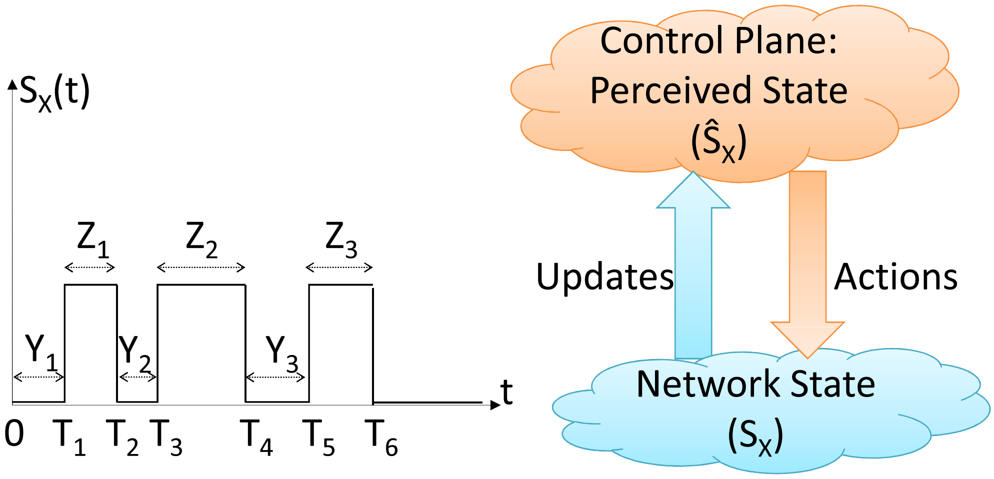

Figure 12.1: **Time diagram of the state of binary random process** X at time t ($S_X(t)$).

notified ($\hat{S}_X = 0,\ S_X = 1$); second, when the state of X is changed from $'1'$ to $'0'$ (change type II) and the control plane still has the old information about it ($\hat{S}_X = 1,\ S_X = 0$). Let D_1 and D_2 denote the probability of the distortions corresponding to changes type I and II, respectively. Here we calculate the minimum rate at which the underlying plane has to update the state of X so that the mentioned distortion probabilities are less than some values ϵ_1 for the first type, and ϵ_2 for the second type, respectively. ϵ_1 and ϵ_2 can be viewed as probability of false negative and false positive alarms at the controller.

We make an additional assumption that the delay of the network is much lower than the time duration of the changes in the forwarding layer, and the control plane will be aware of the announced state immediately[2] (the alternative - that the state of

[2] The control packets sent from the data plane to the control plane are very small in size comparing to the data packets. For example in case of transferring video files, no state changes happen in the data plane unless a video file is downloaded. Since the size of the video files are much larger than the update packets (hundreds of megabytes comparing to a few bytes), the download time and thus the duration of state changes is much lower than the delay of the network for update packets. According to [53] the request round-trip latency in Akamai and Cisco are in the order of a few 10ms, while the download

the system changes as fast or faster than the control plane can be notified of these changes - is obviously unmanageable). Thus, the above errors may occur just when the forwarding plane does not send an update about a change.

The main result now can be stated as a Lemma (with the proof in Appendix).

Lemma 12.1.1. *If the ups and downs in the state of X follow some distributions with means τ_X and θ_X, respectively, then the minimum update rate $R_X(\epsilon_1, \epsilon_2)$ (number of update packets per second) satisfying the mentioned distortion criterion is given by*

$$R_X(\epsilon_1, \epsilon_2) \geq \frac{1}{\tau_X+\theta_X}\left(2 - \frac{\epsilon_1 \frac{\theta_X}{\tau_X}}{\frac{\theta_X}{\tau_X+\theta_X} - \epsilon_2} - \frac{\epsilon_2 \frac{\tau_X}{\theta_X}}{\frac{\tau_X}{\tau_X+\theta_X} - \epsilon_1}\right) \tag{12.1}$$

if $\frac{\epsilon_2}{1-\epsilon_2} < \frac{\theta_X}{\tau_X} < \frac{1-\epsilon_1}{\epsilon_1}$ and $\epsilon_2 \tau_X + \epsilon_1 \theta_X < \frac{\tau_X \theta_X}{\tau_X+\theta_X}$. Otherwise the distortion criteria is satisfied with no update at all.

Lemma 12.1.1 shows the minimum update rate for state of a single random variable X in the underlying plane so that an accepted amount of distortion is satisfied. The total rate and consequently the total protocol overhead for keeping the control layer informed of the forwarding layer is the combination of all the overheads needed for all the random processes of the underlying layer, which may be independent of each other or have some mutual impacts.

In the following section, we will use our model to formulate the control traffic needed in the interaction between caches and controller inside a sub-network.

time for a 1GB movie using a very high speed internet of 100Mbps would take around 10s. This is a practical assumption as well: if the delay in the network is longer than the duration of the changes, then a message sent to update the controller would be carrying obsolete information by the time it reaches the controller. Most practical systems are such that the time to notify the control is sufficient for the controller to use the information when it receives it. However, in any system, there is a chance that the actual state changes while the notification of the previous state is still underway, and there is always some distortion in the state representation at the control vs. the actual state in the forwarding plane.

The notations used here can be found in Tables 12.1-12.2.

12.2 Content Location in Cache Networks

Information-centric networks [109] usually employ resolution-based [39, 40, 65, 99] or routing-based [59] methods for content discovery purposes. In the routing-based discovery methods, like CCN, the required items are found exploring some areas of the network opportunistically or using other solutions like flooding. [9, 68, 102] have studied these methods and proposed solutions to have the best performance. Resolution-based methods, on the other hand, require the control layer to know at least one location for each piece of data. PSIRP, DONA, and NetInf (partly) are some models which use the resolution-based methods. For instance, [99] attempts to set up a route to a nearby copy by requesting the content from a pub/sub mechanism. The pub/sub rendezvous point needs to know the location of the content. This is highly dynamic, as content can be cached, or expunged from the cache at any time. NDN [112] also assumes that the routing plane is aware of multiple locations for a piece of content[3].

In a cache network, the addition/removal of an item (pieces of data which are requested and used by the users) to/from a cache may affect the timings of the other items in that cache; caching a piece of content somewhere may force another content out of the cache, and the caching policy will thus influence the network state (the existing

[3] The routing (in NDN in particular) could know only one route to the content publisher or to an origin server and find cached copies opportunistically on the path to this server. But Fayazbakhsh et. al. [45] have demonstrated that the performance of such an ICN architecture would bring little benefit over that of strict edge caching.

items information), so we need to consider this effect in our calculations. It is worth noting that this framework may be used for CDNs as well, since the basics are the same, the point is that the update traffic for reporting the state of the caches in CDN would be very close to zero, since there are not a large number of changes in their states, unless the acceptable distortion is very low. We assume from now on that the Least-Recently-Used (LRU) replacement policy is used in the caches, as it is a common policy and has been suggested in some ICN architectures [59]. However, based on [77], other policies can be handled in a similar manner by generalizing the decoupling technique of Che's approximation [35].

The request process also impacts the cache state, and we make the usual assumption that the items are requested according to a Zipf distribution with parameter α; meaning that the popularity of an item i is $\alpha_i = \frac{i^{-\alpha}}{\sum_{k=1}^{M} k^{-\alpha}}$, where M is the size of the content set.

In the following sections we first introduce our framework to model the protocol overhead in section 12.1. Then, in section 12.2.1, we use our model to study the total data retrieval cost including the control overhead and data downloading costs in a network of caches, where the nodes update the control plane of a domain (say, an AS) so as to route content to a copy of the cache within this domain if it is available. We denote the control plane function which locates the content for each request as the Content Resolution System (CRS).

12.2.1 Cache-Controller Interaction

Assume that we partition the network into smaller sub-networks each with its own control plane, such that all the nodes in each one of them have similar request patterns. A possible example of such partitioning are the Autonomous Systems (ASs) in the Internet.

Whenever a client has a request for an item, it needs to discover a location of that item, preferably within the AS, and it downloads it from there. To do so, it asks a (logically) centralized *Content Resolution System (CRS)* by sending a Content Resolution Request (CRReq) or locates the content by any other non-centralized locating protocol. The Content Resolution Reply (CRRep) sent back to the client contains the location of the item, then the client starts downloading from the cache identified in CRRep.

If the network domain is equipped with a CRS, it is supposed to have the knowledge of all the caches, meaning that each cache sends its item states (local presence or absence of each item) to the CRS whenever some state changes.

Depending on the caching policy, whenever a piece of content is being downloaded, either no cache, all the intermediate caches on the path, or just the closest cache to the requester on the download path stores it in its content store, independently of the content state in the other caches, or refresh it if it already contains it.

We consider an autonomous network containing N nodes (terminals), each sending requests for items $i = 1, ..., M$ with sizes B_i according to a Poisson distributed process with rate of γ_i. The total request rate for all the items from each node is denoted by $\gamma = \sum_{i=1}^{M} \gamma_i$. Note that all the nodes in an AS have the same request pattern, i.e. content locality is assumed uniform in each AS[4], and that the total request rate of each terminal is a fixed rate independent of the total number of nodes and items while the total requests for all the nodes is a function of N (namely $N\gamma$). If different users have different request distributions, then less cached contents will be reused, and thus there will be more changes in the cache states, and consequently more update traffic will be needed. The uniform content locality will give us the minimum required update rate.

Suppose that there are N_c caches in the system ($\mathcal{V}_c = \{v_1, ..., v_{N_c}\}$) each with size L_c that can keep (and serve) any item i for some limited amount of time τ_i, which depends on the cache replacement policy. Based on the the rate at which each item i enters a cache and the time it stays there, each cache may have item i with some probability ρ_i. For simplicity, we assume that the probability distribution of the contents in all the caches are similar to each other. We can easily extend to the case of heterogeneous caches at the cost of notation complexity. For instance, Theorem 12.2.1 below can be stated as a sum over all N_c possible types of caches with N_c different ρ_is for each type of cache, instead of a product by N_c of identical terms. Our purpose is to describe the homogeneous case, and let the reader adapt the heterogeneous case to suit

[4] This assumption is widely used in works using the mathematical modeling for the networks [30]. This comes from the fact that 1) The requests coming from a specific region are very likely to follow similar patterns, because the users' interest in one area are highly correlated and can be predicted by having the information about just part of them [24]. For example, some certain news title might be of special interest in a certain area, or some new TV series might be very popular in a certain country. 2) each user in this work can actually be a hot spot or a base station, so a request generated from a node is not coming from one specific user but a group of users. So since we assume random users per station, then the assumption of uniform user locality is the best fitted assumption.

her/his specific needs.

In the following Theorem we want to compute the update rate for this system. Let $\bar{N_c}$ denote the number of caches where each downloaded piece of content is stored in (and thus need to send an update), either on the downloaded path, or off-path (Caching policy and network topology are the two factors that determine this number.). Thus, the rate of requests for item i received by each cache is $\lambda_i = \gamma_i N \bar{N_c}/N_c$.

Parameter	Definition
N	Number of users
M	Total number of items
N_c	Number of content stores/caches
$\bar{N_c}$	Number of caches storing the downloaded content
L_c	Storage size per content store (bits)
B_i	Average size of item i (bits)
α_i	Popularity of item i (Zipf)
γ_i	Total request rate for item i per user
γ	Total request rate per user
λ_i	Total request rate for item i received per cache
λ	Total request rate received per cache

Table 12.1: Parameters of the Network Model

Theorem 12.2.1. *The minimum total update rate for each item i in the worst case is*

$$R_i(\epsilon_1,\epsilon_2) \geq N_c\lambda_i(1-\rho_i)$$
$$\{2 - \frac{\epsilon_1(1-\rho_i)}{\rho_i(1-\rho_i-\epsilon_2)} - \frac{\epsilon_2\rho_i}{(1-\rho_i)(\rho_i-\epsilon_1)}\} \tag{12.2}$$

if $\epsilon_1 < \rho_i < 1-\epsilon_2$ *and* $\epsilon_1(1-\rho_i)+\epsilon_2\rho_i < \rho_i(1-\rho_i)$. *Otherwise no update is needed.*

Proof. Let the random process X in the forwarding plane denote the existence of item i in a cache v_j at time t, which is needed to be reported to the control plane (CRS). Let τ_{ij} denote the mean duration of time item i spends in any cache v_j, and θ_{ij} denote the mean duration of time item i not being in the cache v_j.

In order to keep the CRS updated about the content states in the network, all the nodes have to send update packets regarding their changed items to CRS. All the assumptions of section 12.1 are valid here. Thus, by replacing τ_X and θ_X in equation (12.1) with τ_{ij} and θ_{ij} respectively, the result ($R_{ij} = R_X$) shows the minimum rate at which each cache v_j has to send information about item i to the CRS.

It can be seen that at the steady-state, the probability that cache v_j contains item i will be $\rho_{ij} = \frac{\tau_{ij}}{\theta_{ij}+\tau_{ij}}$. On the other hand, the total rate of generating (or refreshing) copies of item i at each cache v_j, denoted by λ_{ij}, equals to $\frac{1}{\theta_{ij}}$. Replacing the values of $\frac{\tau_{ij}}{\tau_{ij}+\theta_{ij}}$ and $\frac{1}{\theta_{ij}}$ in R_{ij} with ρ_{ij} and λ_{ij} respectively, we obtain

$$R_{ij}(\epsilon_1,\epsilon_2) \geq \lambda_{ij}(1-\rho_{ij})$$
$$\{2 - \frac{\epsilon_1(1-\rho_{ij})}{\rho_{ij}(1-\rho_{ij}-\epsilon_2)} - \frac{\epsilon_2\rho_{ij}}{(1-\rho_{ij})(\rho_{ij}-\epsilon_1)}\} \tag{12.3}$$

for $\epsilon_1 < \rho_{ij} < 1 - \epsilon_2$ and $\epsilon_2 \rho_{ij} + \epsilon_1(1 - \rho_{ij}) < \rho_{ij}(1 - \rho_{ij})$.

It is worth noting that we are not assuming any specific topology or caching policy here; the items may be cached on-path or off-path; just one cache may keep the downloaded content or a few caches may keep it. We are looking for the minimum amount of update packets in the worst case, which happens when each cache stores items independent of the items in other caches. It is obvious that topologies like a line of caches which result in strongly dependent caches are not in the scope of this work. Thus, the total update rate for item i, is the sum of the update rates in all caches which is $R_i(\epsilon_1, \epsilon_2) = \sum_{j=1}^{N_c} R_{ij}(\epsilon_1, \epsilon_2)$. Recalling the assumption of (probabilistically) similar caches, we can drop the index j and express the total update rate of item i in terms of the probability of this item being in a cache. This yields the result of equation (12.2) and the total update rate for all the items is the summation of these rates. □

12.2.2 Model Evaluation and Simulation Results

To figure out how the calculated rates perform in practice and evaluate our model, we simulate an LRU cache with capacity $L_c = 20$ items. We use the MovieLens dataset [56], which contains $100,000$ ratings together with their time stamps collected for $M = 1,682$ movies from 943 users during a seven-month period. We took the ratings as a proxy for content requests, assuming that the users who reviewed the movie have requested them shortly prior to the review. In these simulations we first estimate the item availability in the cache ρ_i (by dividing the total time that item is in the cache by the total simulation time), then using the estimated ρ_i and according to equations

Parameter	Definition
τ_i	Average time item i stays in a cache
θ_i	Average time a cache does not have item i
ρ_i	Probability of item i being in a cache
$\epsilon_{1,2}$	Distortion thresholds
$R_{ij}(\epsilon_1, \epsilon_2)$	Minimum rate at which each cache v_j must send update state of item i to CRS so the defined distortion criteria is satisfied
$R_i(\epsilon_1, \epsilon_2)$	Minimum total update rate for item i that satisfies the defined distortion criteria
$R(\epsilon_1, \epsilon_2)$	Minimum total update rate that satisfies the defined distortion criteria

Table 12.2: Parameters Used in Cache-Controller Interaction

(B.29) and (B.30), we calculate the update rate in case of a change. Then we run the simulation for $100,000$ requests. In this simulation we update the CRS according to the calculated rates, which can be interpreted as the chances of update, whenever a change occurs in the cache. Then we measure the total time that the CRS information does not match the actual cache state for each item, and calculate the average generated distortion during 10 rounds of simulation. The top figures in Figure 12.2 illustrate the results for the case where $\epsilon_1 = \epsilon_2 = 0.01$.

Since, according to [65] and [20], the number of data objects is very large, and is becoming larger, we repeated similar evaluation with a relatively large synthetic dataset, containing 10 million Poisson requests for contents picked from a catalog of $100K$ movies, according to a Zipf distribution with skew parameter $\alpha = 0.7$. Bottom figures in Figure 12.2 show the results for the synthetic dataset allowing $\epsilon_1 = \epsilon_2 = 10^{-4}$ distortion accepted (larger number of contents leads to lower cache availability, thus we allowed lower distortion here).

It must be noted here that we are estimating ρ_i based on the past cache states, so it is not the exact ρ_i.Thus the generated distortion may exceed the tolerable values for some items, while they are in the safe zone for the others. It is observed that for a large portion of the items the distortion type I satisfies the distortion criteria. Distortion type II, however, does not satisfy the distortion criteria for more items. The reason is that the calculated update rates are strongly dependent on the availability of the items in the cache and any small error in the estimation of ρ_i may lead to some extra distortions. Since the ρ_i's are mostly very small, not updating just one type II change may cause an

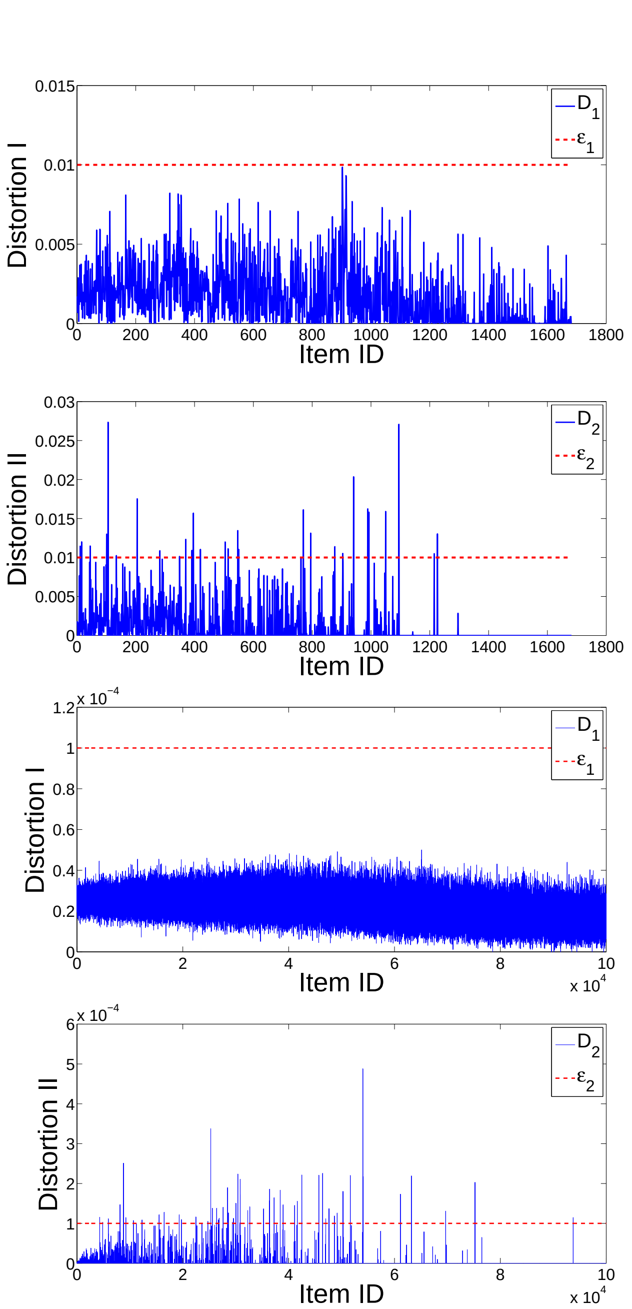

Figure 12.2: **Measured distortion type I (D_1) and II (D_2)** top) for MovieLens dataset ($100K$ requests for $M = 1,682$ movies), with $\epsilon_1 = \epsilon_2 = 0.01$ as accepted distortions, and bottom) for synthetic dataset ($10M$ Poisson requests for $M = 100K$ Zipf distributed movies with skew factor $\alpha = 0.7$), with $\epsilon_1 = \epsilon_2 = 10^{-4}$ as accepted distortions.

error which remain in the system for a long time, and thus creating a large distortion.

Figure 12.3 illustrates the number of needed updates per generated request for each item i in the network $\frac{R_i}{N_c\lambda_i} = \frac{R_i}{N\gamma_i}$ when the caches does not contain it with a known probability $(1-\rho_i)$. The only variable parameters in this graph are ϵ_1 and ϵ_2. The higher distortion we tolerate, the less update announcements for each item i we need to handle. Moreover, the number of items which need some updates is decreasing when higher distortions are accepted. As can be seen the update rate starts from zero for those items which are in the cache with high probability. Status of these items are permanently set to $'1'$ in CRS, and no update is needed. At the other end of the graph, for the items which are almost surely not in the cache, The presence probability is close to zero ($\rho_i = 0$ and thus $1-\rho_i = 1$)), and the status of those items can be permanently set to $'0'$ in the control plane, thus the caches don't need to send any more information regarding those items to the control plane. Therefore, again no update is needed.

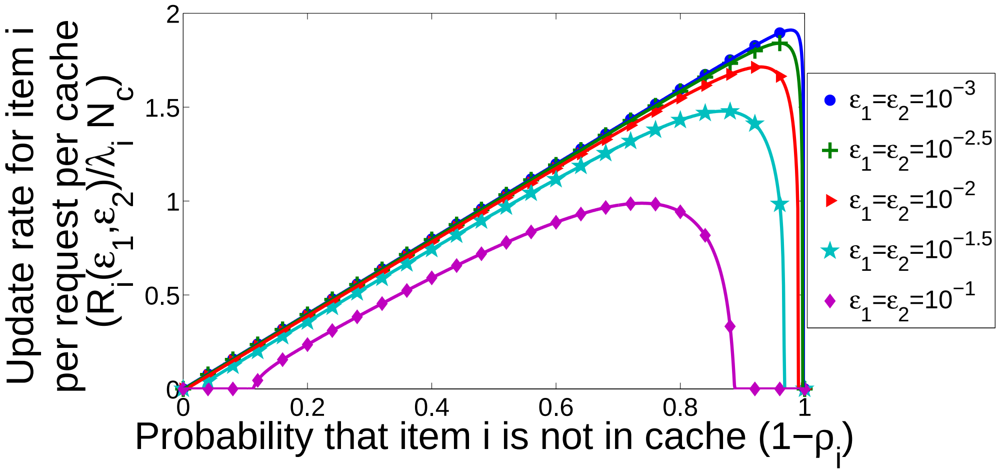

Figure 12.3: **Number of needed updates** for item i sent from all caches to CRS per generated request for that item versus $1-\rho_i$ for different distortion criteria.

The probability ρ_i is strongly dependent on the cache replacement policy. We consider LRU as the cache replacement method used in the network. Clearly, in LRU caches (similarly in other policies like FIFO, LFU, etc.) ρ_i is just a function of the probability of item i coming to the cache (α_i), and the cache capacity (L_c). Figure 12.4 shows the changes of the total update rate (scaled by $\frac{1}{\lambda N_c}$) versus the cache storage size, in a network of LRU caches, such that the distortion criteria defined by (ϵ_1, ϵ_2) is satisfied. In this simulation $M = 10^3$. Note that each change in a cache consists of one item entering into and one other item being expunged from the cache, therefore if no distortion is tolerable, this rate will be 2 updates per change per cache.

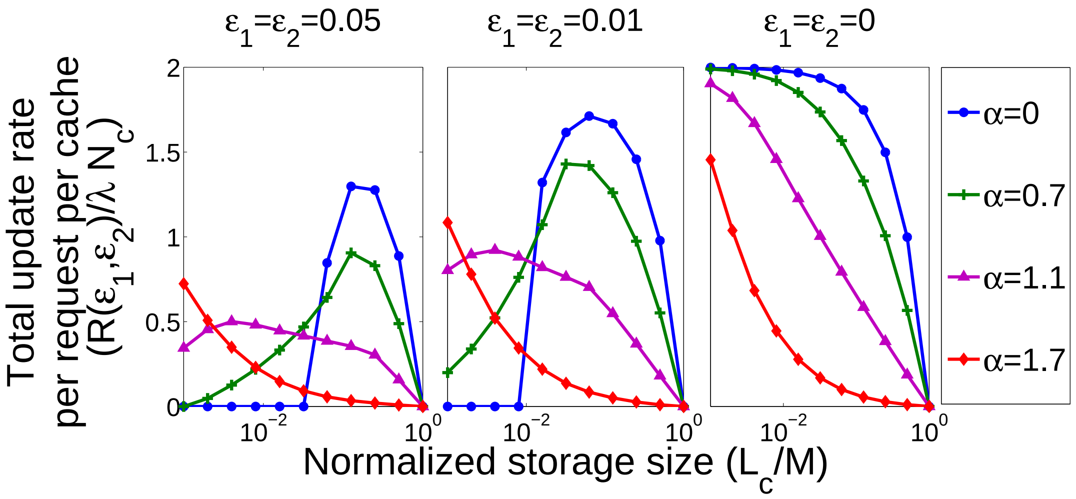

Figure 12.4: **Total cache-CRS update rate** (Updates per generated request per cache) for different cache storage capacities and different acceptable distortions. Content set contains $M = 1,000$ contents.

It can be observed that for very small storage sizes and small popularity index, almost each incoming item changes the status of the cache and triggers an update. When the storage size is still very small, the caches do not provide enough space for storing the items and reusing them when needed, so increasing the size will increase the update rate. At some point, the items will move down and up in the cache before going out, so increasing the storage size more than that will reduce the need to update. However, if the popularity index is large, then increasing cache size from the very small sizes will decrease the need to update since there are just a few most popular items which are being requested.

Moreover, as it is expected, when more distortion is tolerable, the CRS needs fewer change notifications. However, if the cache size is too big, or the popularity exponent is too high, fewer changes will occur, but almost all the changes are needed to be announced to the CRS. On the other hand, for small cache sizes accepting a little distortion will significantly decrease the update rate.

12.3 Application to Cost Analysis

In this section we use Theorem 12.2.1 to study trade-offs involved in updating the content control layer. More specifically, we try to calculate the bounds on the total cost (required bandwidth for download + CRS update) and look at the trade-offs between the cost, the size of the information chunks, the number of caches, and the size of caches.

12.3.1 Layered Network Model

Figure 12.5 illustrates the network model studied in this section. This model consists of entities in three substrates: users are located in first layer; a network of caches with the CRS on the second level; external resources (caches in other networks, Internet, etc.) on the third.

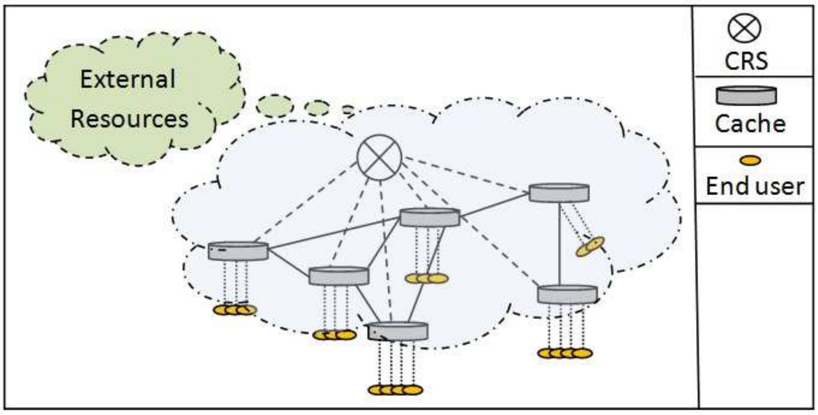

Figure 12.5: **Layered network model.**

We need to define the relative costs of the different actions. We assume that the state update process for item i has a per bit cost of ξ_i^{up} for sending data from the cache to CRS, and per bit cost of $\xi_i^{extup} > \xi_i^{up}$ for sending data from one CRS to another one. On the other hand, the requested piece of content i may be downloaded from a cache inside the network with some per bit cost ξ_i^{int}, or it may be downloaded from an external server with some other cost $\xi_i^{ext} > \xi_i^{int}$. These costs may be a function of the number of hops in the network. Note that the exact costs for each cache are determined based on the network topology, and may not be the same for all the caches. In this work we use the average cost over all the caches [5].

[5] The other option for defining the distortion and correspondingly cost is the worst case, which will map to the maximum update cost. There are many few caches that may undergo some maximum number of changes and need some maximum update transfers, thus, this is clearly not illustrating the performance of a cache network correctly. We have decided to work with the average method, which is very common in the literature (References [58,100,101]) and we believe it can represent the performance of the entire network better in our specific application and many others.

12.3.2 Total Cost in Cache-Controller Interaction

Lemma 12.3.1. *The length of each update packet for content i is*

$$l_i \geq \log N_c - \log \frac{\lambda_i(1-\rho_i)}{\sum_{i=1}^{M} \lambda_k(1-\rho_k)} + 1 \tag{12.4}$$

Proof. Each update packet contains the ID of the cache issuing the query, the ID of the updated item and its new state. There are N_c caches in the network, hence, $\log N_c$ bits are needed to represent the cache. Item i is updated with probability $\beta_i = \frac{\lambda_i(1-\rho_i)}{\sum_{i=1}^{M} \lambda_k(1-\rho_k)}$, which results in a code length of at least $-\log \beta_i$ bits. Thus, the length of each update packet is $l_i \geq \log N_c - \log \beta_i + 1$. □

Lemma 12.3.2. *The total update cost in the defined network is*

$$\varphi^{up} = \sum_{i=1}^{M} R_i(\epsilon_1, \epsilon_2) l_i \xi_i^{up}. \tag{12.5}$$

where $R_i(\epsilon_1, \epsilon_2)$ is the minimum rate at which the update state of item i must be reported to CRS so that a distortion criteria defined by (ϵ_1, ϵ_2) is satisfied.

Proof. Each cache sends update packets at rate $R_i(\epsilon_1, \epsilon_2)$ to provide its CRS with the state of item i in its local content store. Each update packet contains l_i bits, and there is a per bit cost of ξ_i^{up} for the update packets. Therefore, the cost for updating information about item i in the sub-network is $\varphi_i^{up} = R_i(\epsilon_1, \epsilon_2) l_i \xi_i^{up}$, and the total update cost is $\varphi^{up} = \sum_{i=1}^{M} R_i(\epsilon_1, \epsilon_2) l_i \xi_i^{up}$. □

Lemma 12.3.3. *The total download cost in the defined network is*

$$\varphi^{dl} = \sum_{i=1}^{M} N\gamma_i B_i((P_i - \rho_i)\xi_i^{int} + (1 - P_i)\xi_i^{ext}) \tag{12.6}$$

if $\epsilon_1 < \rho_i < 1 - \epsilon_2$ *and* $\epsilon_1(1-\rho_i) + \epsilon_2\rho_i < \rho_i(1-\rho_i)$*. Otherwise no update is needed.*

Proof. The requested piece of content i may be downloaded from the local cache with cost 0 (with probability ρ_i of being in this cache), from another cache inside the same network with some per bit cost ξ_i^{int} (with a probability we denote by $P_i - \rho_i$, where P_i is the probability that content i is within the AS's domain), or it must be downloaded from an external server with some other cost ξ_i^{ext} (with probability $(1-P_i)$). Obviously, $\rho_i \leq P_i \leq 1$. Thus, the download cost for item i with size B_i bits for each user in the sub-network is

$$\varphi_{ij}^{dl} = \gamma_i B_i((P_i - \rho_i)\xi_i^{int} + (1-P_i)\xi_i^{ext}), \tag{12.7}$$

The total download cost for item i is $\varphi_i^{dl} = N\varphi_{ij}^{dl}$, and the total download cost for all the items is the summation of φ_i^{dl}'s over all the contents. □

Theorem 12.3.4. *The total cost in the defined network including update and download costs is*

$$\begin{aligned}\varphi &= \sum_{i=1}^{M} N\gamma_i B_i((P_i - \rho_i)\xi_i^{int} + (1-P_i)\xi_i^{ext}) \\ &+ \sum_{i=1}^{M} R_i(\epsilon_1, \epsilon_2) l_i \xi_i^{up}. \end{aligned} \tag{12.8}$$

Proof. Adding Lemmas 12.3.2 and 12.3.3 proves the Theorem. □

It can be seen that the total cost is strongly dependent on where each query is served from, and consequently on the probability of each item being internally served

(P_i). This probability depends on the probability of that item being in an internal cache, which is in turn a function of the caching criteria and the replacement policy. Lemma 12.3.5 presents some bounds on P_i based upon the allowed distortion. The proof can be found in appendix.

Lemma 12.3.5. *The probability that each content i is internally served is bounded by*

$$[1 - (1 - \rho_i + \epsilon_1)^{N_c}]^+ \leq P_i \leq 1 - (1 - \rho_i)^{N_c}. \tag{12.9}$$

where $[x]^+ = max(x, 0)$,

Note that the above P_i may take any value in the calculated bounds depending on the value of ρ_i. For example if $\rho_i < \epsilon_1$ then $D_{1_i} = \rho_i$, and $P_i = 0$, which is the lowest value of this bound. On the other hand, if $\rho_i > 1 - \epsilon_2$ then $D_{1_i} = 0$, and $P_i = 1 - (1 - \rho_i)^{N_c}$, which is the highest value in this bound. All the other values of ρ_i will lead to P_i between those two boundaries.

These two extreme cases of P_i result in some bounds on the download cost. Let φ_L^{dl} and φ_H^{dl} denote the lower and upper bounds of the download cost corresponding to the upper and lower bounds of P_i, respectively, and φ_L and φ_H denote the lower and upper bounds of the total cost. Note that for small values of the tolerable distortion ϵ_1 the upper and lower limits of P_i and correspondingly the bounds of download cost are very close to each other.

Figure 12.6 the left plot illustrates the changes of update and download costs in a network with a content set of a total of 1 million contents, when the size of each cache is limited to $L_c = 100$ contents. The length of the data packets is assumed to be

$100KB$ in average, while the update packets are l_i bytes each. Note that increasing the data (or update) packet lengths will increase the download (or update) cost linearly.

Here we assume that whenever an item is downloaded, it is stored in $\bar{N_c} = \log N_c$ caches, which have to report the changes to the controller[6]. If these caches are selected randomly, the total update rate would be $\bar{N_c}$ times the rate of update of each cache resulting in max φ^{up}. On the other side, if they are completely dependent, for example if all the caches on the download path keep it, then just one update may be enough, resulting in min φ^{up}. So depending on the caching policy, the update cost will be something between min and max φ^{up}.

The request rate received by each cache is inversely proportional to the number of caches (the request rate per user is assumed to be fixed and independent of N_c), and the update packet length increases logarithmically with the number of caches. The total update rate per cache is almost linearly decreasing with N_c, hence the minimum of the total update rate, or the total update rate if just one cache keeps the downloaded item, will almost be stable when N_c varies (changes are in the order of $\log N_c$). The maximum total update rate will linearly change with the number of copies $\bar{N_c}$ per download. Thus, increasing the number of caches in the network increases the update cost by a factor of at least $\log N_c$ and at most $\bar{N_c} \log N_c$.

Increasing N_c, however, increases the probability of an item being served internally and thus decreases the download cost. Nevertheless, as it can be observed, the rate of decrease is so low that it can be assumed as stable.

[6] This happens in largely used network models like binary tree or grid topology, when all the caches on the download path store the content.

In the right plot of figure 12.6, we fix the number of caches in the AS ($N_C = 10$) and study the effects of cache storage size on the update and total cost. Increasing the cache size, simply increases the probability of an item being served internally and decreases the download cost. Again similar to the left plot, the rate of changes in the download cost is very low. As expected, on the other side, the update cost shows an increase when increasing the storage size. Looking at each cache, very small cache size leads to very large durations where that item is not in that cache and consequently, the update rate would be low. Increasing the storage size will increase the probability of that item being in the cache, and thus increases the update rate. If we let more cache storage, this increase reaches its highest value for a certain value of cache size, and for larger values of cache beyond a threshold, the item is in the cache most of the time. Therefore, we need less updates and increasing the cache size will increase the duration of the item being in the cache leading to fewer update messages. Since the total cost mostly depends on the download cost, by increasing the cache size, this value reaches its minimum value.

It is worth noting here that the cost of download from another AS has been assumed 5 times bigger than the cost of download from inside the AS, which in turn is assumed to be the same as the update cost per bit. Figure 12.7 shows the impact of the external and internal costs on the total download cost. Higher external costs result in higher total download costs, as it is expected, and show more decrease rate when the number of caches increases. Thus, if the external download cost comparing to the internal download cost is very high, having more caches may make sense, although, the total cost decrease rate is still very low.

Another important result shown by figures 12.6 and 12.7 is that having big data packets, the download cost is always much higher than the update cost, which is reasonable. In other words, having the resolution-based content discovery when the data packets are large, will add just negligible cost to update the controller. In the following we study the affect of chunk-based caching to obtain some insight on the reasonable size of chunks, such that the update cost remains negligible.

The top plot in figure 12.8 shows the total cost versus the number of caches, when the LRU cache replacement policy is used and the total storage of the cache sublayer is limited to half of the total number of items. The parameters are set as follows: $M = 10,000$, $\epsilon_{1,2} = 1e - 4$, $B_i = 100K$, $\xi_i^{int} = 1$, $\xi_i^{ext} = 5$, $\xi_i^{up} = 1$, and $\alpha = 0.7$. Since the lower and upper bounds of the total cost are very close to each other, we just plot the upper limit here. In the bottom plot of this figure, the total cost is plotted versus the size of each cache. It can be observed that with a fixed total storage size, concentrating all the caches in one node and increasing the size of it will lead to better overall performance (least cost). Note that in these figures the total cost value shown is just a relative value, and not the exact one.

12.3.3 Optimized Cache Management

In previous section, the total cost in the described cache network was derived and the impacts of the number or size of the content stores on this cost was studied. We now turn our attention to minimizing the total cost for given N_c and L_c.

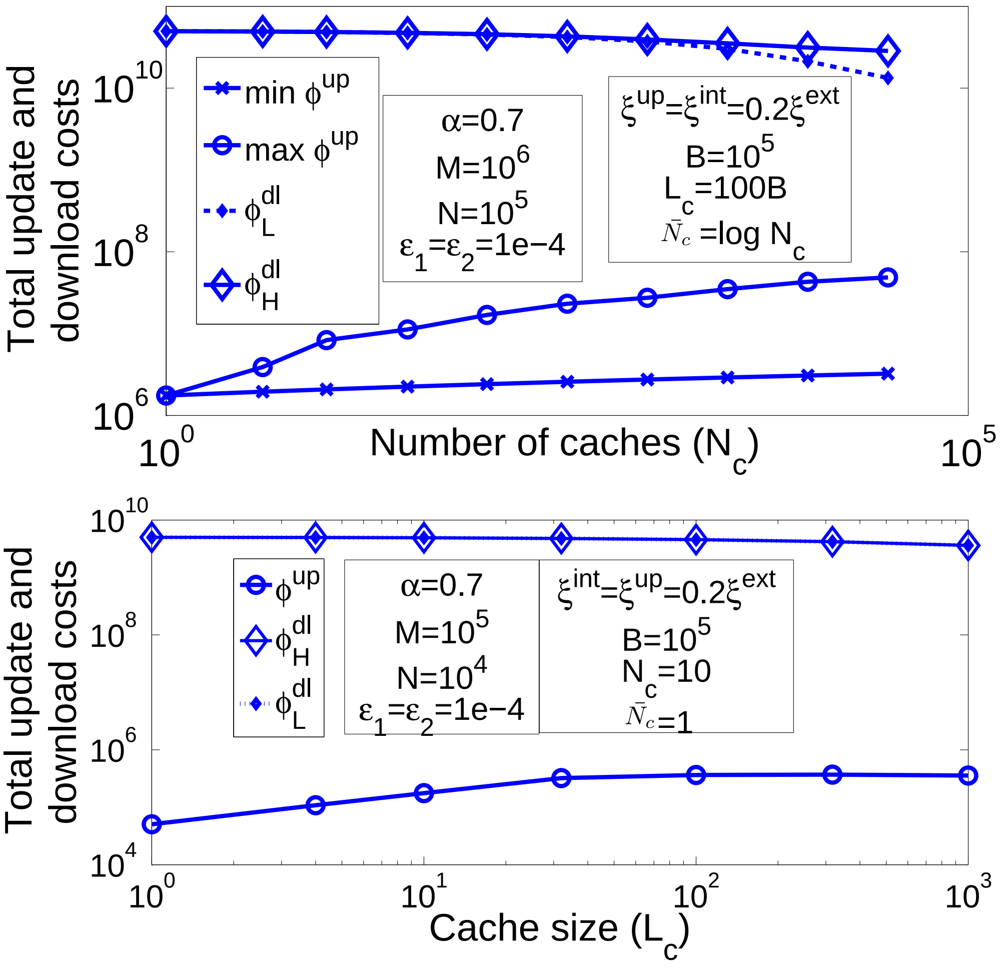

Figure 12.6: **Total update costs vs. number of caches** (minimum and maximum of φ^{up}) and total download cost (lower and upper bounds, φ_L^{dl} and φ_H^{dl}, left) vs. the number of caches (N_c), when each item is $B = 10^5$ units long, the storage size per cache is fixed ($L_c = 100$ items), and each downloaded item is stored in $\bar{N}_c = \log N_c$ caches, and right) vs. the cache size (L_c), when each item is $B = 10^5$ units long, the number of caches is fixed ($N_c = 10$), and each downloaded item is stored in $\bar{N}_c = 1$ cache.

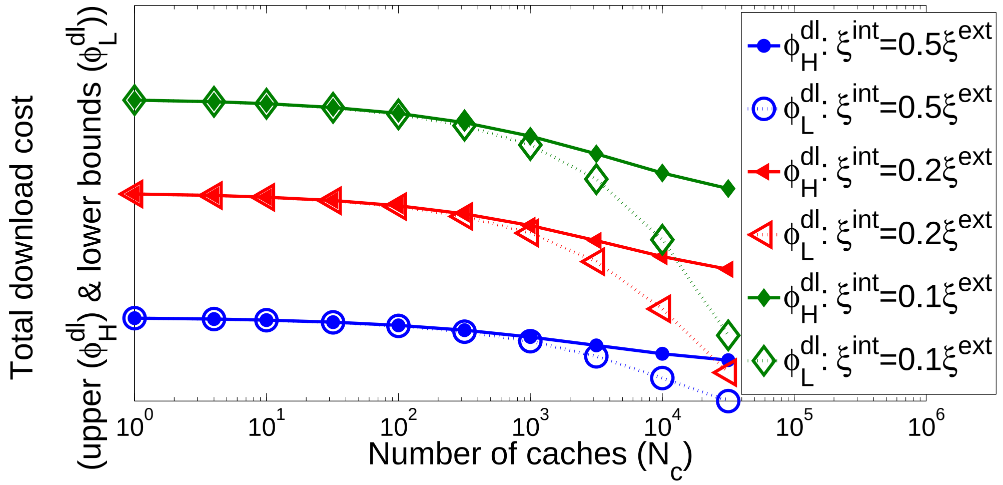

Figure 12.7: **Total download cost vs. the number of caches** (lower and upper bounds, φ_L^{dl} and φ_H^{dl}) vs. the number of caches (N_c), for different download cost values, when each item is $B = 10^5$ units long, and the storage size per cache is fixed ($L_c = 100$ items).

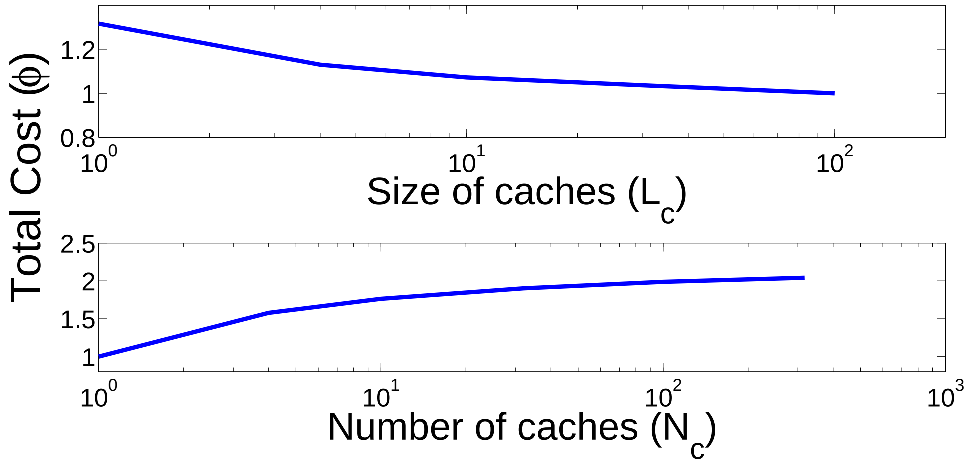

Figure 12.8: **Total cost** (φ),when the total storage size (N_cL_c) is fixed and equal to half of the catalog size, vs. top) the size of caches (L_c), and bottom) the number of caches (N_c).

Under a Zipf popularity distribution, many rare items will not be requested again while they are in the cache under the LRU policy. We can rewrite the total cost if the caches only keep the items with popularity from 1 up to i^*.

$$\begin{aligned}\varphi &= \sum_{i=1}^{i^*} N\gamma_i B_i((P_i-\rho_i)\xi_i^{int}+(1-P_i)\xi_i^{ext}) \\ &+ \sum_{i=i^*+1}^{M} N\gamma_i B_i \xi_i^{ext} + \sum_{i=1}^{i^*} R_i(\epsilon_1,\epsilon_2) l_i \xi_i^{up} \qquad (12.10)\end{aligned}$$

Now just i^* different pieces of content may be stored in each cache. This changes the probability of an item $i = 1, ..., i^*$ being in a cache (ρ_i), which in turn changes P_i and R_i.

Figures 12.9 demonstrates the total cost versus the caching popularity threshold i^*, for different number and size of content stores, and acceptable distortions.

If just a very small number of items (small i^*) are kept inside cache layer, then the download cost for those which are not allowed to be inside caches will be the dominant factor in the total cost and will increase it. On the other hand, if a lot of popularity classes are allowed to be kept internally, then the update rate is increased and also the probability of the most popular items being served internally decreases, so the total cost will increase. There is some optimum caching popularity threshold where the total cost is minimized. This optimum threshold is a function of the number and size of the stores, distortion criteria, per bit cost of downloads and updates.

The benefit of the optimized solution also varies depending on the mentioned parameters. For example according to figure 12.9, the optimized solution can have 17% reduction in cost in case when $N_c = 50, L_c = 10, \epsilon_1 = \epsilon_2 = 1e-4$, while this cost reduction is just 7% when N_c is five times smaller.

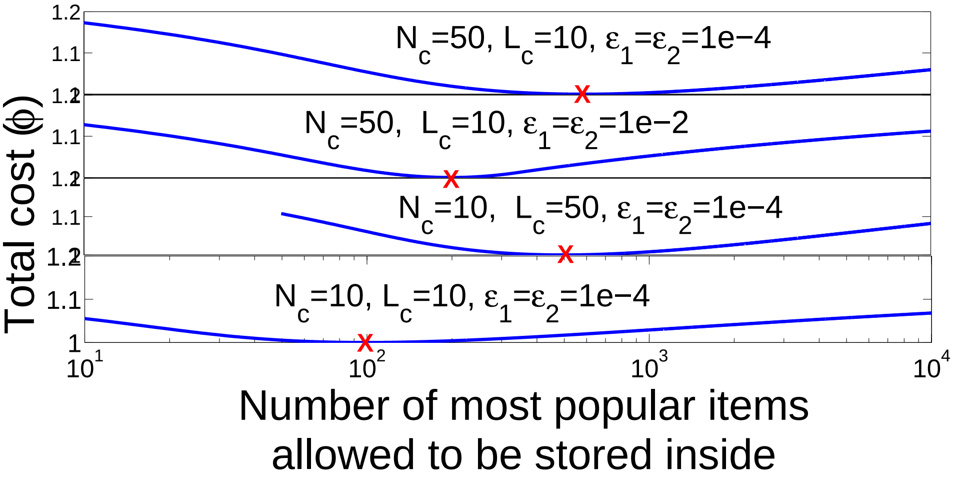

Figure 12.9: **Total cost when just i^* most popular items are allowed** to be stored inside caches ($N = 10^3, M = 10^4, B = 10^5, \alpha = 0.7$).

To find the optimal i^*, assume that all the items have the same size ($B_i = B$) and the per bit costs is fixed for all popularity classes ($\xi_i^{int} = \xi^{int}$, $\xi_i^{ext} = \xi^{ext}$, $\xi_i^{up} = \xi^{up}$). We can rearrange equation (12.10).

$$\varphi = \varphi_1 - \varphi_2 + \varphi_3 \tag{12.11}$$

where $\varphi_1 = BN\gamma\xi^{ext}$ is the total cost if no cache exists and all the requests are served externally; $\varphi_2 = BN\gamma(\xi^{ext} - \xi^{int})\sum_{i=1}^{i^*}\alpha_i P_i + BN\gamma\xi^{int}\sum_{i=1}^{i^*}\alpha_i\rho_i$ corresponds to the benefit of caching (cost reduction due to caching); and $\varphi_3 = \xi^{up}\sum_{i=1}^{i^*}R_i(\epsilon_1, \epsilon_2)l_i$ is the caching overhead cost due to the updates. The last term is the cost we pay because of caching (updates). We need to calculate the value of i^* such that the cost of caching is dominated by its advantage; i.e. we need to maximize $\varphi_2 - \varphi_3$.

This can be done using numerical methods which will lead to a unique i^* for each network setup (fixed parameters). However, the network characteristics and the request pattern are changing over time, so it seems that it is better to have a mechanism to dynamically optimize the cost by selecting the caching threshold (i^*) according to the varying network features.

In such a mechanism, the CRS can keep track of requests and have an estimation of their popularity. For those requests which are served locally the CRS can have an idea of the popularity based on the updates that receives from all the caches; i.e. the longer an item stays in a cache, the more popular it is. It can also take into account the local popularity of the items. The CRS can then dynamically search for the caching threshold which minimizes the total cost by solving equation (12.11). Once the CRS determines which items to keep internally, it will set/reset a flag in each CRRep so that the local cache knows to store or not to store the requested piece of content.

Chapter 13

Conclusion and Future View

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 $\underset{k}{min}\rho^{(k)}(n)$ is sufficiently small ($\underset{k}{min}\rho^{(k)}(n) \preceq \frac{1}{n}$) or big enough ($\underset{k}{min}\rho^{(k)}(n) \nrightarrow 0$). For other values of $\underset{k}{min}\rho^{(k)}(n)$, 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.

We also formulated a distortion-based protocol overhead model. Some simple content distribution networks were then considered as examples to show how this framework can be used, and based on this model the overhead of keeping the control plane informed about the states of the contents in these networks was calculated. It was confirmed that with big data packets, or in large un-chunked data transfer scenarios, the cost of updating the control layer is much lower than the cost of data download, so resolution-based content discovery can be a good solution.

We also studied the total cost of data retrieval and observed that with limited cache storage sizes, allowing all the items to have the opportunity to be stored inside the sub-network's caches is not always the most efficient way of using the caching feature.

For the case with a central resolution system in each sub-network and with LRU cache replacement policy, an algorithm has been proposed that can dynamically determine which items not to be cached inside the AS at any time such that the total cost of data retrieval is minimized.

In this work, our overhead model focuses on systems with Boolean states. Our future work involves systems with other state distributions. In addition, we have assumed uniform distribution of caches in the studied example. This assumption means that the probability of an item being in all the caches are the same. Future study can consider some structure like tree or power-law for the caches inside each sub-network, and using the described framework, investigate how this assumption changes the results.

As a final note, distributed caching may introduce security issues, solving these issues is another path for future work.

Bibliography

[1] PURSUIT: Pursuing a pub/sub internet. http://www.fp7-pursuit.eu/, September 2010.

[2] Lada A Adamic. The small world web. In *International Conference on Theory and Practice of Digital Libraries*, pages 443–452. Springer, 1999.

[3] Lada A Adamic and Bernardo A Huberman. Power-law distribution of the world wide web. *science*, 287(5461):2115–2115, 2000.

[4] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, and B. Ohlman. A survey of information-centric networking. *Communications Magazine, IEEE*, 50(7), July 2012.

[5] Bengt Ahlgren, Matteo D'Ambrosio, Marco Marchisio, Ian Marsh, Christian Dannewitz, Börje Ohlman, Kostas Pentikousis, Ove Strandberg, René Rembarz, and Vinicio Vercellone. Design considerations for a network of information. In *ACM CoNEXT*, pages 1–6, 2008.

[6] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[7] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Internet: Diameter of the world-wide web. *nature*, 401(6749):130–131, 1999.

[8] G. Alfano, M. Garetto, and E. Leonardi. Content-Centric Wireless Networks With Limited Buffers: When Mobility Hurts. *Networking, IEEE/ACM Transactions on*, 20, 2014.

[9] Carlos Anastasiades, Arian Uruqi, and Torsten Braun. Content discovery in opportunistic content-centric networks. In *IEEE LCN WKSHPS*, pages 1044–1052, 2012.

[10] B. Azimdoost, Hamid R. Sadjadpour, and J.J. Garcia-Luna. Capacity of wireless networks with social behavior. In *IEEE Transaction on Wireless Communications*, January, vol. 12, no 1 2013.

[11] B. Azimdoost, C. Westphal, and H. R. Sadjadpour. On the throughput capacity of information-centric networks. In *IEEE ITC25*, pages 1–9, 2013.

[12] Bita Azimdoost, Golnaz Farhadi, Noor Abani, and Akira Ito. Optimal in-network cache allocation and content placement. In *IEEE INFOCOM WKSHPS*, pages 263–268, 2015.

[13] Bita Azimdoost and Hamid R. Sadjadpour. Capacity of scale free wireless net-

works. In *Proceedings of the Global Communication Conference (Globecom)*, Anaheim, California, USA, December 2012.

[14] Bita Azimdoost, Hamid R Sadjadpour, and JJ Garcia-Luna-Aceves. The impact of social groups on the capacity of wireless networks. In *Network Science Workshop (NSW), 2011 IEEE*, pages 30–37. IEEE, 2011.

[15] Bita Azimdoost, Hamid R Sadjadpour, and Jose Joaquin Garcia-Luna-Aceves. Capacity of composite networks: Combining social and wireless ad hoc networks. In *Wireless Communications and Networking Conference (WCNC), 2011 IEEE*, pages 464–468. IEEE, 2011.

[16] Bita Azimdoost, Cedric Westphal, and Hamid R Sadjadpour. Fundamental limits on throughput capacity in information-centric networks. *IEEE Transactions on Communications*, 64(12):5037–5049, 2016.

[17] Bita Azimdoost, Cedric Westphal, and Hamid R Sadjadpour. Resolution-based content discovery in network of caches: Is the control traffic an issue? *IEEE Transactions on Communications*, 2017.

[18] Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pages 61–70. ACM, 2010.

[19] Ejder Baştuğ, Mehdi Bennis, and Mérouane Debbah. A Transfer Learning Ap-

proach for Cache-Enabled Wireless Networks. In *IEEE WiOpt*, pages 161–166, September 2015.

[20] James Bankoski. ICNC'17 Panel on Future Video Distribution, 2017.

[21] A Barbir, B Cain, R Nair, and O Spatscheck. Known content network (CN) request-routing mechanisms. *IETF RFC 3568, Network Working Group*, July 2003.

[22] Md Faizul Bari, Arup Raton Roy, Shihabur Rahman Chowdhury, Qi Zhang, Mohamed Faten Zhani, Reaz Ahmed, and Raouf Boutaba. Dynamic controller provisioning in software defined networks. In *IEEE CNSM*, pages 18–25, 2013.

[23] Ejder Bastug, Mehdi Bennis, and Mérouane Debbah. Living on the edge: The role of proactive caching in 5g wireless networks. *IEEE Communications Magazine*, 52(8):82–89, 2014.

[24] Ejder Baştuğ, Mehdi Bennis, Engin Zeydan, Manhal Abdel Kader, Ilyas Alper Karatepe, Ahmet Salih Er, and Mérouane Debbah. Big data meets telcos: A proactive caching perspective. *Journal of Communications and Networks*, 17(6):549–557, 2015.

[25] B. N. Bharath, K. G. Nagananda, and H. V. Poor. A Learning-Based Approach to Caching in Heterogenous Small Cell Networks. In *arXiv preprint arXiv: 1508.03517*, October 2015.

[26] S. Bhattacharkee, K. Calvert, and E. Zegura. Self-organizing wide-area network caches. In *IEEE INFOCOM*, 1998.

[27] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D-U Hwang. Complex networks: Structure and dynamics. *Physics reports*, 424(4):175–308, 2006.

[28] F. Boccardi, R. W. Heath Jr, A. Lozano, T. L. Marzetta, and P Popovski. Five disruptive technology directions for 5g. *IEEE Communication Magazine*, pages 74 – 80, 2014.

[29] S. Borst, V. Gupta, and A. Walid. Distributed Caching Algorithms for Content Distribution Networks. In *IEEE INFOCOM*, March 2010.

[30] Lee Breslau, Pei Cao, Li Fan, Graham Phillips, and Scott Shenker. Web caching and zipf-like distributions: Evidence and implications. In *IEEE INFOCOM*, volume 1, pages 126–134, 1999.

[31] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web. *Computer networks*, 33(1):309–320, 2000.

[32] G. Carofiglio, M. Gallo, L. Muscariello, and D. Perino. Modeling data transfer in content-centric networking. In *IEEE ITC23*, pages 111–118, 2011.

[33] Wei Koong Chai, Diliang He, Ioannis Psaras, and George Pavlou. Cache "less for

more" in information-centric networks. In *International Conference on Research in Networking*, pages 27–40. Springer, 2012.

[34] H. Che, Z. Wang, and Y. Tung. Analysis and design of hierarchical web caching systems. In *IEEE INFOCOM*, pages 1416–1424, 2001.

[35] Hao Che, Ye Tung, and Zhijun Wang. Hierarchical web caching systems: Modeling, design and experimental results. *IEEE Journal on Selected Areas in Communications*, 20(7):1305–1314, 2002.

[36] Kideok Cho, Munyoung Lee, Kunwoo Park, Ted Taekyoung Kwon, Yanghee Choi, and Sangheon Pack. Wave: Popularity-based and collaborative in-network caching for content-oriented networks. In *IEEE INFOCOM WKSHPS*, pages 316–321, 2012.

[37] Andrew R. Curtis, Jeffrey C. Mogul, Jean Tourrilhes, Praveen Yalagandula, Puneet Sharma, and Sujata Banerjee. Devoflow: scaling flow management for high-performance networks. *ACM SIGCOMM CCR*, 41(4), August 2011.

[38] Ali Dabirmoghaddam, Maziar M. Barijough, and Garcia Luna Aceves. Understanding Optimal Caching and Opportunistic Caching at "the Edge" of Information-centric Networks. In *ACM ICN*, pages 47–56, 2014.

[39] Matteo D'Ambrosio, Christian Dannewitz, Holger Karl, and Vinicio Vercellone. Mdht: a hierarchical name resolution service for information-centric networks. In *ACM SIGCOMM ICN*, pages 7–12, 2011.

[40] Christian Dannewitz, Dirk Kutscher, BöRje Ohlman, Stephen Farrell, Bengt Ahlgren, and Holger Karl. Network of information (netinf)–an information-centric networking architecture. *Computer Communications*, 36(7):721–735, 2013.

[41] Mostafa Dehghan, Anand Seetharam, Bo Jiang, Ting He, Theodoros Salonidis, Jim Kurose, Don Towsley, and Ramesh Sitaraman. On the complexity of optimal routing and content caching in heterogeneous networks. In *IEEE INFOCOM*, pages 936–944, April 2015.

[42] Martin Dietzfelbinger and Philipp Woelfel. Tight lower bounds for greedy routing in uniform small world rings. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 591–600. ACM, 2009.

[43] Rick Durrett. *Probability: theory and examples.* Cambridge university press, 2010.

[44] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the internet topology. In *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication (SIGCOMM)*, volume 29, Cambridge, Massachusetts, USA, October 1999.

[45] Seyed Kaveh Fayazbakhsh, Yin Lin, Amin Tootoonchian, Ali Ghodsi, Teemu Koponen, Bruce Maggs, KC Ng, Vyas Sekar, and Scott Shenker. Less pain, most of the gain: Incrementally deployable icn. In *ACM SIGCOMM CCR*, volume 43, pages 147–158, 2013.

[46] N. Choungmo Fofack, M. Dehghan, D. Towsley, M. Badov, and D. L. Goeckel.

On the Performance of General Cache Networks. In *IEEE VALUETOOLS*, pages 106–113, 2014.

[47] N. Choungmo Fofack, Philippe Nain, Giovanni Neglia, and Don Towsley. Analysis of TTL-based cache networks. In *IEEE VALUETOOLS*, pages 1–10, 2012.

[48] Pierre Fraigniaud and George Giakkoupis. On the searchability of small-world networks with arbitrary underlying structure. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 389–398. ACM, 2010.

[49] Robert Gallager. Basic limits on protocol information in data communication networks. *IEEE Transactions on Information Theory*, 22(4):385–398, 1976.

[50] Ali Ghodsi, Scott Shenker, Teemu Koponen, Ankit Singla, Barath Raghavan, and James Wilcox. Information-centric networking: seeing the forest for the trees. In *ACM HotNets-X*, November 2011.

[51] S. Gitzenis, G. S. Paschos, and L. Tassiulas. Asymptotic Laws for Joint Content Replication and Delivery in Wireless Networks. *Information Theory, IEEE Transactions on*, 59(5):2760–2776, May 2013.

[52] Negin Golrezaei, Karthikeyan Shanmugam, Alexandros G. Dimakis, Andreas F. Molisch, and Giuseppe Caire. FemtoCaching: Wireless video content delivery through distributed caching helpers. In *IEEE INFOCOM*, pages 1107–1115, March 2012.

[53] Mark Gritter and David R Cheriton. An architecture for content routing support in the internet. In *USITS*, volume 1, pages 4–4, 2001.

[54] Matthias Grossglauser and David Tse. Mobility increases the capacity of ad hoc wireless networks. *Networking, IEEE/ACM Transactions On*, 10(4):477–486, 2002.

[55] Piyush Gupta and Panganmala R Kumar. The capacity of wireless networks. *IEEE Transactions on information theory*, 46(2):388–404, 2000.

[56] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4):19, 2016.

[57] Jeffrey D Herdtner and Edwin KP Chong. Throughput-storage tradeoff in ad hoc networks. In *IEEE INFOCOM*, pages 2536–2542, 2005.

[58] Jun Hong and Victor OK Li. Impact of information on network performance-an information-theoretic perspective. In *IEEE GLOBECOM*, pages 1–6, 2009.

[59] Van Jacobson, Diana K. Smetters, James D. Thornton, Michael F. Plass, Nicholas H. Briggs, and Rebecca L. Braynard. Networking named content. In *ACM CoNEXT*, pages 1–12, 2009.

[60] Mingyue Ji, Giuseppe Caire, and Andreas F. Molisch. Fundamental Limits of Caching in Wireless D2D Networks. *Information Theory, IEEE Transactions on*, 62(2):849–869, February 2016.

[61] Mingyue Ji, Giuseppe Caire, and Andreas F. Molisch. Wireless Device-to-Device Caching Networks: Basic Principles and System Performance. *Selected Areas in Communications, IEEE Journal on*, 34(1):176–189, January 2016.

[62] Mohsen Karimzadeh Kiskani, Bita Azimdoost, and Hamid R Sadjadpour. Effect of social groups on the capacity of wireless networks. *IEEE Transactions on Wireless Communications*, 15(1):3–13, 2016.

[63] Mohsen Karimzadeh Kiskani, Hamid Sadjadpour, and Mohsen Guizani. Social interaction increases capacity of wireless networks. In *Wireless Communications and Mobile Computing Conference (IWCMC), 2013 9th International*, pages 467–472. IEEE, 2013.

[64] Jon Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 163–170. ACM, 2000.

[65] Teemu Koponen, Mohit Chawla, Byung G. Chun, Andrey Ermolinskiy, Kye H. Kim, Scott Shenker, and Ion Stoica. A data-oriented (and beyond) network architecture. In *ACM SIGCOMM*, pages 181–192, 2007.

[66] S. Kulkarn and P. Viswanath. A deterministic approach to throughput scaling in wireless networks. *IEEE Transactions on Information Theory*, 50:1041–1049, 2004.

[67] Bibb Latané, James H Liu, Andrzej Nowak, Michael Bonevento, and Long Zheng.

Distance matters: Physical space and social impact. *Personality and Social Psychology Bulletin*, 21(8):795–805, 1995.

[68] Munyoung Lee, Junghwan Song, Kideok Cho, Sangheon Pack, Jussi Kangasharju, Yanghee Choi, et al. Content discovery for information-centric networking. *Computer Networks*, 83:1–14, 2015.

[69] Dan Levin, Andreas Wundsam, Brandon Heller, Nikhil Handigol, and Anja Feldmann. Logically centralized?: state distribution trade-offs in software defined networks. In *ACM HotSDN*, August 2012.

[70] Jinyang Li, Charles Blake, Douglas SJ De Couto, Hu Imm Lee, and Robert Morris. Capacity of ad hoc wireless networks. In *Proceedings of the 7th annual international conference on Mobile computing and networking*, pages 61–69. ACM, 2001.

[71] Xiang-Yang Li. Multicast capacity of wireless ad hoc networks. *IEEE/ACM Transactions on Networking (TON)*, 17(3):950–961, 2009.

[72] David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628, 2005.

[73] A. Liu and V. Lau. Asymptotic Scaling Laws of Wireless Adhoc Network with Physical Layer Caching. *Wireless Communications, IEEE Transactions on*, 2015.

[74] Benyuan Liu, D. Towsley, and A. Swami. Data gathering capacity of large scale

multihop wireless networks. In *Mobile Ad Hoc and Sensor Systems, 2008. MASS 2008. 5th IEEE International Conference on*, pages 124–132. IEEE, 2008.

[75] H. Liu, Y. Zhang, and D. Raychaudhuri. Performance evaluation of the cache-and-forward (CNF) network for mobile content delivery services. In *ICC Workshop*, pages 1–5, 2009.

[76] Mohammad Ali Maddah-Ali and Urs Niesen. Fundamental limits of caching. *IEEE Transactions on Information Theory*, 60(5):2856–2867, 2014.

[77] Valentina Martina, Michele Garetto, and Emilio Leonardi. A unified approach to the performance analysis of caching systems. In *IEEE INFOCOM*, pages 2040–2048, April 2014.

[78] Nick McKeown, Tom Anderson, Hari Balakrishnan, Guru Parulkar, Larry Peterson, Jennifer Rexford, Scott Shenker, and Jonathan Turner. Openflow: enabling innovation in campus networks. *ACM SIGCOMM CCR*, 38(2), March 2008.

[79] Stanley Milgram. The small world phenomenon. *Psychology Today*, 1:61, 1967.

[80] Todor P Mitev. New inequalities between elementary symmetric polynomials. *J. Ineq. Pure and Appl. Math*, 4(2), 2003.

[81] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 54:167 – 256, 2003.

[82] Mark EJ Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.

[83] U. Niesen, D. Shah, and G. Wornell. Caching in wireless networks. *Information Theory, IEEE Transactions on*, 2011.

[84] Urs Niesen, Piyush Gupta, and Devavrat Shah. On capacity scaling in arbitrary wireless networks. *Information Theory, IEEE Transactions on*, 55(9):3959–3982, 2009.

[85] Felipe Olmos, Bruno Kauffmann, Alain Simonian, and Yannick Carlinet. Catalog dynamics: Impact of content publishing and perishing on the performance of a LRU cache. In *IEEE ITC26*, pages 1–9, 2014.

[86] Afif Osseiran, Federico Boccardi, Volker Braun, Katsutoshi Kusume, Patrick Marsch, Michal Maternia, Olav Queseth, Malte Schellmann, Hans Schotten, Hidekazu Taoka, Hugo Tullberg, Mikko A. Uusitalo, Bogdan Timus, and Mikael Fallgren1. Scenarios for 5g mobile and wireless communications: the vision of the metis project. *IEEE Communication Magazine*, 52:26 – 35, 2014.

[87] Mathew D. Penrose. The longest edge of the random minimal spanning tree. *The Annals of Applied Probability*, 7(2):340 – 361, 1997.

[88] L Peterson and B Davie. Framework for cdn interconnection draft-ietf-cdni-framework-08. *Internet draft in Network Working Group of Internet Engineering Task Force (IETF)*, 2014.

[89] Ioannis Psaras, Richard G Clegg, Raul Landa, Wei Koong Chai, and George

Pavlou. Modelling and evaluation of ccn-caching trees. In *International Conference on Research in Networking*, pages 78–91. 2011.

[90] E. J. Rosensweig and J Kurose. Breadcrumbs: Efficient, Best-Effort content location in cag networks. In *IEEE INFOCOM*, pages 2631–2635, 2009.

[91] E.J. Rosensweig, J. Kurose, and D. Towsley. Approximate models for general cache networks. In *IEEE INFOCOM*, pages 1–9, 2010.

[92] Elisha J. Rosensweig and Jim Kurose. A network calculus for cache networks. In *INFOCOM, 2013 Proceedings IEEE*, pages 85–89. IEEE, April 2013.

[93] Dario Rossi and Giuseppe Rossini. Caching performance of content centric networks under multi-path routing (and more). *Relatório técnico, Telecom ParisTech*, 2011.

[94] Dario Rossi and Giuseppe Rossini. On sizing CCN content stores by exploiting topological information. In *Computer Communications Workshops (INFOCOM WKSHPS), 2012 IEEE Conference on*, pages 280–285. IEEE, 2012.

[95] V. Sourlas, P. Flegkas, L. Gkatzikis, and L. Tassiulas. Autonomic cache management in Information-Centric Networks. In *IEEE NOMS*, April 2012.

[96] Bin Tang, Himanshu Gupta, and Samir R Das. Benefit-based data caching in ad hoc networks. *IEEE transactions on Mobile Computing*, 7(3):289–304, 2008.

[97] Amin Tootoonchian, Sergey Gorbunov, Yashar Ganjali, Martin Casado, and Rob

Sherwood. On controller performance in software-defined networks. *Hot-ICE*, 12:1–6, 2012.

[98] M. Tortelli, I. Cianci, L. A. Grieco, G. Boggia, and P. Camarda. A fairness analysis of content centric networks. In *IEEE NOF*, pages 117–121, November 2011.

[99] Dirk Trossen, George Parisis, Kari Visala, Borisalva Gajic, Janne Riihijarvi, Paris Flegkas, Pasi Sarolahti, Petri Jokela, Xenofon Vasilakos, Christos Tsilopoulos, and Somaya Arianfar. PURSUIT Conceptual Architecture: Principles, Patterns and sub-Components Descriptions, May 2011.

[100] Di Wang and A. Abouzeid. Link State Routing Overhead in Mobile Ad Hoc Networks: A Rate-Distortion Formulation. In *IEEE INFOCOM*, pages 1337–1345, April 2008.

[101] Di Wang and A. Abouzeid. On the cost of knowledge of mobility in dynamic networks: An information-theoretic approach. *IEEE Transactions on Mobile Computing*, 11(6):995–1006, June 2012.

[102] Liang Wang, Suzan Bayhan, Jörg Ott, Jussi Kangasharju, Arjuna Sathiaseelan, and Jon Crowcroft. Pro-diluvian: Understanding scoped-flooding for content discovery in information-centric networking. In *ACM ICN*, pages 9–18, 2015.

[103] Zheng Wang, Hamid R Sadjadpour, and JJ Garcia-Luna-Aceves. Fundamental limits of information dissemination in wireless ad hoc networks-part ii: multi-

packet reception. *IEEE Transactions on Wireless Communications*, 10(3):803–813, 2011.

[104] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. *nature*, 393(6684):440–442, 1998.

[105] Cédric Westphal. On maximizing the lifetime of distributed information in ad-hoc networks with individual constraints. In *ACM MobiHoc*, pages 26–33, 2005.

[106] Walter Willinger, D. Alderson, and John C. Doyle. Mathematics and the internet: A source of enormous confusion and great potential. *Notices American Mathematical Society (AMS)*, 56(5):586 – 599, May 2009.

[107] Alec Wolman, M. Voelker, Nitin Sharma, Neal Cardwell, Anna Karlin, and Henry M. Levy. On the scale and performance of cooperative Web proxy caching. *SIGOPS Oper. Syst. Rev.*, 33(5):16–31, December 1999.

[108] Feng Xue and P. R. Kumar. Scaling Laws for Ad Hoc Wireless Networks: an Information Theoretic Approach. In *Foundations and Trends in Networking*, pages 16–47, 2006.

[109] George Xylomenos, Christopher N Ververidis, Vasilios A Siris, Nikos Fotiou, Christos Tsilopoulos, Xenofon Vasilakos, Konstantinos V Katsaros, and George C Polyzos. A survey of information-centric networking research. *IEEE Communications Surveys & Tutorials*, 16(2):1024–1049, 2014.

[110] L. Yin and G. Cao. Supporting cooperative caching in ad hoc networks. *Mobile Computing, IEEE Transactions on*, (1):77–89, 2005.

[111] Minlan Yu, Jennifer Rexford, Michael J. Freedman, and Jia Wang. Scalable flow-based networking with difane. *ACM SIGCOMM CCR*, 41(4), August 2010.

[112] Lixia Zhang, Deborah Estrin, Jeffrey Burke, Van Jacobson, James D Thornton, Diana K Smetters, Beichuan Zhang, Gene Tsudik, Dan Massey, Christos Papadopoulos, et al. Named data networking (ndn) project. *Relatório Técnico NDN-0001, Xerox Palo Alto Research Center-PARC*, 2010.

Appendix A

Detailed Derivation of Equations and Proofs in Part I

A.1 Detailed derivation of equation (5.25)

$$\sum_{\lceil \frac{1}{C_1}+1 \rceil}^{\lceil \frac{2}{C_1 r(n)} \rceil} x \sum_{l=1}^{4x} \sum_{v_k \ in \ s_l} d_k^{-\alpha} \equiv nr^{2-\alpha}(n) \int_{\lceil \frac{1}{C_1}+1 \rceil}^{\lceil \frac{2}{C_1 r(n)} \rceil + 1} u^{2-\alpha} du$$

$$= \frac{nr^{2-\alpha}(n)}{3-\alpha}((\lceil \frac{2}{C_1 r(n)} \rceil + 1)^{3-\alpha} - (\lceil \frac{1}{C_1} + 1 \rceil)^{3-\alpha}) \quad (A.1)$$

If the transmission range decreases with increasing n, then for sufficiently large n, we have

$$\left(\lceil \frac{2}{C_1 r(n)} \rceil + 1 \right)^{3-\alpha} = \Theta(\frac{1}{r^{3-\alpha}(n)}). \quad (A.2)$$

If $\alpha < 3$,[1]

$$
\begin{aligned}
(\lceil \frac{2}{C_1 r(n)} \rceil + 1)^{3-\alpha} - (\lceil \frac{1}{C_1} + 1 \rceil)^{3-\alpha} &\equiv \Theta(\frac{1}{r^{3-\alpha}(n)}) - \Theta(1) \\
&= \Theta(\frac{1}{r^{3-\alpha}(n)}) \qquad \text{(A.3)}
\end{aligned}
$$

Therefore,

$$
\sum_{\lceil \frac{1}{C_1}+1 \rceil}^{\lceil \frac{2}{C_1 r(n)} \rceil} x \sum_{l=1}^{4x} \sum_{v_k \ in \ s_l} d_k^{-\alpha} \equiv \frac{n r^{2-\alpha}(n)}{3-\alpha} \Theta(\frac{1}{r^{3-\alpha}(n)}) \equiv \Theta(\frac{n}{r(n)}) \qquad \text{(A.4)}
$$

For dense social networks in which $\alpha > 3$, we have

$$
\begin{aligned}
& \sum_{\lceil \frac{1}{C_1}+1 \rceil}^{\lceil \frac{2}{C_1 r(n)} \rceil} x \sum_{l=1}^{4x} \sum_{v_k \ in \ s_l} d_k^{-\alpha} \\
\equiv \ & \frac{n r^{2-\alpha}(n)}{\alpha - 3} ((\frac{1}{\lceil \frac{1}{C_1} + 1 \rceil})^{\alpha-3} - (\frac{1}{\lceil \frac{2}{C_1 r(n)} \rceil + 1})^{\alpha-3}), \qquad \text{(A.5)}
\end{aligned}
$$

and for large n

$$
(\frac{1}{\lceil \frac{1}{C_1} + 1 \rceil})^{\alpha-3} - (\frac{1}{\lceil \frac{2}{C_1 r(n)} \rceil + 1})^{\alpha-3} \equiv \Theta(1) - \Theta(r^{\alpha-3}(n)) \equiv \Theta(1) \qquad \text{(A.6)}
$$

Thus, the above summation is equivalent to

$$
\frac{n r^{2-\alpha}(n)}{\alpha - 3} \Theta(1) \equiv \Theta(\frac{n}{r^{\alpha-2}(n)}). \qquad \text{(A.7)}
$$

A.2 Detailed derivation of equation (5.27)

For large n with minimum transmission range, we have

$$
\sigma_{1,n} \equiv \int_{r(n)}^{\gamma d_{max}} n u^{1-\alpha} du = \frac{n}{2-\alpha} ((\gamma d_{max})^{2-\alpha} - r^{2-\alpha}(n)) \qquad \text{(A.8)}
$$

[1] Note that for $\alpha = 3$, both Cases I and II give the same result.

For $\alpha < 2$ and small $r(n)$, we arrive at $\sigma_{1,n} \equiv \frac{n}{2-\alpha}(\gamma d_{max})^{2-\alpha} \equiv \Theta(n)$.

And for $\alpha > 2$, $\sigma_{1,n}$ is

$$\begin{aligned} \sigma_{1,n} &\equiv \frac{1}{r^{2-\alpha}(n)} - (\frac{1}{\gamma d_{max}})^{\alpha-2}) \\ &\equiv \frac{n}{(\alpha-2)r^{2-\alpha}(n)} \equiv \Theta(\frac{n}{r^{\alpha-2}(n)}) \end{aligned} \tag{A.9}$$

A.3 Detailed derivation of equation (5.2)

Based on the computations in section 5 and if $\lim_{n\to\infty} q = \infty$, the probability of a node v_k at distance $d_k = x$ away from the source being the destination (equation (3.12)) will be

$$\frac{x^{-\alpha}\sigma^{\overline{x}}_{q-1,n-1}}{q\sigma_{q,n}} \equiv \frac{1}{n}. \tag{A.10}$$

Thus equation (5.1) equals to

$$Pr(d_{st} = \Theta(1)) \quad \equiv \quad \int_{D_1}^{D_2} x dx = \Theta(1). \tag{A.11}$$

For networks with finite number of contacts per node, $\lim_{n\to\infty} q < \infty$, the probability that a node at distance x is selected as the destination is $\frac{nx^{-\alpha}}{(n-q+1)\sigma_{1,n}}$ (equation (5.22)). By replacing this value in equation (5.1) we have

$$\begin{aligned} Pr(d_{st} = \Theta(1)) &\equiv \frac{n}{n-q+1}\int_{D_1}^{D_2} \frac{nx^{1-\alpha}}{\sigma_{1,n}} dx \\ \equiv \Theta(\frac{n}{\sigma_{1,n}}) &\equiv \begin{cases} 1 & for\ 0 \le \alpha \le 2 \\ \sqrt{\frac{\log n}{n}}^{\alpha-2} & for\ 2 \le \alpha \end{cases} \end{aligned} \tag{A.12}$$

As can be seen, for both finite and infinite values of q when $0 \leq \alpha \leq 2$, with high probability the destinations are at distance of $\Theta(1)$ from the sources, while for concentrated social networks($2 \leq \alpha$) with finite q, this probability is asymptotically negligible.

A.4 Detailed calculation of E_2

Here we calculate the upper bound for E_2. $d_k^{-\alpha}\sigma_{q-1,n-1}^{\overline{k}}(v)$ in (5.53) can be written as

$$\begin{aligned} d_k^{-\alpha}\sigma_{q-1,n-1}^{\overline{k}}(v) &= d_k^{-\alpha}(\sigma_{q-1,n}(v) - d_k^{-\alpha}\sigma_{q-2,n-1}^{\overline{k}}(v)), \\ &= d_k^{-\alpha}(\sigma_{q-1,n}(v) - d_k^{-\alpha}(\sigma_{q-2,n}(v) - d_k^{-\alpha}\sigma_{q-3,n-1}^{\overline{k}}(v))). \end{aligned} \tag{A.13}$$

Since $d_k^{-\alpha}$ and $\sigma_{q-2,n-1}$ are positive values, $d_k^{-\alpha}\sigma_{q-1,n}(v)$ provides an upper bound for $d_k^{-\alpha}\sigma_{q-1,n-1}^{\overline{k}}(v)$.

Lemma A.4.1. *[80] Let $\mathbf{\Psi} = \{\psi_1, ..., \psi_n\}$ be a set of $n \geq 2$ non-negative real numbers. Then for $1 \leq p \leq n-1$ we have $\sigma_1(\mathbf{\Psi})\sigma_p(\mathbf{\Psi}) \geq \frac{n(p+1)}{n-p}\sigma_{p+1}(\mathbf{\Psi})$.*

In Section 5 (Lemma 5.1.3), we prove that this is a tight bound for values of p that do not grow as fast as n.

According to Lemma A.4.1, p requires to be very small such that $\frac{n}{p}$ is sufficiently large and we use Law of Large Numbers. Therefore, we select q_0 such that when n goes to infinity and for $q \leq q_0$, we have

$$\frac{\sigma_{q-1,n}(v)}{\sigma_{q,n}(v)} \equiv \frac{nqd_k^{-\alpha}}{(n-q+1)\sigma_{1,n}(v)}. \tag{A.14}$$

By incorporating the upper bound of (A.13) into (5.53), we arrive at

$$E_2 < \sum_{x=1}^{\frac{1}{r(n)}} x \sum_{l=1}^{4x} \sum_{v_k \ in \ s_l} \sum_{q=1}^{q_0} \frac{q^{-\gamma-1}}{\sum_{b=1}^{n-1} b^{-\gamma}} \frac{nqd_k^{-\alpha}}{(n-q+1)\sigma_{1,n}(v)}. \tag{A.15}$$

It can be easily seen that the order of the distances between the source and all the nodes inside s_l for all $l = 1, .., 4x$ equals to $xr(n)$ and the number of nodes with a distance of x equals to $\Theta(xnr^2(n))$, thus

$$E_2 < \frac{n^2 r^{2-\alpha}(n)}{\sigma_{1,n}(v) \sum_{b=1}^{n-1} b^{-\gamma}} \sum_{x=1}^{\frac{1}{r(n)}} x^{2-\alpha} \sum_{q=1}^{q_0} \frac{q^{-\gamma}}{n-q+1}. \tag{A.16}$$

Since q_0 is not growing with n, then $n - q + 1 = \Theta(n)$ for $q \leq q_0$. Therefore,

$$E_2 < \frac{nr^{2-\alpha}(n)}{\sigma_{1,n}(v) \sum_{b=1}^{n-1} b^{-\gamma}} \sum_{x=1}^{\frac{1}{r(n)}} x^{2-\alpha} \sum_{q=1}^{q_0} q^{-\gamma}. \tag{A.17}$$

According to equation (5.27) the order of E_2, the average number of hops for sources with low number of social connections is

$$E_2 < \frac{1}{\sum_{b=1}^{n-1} b^{-\gamma}} \sum_{q=1}^{q_0} q^{-\gamma} \begin{cases} \frac{1}{r(n)} & for \ 0 \leq \alpha \leq 2 \\ \frac{1}{r^{3-\alpha}(n)} & for \ 2 \leq \alpha \leq 3 \\ 1 & for \ 3 \leq \alpha \end{cases} . \tag{A.18}$$

By utilizing (A.13) and the results from section 5.1, we arrive at

$$\frac{\sigma_{q-2,n}(v)}{q\sigma_{q,n}(v)} = \Theta(\frac{n^2}{\sigma_{1,n}^2(v)}\frac{q-1}{(n-q+1)(n-q+2)}). \tag{A.19}$$

We can use similar calculations to prove that this upper bound is actually a tight bound. Therefore for $\gamma > 1$.

$$E_2 = \begin{cases} \Theta(\frac{1}{r(n)}) & for\ 0 \leq \alpha \leq 2 \\ \Theta(\frac{1}{r^{3-\alpha}(n)}) & for\ 2 \leq \alpha \leq 3 \\ \Theta(1) & for\ 3 \leq \alpha \end{cases} \tag{A.20}$$

It is obvious that for $\alpha > 2$, the value of E_2 is much less than E_1. Under this condition, a small number of nodes ($N_{>q_0}$) are using a large portion of the resources and are limiting the total throughput of the network.

A.5 Proof of Lemma 5.3.3

Proof. Define the random variables $Y_i = d_i^{-\alpha}$ and $Z_i = \log Y_i$ for $1 \leq i \leq n$. Since Y_i's are i.i.d random variables, Z_i's are also i.i.d. random variables. By using the law of large numbers, we have $\lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} Z_i = \bar{Z}$ where $\bar{Z}$ is the expected value of the random variable Z_i. Hence,

$$\begin{aligned}\frac{d_k^{-\alpha}\sigma_{q-1}(\mathbf{d}_{\mathbf{n}}^{\bar{\mathbf{k}}})}{\sigma_q(\mathbf{d}_{\mathbf{n}})} &\equiv \frac{\sum_{1\le i_1<..<i_q\le n,\exists h:i_h=k}\prod_{j=1}^{q} Y_{i_j}}{\sum_{1\le i_1<..<i_q\le n}\prod_{j=1}^{q} Y_{i_j}} \\ &\equiv \frac{\sum_{1\le i_1<..<i_q\le n,\exists h:i_h=k}\exp\sum_{j=1}^{q} Z_{i_j}}{\sum_{1\le i_1<..<i_q\le n}\exp\sum_{j=1}^{q} Z_{i_j}} \\ &\equiv \frac{\sum_{1\le i_1<..<i_q\le n,\exists h:i_h=k}\exp q\overline{Z}}{\sum_{1\le i_1<..<i_q\le n}\exp q\overline{Z}} \equiv \frac{\binom{n-1}{q-1}}{\binom{n}{q}} = \frac{q}{n} \qquad \text{(A.21)}\end{aligned}$$

Now if $q = \Theta(n)$, we have $\frac{q}{n} \equiv \Theta(1)$ and therefore, $\frac{d_k^{-\alpha}\sigma_{q-1}(\mathbf{d}_{\mathbf{n}}^{\bar{\mathbf{k}}})}{\sigma_q(\mathbf{d}_{\mathbf{n}})} \equiv \frac{q}{n} \equiv \Theta(1)$. □

A.6 Proof of Lemma 5.3.6

Proof. Suppose that the i-th member of the long-range social group is located in the distance of x_{q_i} hops from the source, then we can say that

$$\sigma_1(\mathbf{d_q}) = \sum_{i=1}^{q} d_{q_i}^{-\beta} = \sum_{i=1}^{q} (c_i r(n) x_{q_i})^{-\beta} = (r(n))^{-\beta}\sum_{i=1}^{q}(c_i x_{q_i})^{-\beta}. \qquad \text{(A.22)}$$

Since, x_{q_i} can be every integer between one and $\frac{1}{r(n)}$, the order of $\sigma_1(\mathbf{d_q})$ may range from $\Theta(1)$ to $\Theta(r(n)^{-\beta})$. However, note that when n goes to infinity, with probability approaching one at least one of the long-range contacts lies within a lattice distance of $\Theta(1)$ to the source.

To prove this, it is enough to show that with probability approaching zero, all of the long-range contacts lie outside a lattice distance of $f(n) = \Omega(1)$ to the source. Assuming $q = \Theta(1)$ or $q = \Theta(f(n))$ where $\lim_{n\to\infty}\frac{f(n)}{n} = 0$, we can argue that the probability of selecting long-range social contacts is independent of each other. Thus, using lemma 5.3.5 we have

$$\begin{aligned}
\Pr(x_{q_1} = \Theta(f(n)), x_{q_2} &= \Theta(f(n)), ..., x_{q_q} = \Theta(f(n))) \\
&= \prod_{i=1}^{q} \Pr(x_{q_i}\Theta(f(n)) \\
&\equiv \prod_{i=1}^{q} \frac{(f(n)r(n))^{-\alpha}}{\sigma_1(\mathbf{d_n})} \\
&\equiv O(\frac{(r(n))^{-q\alpha}}{(\sigma_1(\mathbf{d_n}))^q}) \\
&= \begin{cases} O\left((nr^{\alpha}(n))^{-q}\right), & 0 \leq \alpha \leq 2 \\ O\left((\log n)^{-q}\right). & 2 \leq \alpha \end{cases}
\end{aligned} \tag{A.23}$$

It is not difficult to verify that the right hand side which is an upper bound for this probability goes to zero as n approaches infinity thus the aforementioned probability tends to zero. Thus with probability approaching one, there exists at least one long-range contact in the lattice distance of $\Theta(1)$ to the source which will be the dominant term in $\sigma_1(\mathbf{d_q})$. Therefore, $\sigma_1(\mathbf{d_q}) = \Omega\left((r(n))^{-\beta}\right)$ and since in the case of $q = \Theta(1)$, $\sigma_1(\mathbf{d_q})$ is only composed of $\Theta(1)$ terms we have $\sigma_1(\mathbf{d_q}) = \Theta\left((r(n))^{-\beta}\right)$.

For the case of $q = \Theta(f(n))$ we know that at least one long range contact exists within a distance of $\Theta(1)$ to the source. Therefore, $\sigma_1(\mathbf{d_q})$ can have the order of $\Theta\left((r(n))^{-\beta}\right)$ when it only has $\Theta(1)$ social contacts within a distance of $\Theta(1)$ to the source or it can have the order of $\Theta\left(f(n)(r(n))^{-\beta}\right)$ when almost all of the $\Theta(f(n))$ social contacts lie within a distance of $\Theta(1)$ to the source. We will now show that with a probability close to one the latter almost never happens and therefore in the case of

$q = \Theta(f(n))$, almost surely we have, $\sigma_1(\mathbf{d_q}) = \Theta\left((r(n))^{-\beta}\right)$. To prove this, using the same approach as above, we will compute the probability that almost all of the social contacts lie within a distance $\Theta(1)$ to the source,

$$\begin{aligned}
\Pr(x_{q_1} = \Theta(1), x_{q_2} = \Theta(1), ..., x_{q_q} = \Theta(1)) &= \prod_{i=1}^{q} \Pr(x_{q_i} = \Theta(1)) \\
&\equiv \prod_{i=1}^{q} \frac{(r(n))^{-\alpha}}{\sigma_1(\mathbf{d_n})} \\
&\equiv \Theta(\frac{(r(n))^{-q\alpha}}{(\sigma_1(\mathbf{d_n}))^q}) \\
&= \begin{cases} \Theta\left((nr^{\alpha}(n))^{-q}\right), & 0 \le \alpha \le 2 \\ \\ \Theta\left((\log n)^{-q}\right). & 2 \le \alpha \end{cases}
\end{aligned} \tag{A.24}$$

Clearly, when n is a large number, this probability goes to zero and therefore this scenario almost surely never happens. □

A.7 Proof of Lemma 5.3.7

Proof. This lemma can be proved by expanding the polynomials and considering the non-negativity of elements in $\mathbf{d_n}$. We will use this lemma to find the upper and lower bounds for $E[X]$. □

Appendix B

Detailed Derivation of Equations and Proofs in Part II

B.1 Proof of Lemma 10.1.3

Proof. 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

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

This case can be considered as a special case of the network studied in Theorem 10.1.1, where $\rho_i^{(k)}(n)$ is the same for all i[1]. Thus, we can drop the index i and let $\rho^{(k)}(n)$

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

denote the common value of this probability. Using equation (10.2) and (10.3) we will have $E[h_k]$ equal to

$$\frac{4}{n}\sum_{i=1}^{\sqrt{n}} i\{i(1-\rho^{(k)}(n))^i + \sum_{j=1}^{i-1}(i-j)(1-\rho^{(k)}(n))^{i-j}\rho^{(k)}(n)\} \tag{B.2}$$

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

B.2 Proof of Lemma 10.2.2

Proof. 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 $(1-\rho^{(k)}(n))^{1+4(1+2+\ldots+(i-1))} = (1-\rho^{(k)}(n))^{2i^2-2i+1}$, 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 equations (B.1), (10.2), and (10.3),

$E[h_k]$ is equal to

$$\frac{1}{n}\sum_{i=2}^{\sqrt{n}} i \sum_{l=1}^{i-1} l(1-(1-\rho^{(k)}(n))^{4l})(1-\rho^{(k)}(n))^{2l^2-2l+1} + \frac{1}{n}\sum_{i=1}^{\sqrt{n}} i^2(1-\rho^{(k)}(n))^{2i^2-2i+1} \tag{B.3}$$

□

B.3 Proof of Lemma 10.3.2

Proof. 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 (B.1) $E[h_k]$ equals to

$$
\begin{aligned}
r^2(n)\{(1-\rho^{(k)}(n)) \quad &+ \quad \sum_{i=2}^{\frac{1}{r(n)}} i^2(1-\rho^{(k)}(n))^{inr^2(n)} \\
&+ \quad (1-(1-\rho^{(k)}(n))^{nr^2(n)}) \sum_{i=2}^{\frac{1}{r(n)}} i \sum_{l=1}^{i-1} l(1-\rho^{(k)}(n))^{lnr^2(n)}\}.
\end{aligned}
\tag{B.4}
$$

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. □

B.4 Proof of Lemma 10.1.4

Proof. 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).

Let us define

$$
E_s^i = \frac{1}{n}\sum_{i=1}^{\sqrt{n}} i^2(1-\rho(n))^i \quad , \quad E_c^i = \frac{\rho(n)}{n}\sum_{i=1}^{\sqrt{n}} i \sum_{l=1}^{i-1} l(1-\rho(n))^l. \tag{B.5}
$$

Thus equation (10.5) 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.

$$E_s^i \equiv \frac{1}{n}\{\sum_{i \prec \sqrt{n}} i^2(1-\rho(n))^i + \sum_{i \equiv \sqrt{n}} i^2(1-\rho(n))^i\} \tag{B.6}$$

Assume $\rho(n) \equiv \frac{1}{\sqrt{n}}$, then using first and second region of equation (B.4) we have

$$E_s^i \equiv \frac{1}{n}\{\sum_{i \prec \sqrt{n}} i^2 + \sum_{i \equiv \sqrt{n}} i^2\} \equiv \frac{n^{3/2}}{n} \equiv \sqrt{n}. \tag{B.7}$$

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

$$\begin{aligned} E_s^i &\equiv \frac{(1-\rho(n))(2-\rho(n))}{n\rho^3(n)} \\ &- \frac{(1-\rho(n))^{\sqrt{n}+1}}{n\rho^3(n)}\{n(1-\rho(n))^2 - (1-\rho(n))(2n+2\sqrt{n}-1) + (\sqrt{n}+1)^2\} \end{aligned} \tag{B.8}$$

If $\rho(n) \succ \frac{1}{\sqrt{n}}$, then using third region in equation (B.4), $(1-\rho(n))^{\sqrt{n}+1}$ 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

$$E_s^i \equiv \begin{cases} \sqrt{n} & \rho(n) \preceq \frac{1}{\sqrt{n}} \\ \frac{1}{n\rho^3(n)} & \rho(n) \succ \frac{1}{\sqrt{n}} \end{cases} \tag{B.9}$$

According to equation (B.9) 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 $\rho(n) \succ \frac{1}{\sqrt{n}}$.

$$
\begin{aligned}
E_c^{i1} &= \frac{1}{n} \sum_{i \equiv \sqrt{n}} i \sum_{l=1}^{i-1} l\rho(n)(1-\rho(n))^l, \\
E_c^{i2} &= \frac{1}{n} \sum_{i \prec \sqrt{n}} i \sum_{l=1}^{i-1} l\rho(n)(1-\rho(n))^l
\end{aligned}
\tag{B.10}
$$

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

$$
\begin{aligned}
E_c^{i1} &\equiv \frac{1}{\sqrt{n}} \sum_{l=1}^{\sqrt{n}} l\rho(n)(1-\rho(n))^l \\
&\equiv \frac{1-\rho(n)}{\rho(n)\sqrt{n}}(1-(1-\rho(n))^{\sqrt{n}}(1+\rho(n)\sqrt{n}))
\end{aligned}
\tag{B.11}
$$

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

$$
\begin{aligned}
E_c^{i2} &\equiv \frac{1-\rho(n)}{\rho(n)n} \sum_{i \prec \sqrt{n}} i[1-\{1-\rho(n)+\rho(n)i\}(1-\rho(n))^{i-1}] \\
&\equiv \frac{1-\rho(n)\{1-\frac{1}{n}\sum\limits_{i \prec \sqrt{n}} i(1-\rho(n))^i - \frac{1}{n}\sum\limits_{i \prec \sqrt{n}} i^2\rho(n)(1-\rho(n))^{i-1}\}}{\rho(n)} \\
&\equiv \frac{1-\rho(n)}{\rho(n)} - \frac{(1-\rho(n))^2}{\rho^3(n)n} - \frac{1}{\rho^3(n)n} \equiv \frac{1}{\rho(n)}
\end{aligned}
\tag{B.12}
$$

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 $\rho(n) \succ \frac{1}{\sqrt{n}}$. □

B.5 Proof of Lemma 10.2.3

Proof. Let us define

$$
\begin{aligned}
E_s^{ii} &= \frac{1}{n}\sum_{i=1}^{\sqrt{n}} i^2(1-\rho(n))^{2i^2-2i+1}, \\
E_c^{ii} &= \frac{1}{n}\sum_{i=2}^{\sqrt{n}} i \sum_{k=1}^{i-1} l(1-\rho(n))^{2l^2-2l+1}(1-(1-\rho(n))^{4l})
\end{aligned}
\qquad \text{(B.13)}
$$

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

$$
E_s^{ii} \equiv \frac{1}{n}\sum_{i=1}^{\sqrt{n}} i^2(1-\frac{1}{n})^{2i^2-2i+1} \equiv \frac{1}{n}\sum_{i=1}^{\sqrt{n}} i^2 \equiv \sqrt{n}. \qquad \text{(B.14)}
$$

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.

$$
\begin{aligned}
E_s^{ii} &\equiv \frac{1}{n}\int_{v=1}^{\sqrt{n}} v^2(1-\rho(n))^{2v^2-2v+1} \\
&\equiv \{\frac{(1-\log(1-\rho(n)))\sqrt{2\pi(1-\rho(n))}erf(\frac{(2v-1)\sqrt{-\log(1-\rho(n))}}{\sqrt{2}})}{n\log^{3/2}(1-\rho(n))} \\
&+ \frac{-2\sqrt{-\log(1-\rho(n))}(2v+1)(1-\rho(n))^{2v^2-2v+1}}{n\log^{3/2}(1-\rho(n))}\}|_{v=1}^{\sqrt{n}}
\end{aligned}
\qquad \text{(B.15)}
$$

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 (B.4), 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 $\rho(n) \succ \frac{1}{n}$. □

B.6 Proof of Lemma 10.3.3

Proof. Let us define $E[h] = E_s^{iii} + E_c^{iii}$, where

$$\begin{aligned} E_s^{iii} = & \ r^2(n) \sum_{i=2}^{\frac{1}{r(n)}} i^2 (1-\rho(n))^{inr^2(n)} \\ E_c^{iii} = & \ r^2(n)(1-(1-\rho(n))^{nr^2(n)}) \{ \sum_{i=2}^{\frac{1}{r(n)}} i \sum_{l=1}^{i-1} l(1-\rho(n))^{lnr^2(n)} \} \end{aligned} \tag{B.16}$$

First we check the behavior of E_s^{iii} when $\rho(n) \equiv \frac{1}{nr(n)}$. Using the second region in equation (B.4) 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

$$E_s^{iii} \equiv r^2(n) \int_2^{\frac{1}{r(n)}} v^2 (1-\rho(n))^{vnr^2(n)} dv, \tag{B.17}$$

which equals to

$$\{(v^2 \log^2 (1-\rho(n))^{nr^2(n)} - 2v \log (1-\rho(n))^{nr^2(n)} + 2) \frac{r^2(n)(1-\rho(n))^{vnr^2(n)}}{\log^3 (1-\rho(n))^{nr^2(n)}} \}|_{v=2}^{\frac{1}{r(n)}}. \tag{B.18}$$

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 (B.4), 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 (B.18) 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,

$$E[h] \equiv \begin{cases} E_s^{iii} \equiv \frac{1}{r(n)} & \rho(n) \preceq \frac{1}{nr(n)} \\ E_c^{iii} \equiv \frac{1}{\rho(n)nr^2(n)} & \frac{1}{nr(n)} \preceq \rho(n) \preceq \frac{1}{nr^2(n)} \\ E_c^{iii} \equiv 1 & \frac{1}{nr^2(n)} \preceq \rho(n) \end{cases} \tag{B.19}$$

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., $\lim_{n\to\infty}(1-(1-\rho(n))^{nr^2(n)}) = 1$. □

B.7 Proof of Lemmas 10.1.6, 10.2.5, and 10.3.5

Proof. 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 Lemmas 10.1.6, 10.2.5, and $(\Theta(\frac{1}{r^2(n)}))$ in Lemma 10.3.5. Thus, $nE[h]\gamma \preceq n$ and $\gamma_{max} \equiv \frac{1}{E[h]}$ in Lemmas 10.1.6, and 10.2.5, and $nE[h]\gamma \preceq \frac{1}{r^2(n)}$ and $\gamma_{max} \equiv \frac{1}{E[h]nr^2(n)}$ Lemma 10.3.5. Therefore the maximum download rate is easily derived using the results of Lemmas 10.1.6, 10.2.5, and 10.3.5. □

B.8 Proof of Lemma 10.1.7

Proof. Each link between two nodes 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.

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 (B.4) the above summation can be written as

$$\gamma \frac{(-1+\rho^{(k)}(n))(\sqrt{n}\rho^{(k)}(n)(1-\rho^{(k)}(n))^{\sqrt{n}} + (1-\rho^{(k)}(n))^{\sqrt{n}} - 1)}{(\rho^{(k)}(n))^2} \equiv \frac{\gamma}{(\rho^{(k)}(n))^2}. \tag{B.20}$$

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 ($\psi \equiv \Theta(\frac{\gamma}{\underset{k}{min}(\rho^{(k)}(n))^2)})$), 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, $\psi \equiv min(n\gamma, \frac{\gamma}{\underset{k}{min}(\rho^{(k)}(n))^2)}) \preceq 1$, then $\gamma_{max} \equiv max(\frac{1}{n}, \underset{k}{min}((\rho^{(k)}(n))^2))$.

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. □

B.9 Proof of Lemma 10.2.6

Proof. Each link between two nodes 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.

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

$$\begin{aligned}
& \gamma(1-\rho^{(k)}(n)) + \sum_{i=1}^{\sqrt{n}} 4\gamma i(1-\rho^{(k)}(n))^{(1+4\sum_{j=1}^{i} j)} \\
\equiv\ & \gamma(1-\rho^{(k)}(n)) + \sum_{i=1}^{\sqrt{n}} \gamma i(1-\rho^{(k)}(n))^{2i^2+2i+1}, \\
\equiv\ & \gamma\{(1-\rho^{(k)}(n)) + \frac{(1-\rho^{(k)}(n))^n - (1-\rho^{(k)}(n))^4}{\log(1-\rho^{(k)}(n))/(1-\rho^{(k)}(n))} \\
+\ & \frac{\sqrt{-\frac{\log(1-\rho^{(k)}(n))}{1-\rho^{(k)}(n)}}erf(\sqrt{-n\log(1-\rho^{(k)}(n))})}{\log(1-\rho^{(k)}(n))/(1-\rho^{(k)}(n))} \\
-\ & \frac{\sqrt{-\frac{\log(1-\rho^{(k)}(n))}{1-\rho^{(k)}(n)}}erf(\sqrt{-\log(1-\rho^{(k)}(n))})}{\log(1-\rho^{(k)}(n))/(1-\rho^{(k)}(n))}\}.
\end{aligned} \tag{B.21}$$

If $\rho^{(k)}(n) \preceq \frac{1}{n}$, then $(1-\rho^{(k)}(n))^{2i^2+2i+1} \equiv 1$ for all $1 \le i \le \sqrt{n}$. 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 $\psi \equiv \frac{\gamma}{\underset{k}{min}(\rho^{(k)}(n))}$. If $\rho^{(k)}(n) \preceq \frac{1}{n}$ for at least one item, then $\psi \equiv n\gamma$. Thus, $\psi \equiv min(n\gamma, \frac{\gamma}{\underset{k}{min}(\rho^{(k)}(n))}) \preceq 1$, then $\gamma_{max} \equiv max(\frac{1}{n}, \underset{k}{min}(\rho^{(k)}(n)))$.

Using similar reasoning as in Lemma 10.1.7 other links carry less traffic, so the above capacities are supportable for all the other links. □

B.10 Proof of Lemma 10.3.6

Proof. Each link between two nodes 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.

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

$$\begin{aligned} & \gamma nr^2(n)\{(1-\rho^{(k)}(n)) + \sum_{i=2}^{\frac{1}{r(n)}} i(1-\rho^{(k)}(n))^{inr^2(n)}\} \\ \equiv\ & \gamma nr^2(n)\{(1-\rho^{(k)}(n)) \\ +\ & \frac{(1-\rho^{(k)}(n))^{nr(n)}(nr(n)\log(1-\rho^{(k)}(n))-1)}{\log^2(1-\rho^{(k)}(n))^{nr^2(n)}} \\ -\ & \frac{(1-\rho^{(k)}(n))^{nr^2(n)}(\log(1-\rho^{(k)}(n))^{nr^2(n)}-1)}{\log^2(1-\rho^{(k)}(n))^{nr^2(n)}}\} \end{aligned} \tag{B.22}$$

If $\rho^{(k)}(n) \preceq \frac{1}{nr(n)}$, then $(1-\rho^{(k)}(n))^{inr^2(n)} \to 1$ for $2 \le i \le \frac{1}{r(n)}$, 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

$$\gamma nr^2(n) + \gamma nr^2(n)\frac{1+2\rho^{(k)}(n)nr^2(n)}{(\rho^{(k)}(n))^2n^2r^4(n)} \equiv \frac{\gamma}{(\rho^{(k)}(n))^2nr^2(n)} \tag{B.23}$$

If $\rho^{(k)}(n) \succeq \frac{1}{nr^2(n)}$, then equation (B.22) 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, $\psi \equiv min[n\gamma,\ max(\gamma nr^2(n), \frac{\gamma}{\underset{k}{min}(\rho^{(k)}(n))^2nr^2(n)})] \preceq 1$, $\gamma_{max} \preceq max[\frac{1}{n}, min(\frac{1}{nr^2(n)}, \underset{k}{min}((\rho^{(k)}(n))^2)nr^2(n))]$. 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 $\Theta(\frac{1}{n})$. □

B.11 Proof of Lemma 12.1.1

Proof. The distortion criteria is defined as

$$\begin{aligned} D_1 &= Pr(S_X = 1, \hat{S}_X = 0) \leq \epsilon_1 \\ D_2 &= Pr(S_X = 0, \hat{S}_X = 1) \leq \epsilon_2 \end{aligned} \tag{B.24}$$

It can be seen that $Pr(S_X = 1) = \frac{\tau_X}{\tau_X + \theta_X}$, and $Pr(S_X = 0) = \frac{\theta_X}{\tau_X + \theta_X}$. There are three cases where the distortion criteria is satisfied even when the controller has no information about the underlying plane.

1. If the monitoring state is 'down' with high probability ($Pr(S_X = 1) \leq \epsilon_1$), then having the controller assume that it is always 'down' (keeping $\hat{S}_X$ constantly equal to $'0'$) will satisfy the distortion criteria ($D_1 = Pr(S_X = 1) \leq \epsilon_1$ and $D_2 = 0 < \epsilon_2$).

2. If the monitoring state is 'up' with high probability ($Pr(S_X = 0) \leq \epsilon_2$), then setting the controller to assume it is always 'up' (keeping $\hat{S}_X$ constantly equal to $'1'$) will satisfy the distortion criteria ($D_1 = 0 < \epsilon_1$ and $D_2 = Pr(S_X = 0) \leq \epsilon_2$).

3. If the monitoring variable can take both 'up' and 'down' states with high enough probabilities such that $1 - \frac{\epsilon_1}{Pr(S_X=1)} \leq \frac{\epsilon_2}{Pr(S_X=0)}$, then we pick a value ρ_0 between $1 - \frac{\epsilon_1}{Pr(S_X=1)}$ and $\frac{\epsilon_2}{Pr(S_X=0)}$, and assign $'1'$ to $\hat{S}_X$ with probability ρ_0 independent of the value of S_X. Therefore, since $D_1 = Pr(S_X = 1)Pr(\hat{S}_X = 0) = Pr(S_X = 1)(1 - \rho_0) \leq \epsilon_1$, and $D_2 = Pr(S_X = 0)Pr(\hat{S}_X = 1) = \rho_0 Pr(S_X = 0) \leq \epsilon_2$, the distortion criteria is satisfied.

Thus in the following, we concentrate on the cases where $Pr(S_X = 1) > \epsilon_1$, $Pr(S_X = 0) > \epsilon_2$, and $1 - \frac{\epsilon_1}{Pr(S_X=1)} > \frac{\epsilon_2}{Pr(S_X=0)}$.

Note that we assume that $\epsilon_1 + \epsilon_2 \leq 1$, then $\frac{\epsilon_2}{1-\epsilon_2} \leq \frac{1-\epsilon_1}{\epsilon_1}$, and the first two regions can be summarized in the region where $\frac{\epsilon_2}{1-\epsilon_2} \leq \frac{\theta_X}{\tau_X} \leq \frac{1-\epsilon_1}{\epsilon_1}$. The third region is also mapped to the region where $\epsilon_2\tau_X + \epsilon_1\theta_X < \frac{\tau_X\theta_X}{\tau_X+\theta_X}$.

Let $U_X^1(\epsilon_1)$ (and $U_X^2(\epsilon_2)$) denote the needed update rate per change type I (and II), or in other words the ratio of times that type I (and II) changes have to be reported to the control plane so that the distortion criteria is satisfied. As can be seen in figure 12.1, each 'up' period Z_m starts at time T_{2m-1} and ends at time T_{2m}. The false negative alarm is generated during the m^{th} 'up' period (Z_m) if a type I change in the state of X at time T_{2m-1} is not announced to the control plane while the previous state ('0') was correctly perceived by the control plane; we show this event by W_m^1, and its probability is given by

$$\begin{aligned} Pr(W_m^1) &= (1 - U_X^1(\epsilon_1))Pr(\hat{S}_X = 0|S_X = 0) \\ &= (1 - U_X^1(\epsilon_1))(1 - Pr(\hat{S}_X = 1|S_X = 0)) \\ &= (1 - U_X^1(\epsilon_1))(1 - \frac{Pr(S_X = 0, \hat{S}_X = 1)}{Pr(S_X = 0)} \\ &= (1 - U_X^1(\epsilon_1))(1 - D_2\frac{\tau_X + \theta_X}{\theta_X}) \qquad \text{(B.25)} \end{aligned}$$

In this case, $\hat{S}_X = 0$ during the time where $S_X = 1$. So assuming that the m^{th} such change is perceived wrong by the control plane, Z_m is the time interval where the control plane has the type I wrong information about the state of X. Let N_w be the number of times S_X undergoes type I changes during a time interval $[0, w]$. The probability of type I error, and consequently type I distortion can be calculated as the ratio of total time of type I error over w when $w \to \infty$.

$$\begin{aligned} D_1 &= E[\frac{1}{w}\sum_{m=1}^{N_w} 1_{[W_m^1]} Z_m] \\ &= \frac{1}{w} E[1_{[W_m^1]} Z_m] E[N_w] \\ &= \frac{\tau_X}{\tau_X + \theta_X} Pr(W_m^1) \\ &= \frac{\tau_X}{\tau_X + \theta_X}(1 - U_X^1(\epsilon_1))(1 - D_2 \frac{\tau_X + \theta_X}{\theta_X}) \end{aligned} \tag{B.26}$$

Similarly, a false positive alarm is generated when a type II change is not announced while the previous perceived state ('1') was correct, and assuming that this is the m^{th} such change, Y_{m+1} is the time interval that the control plane has type II wrong information about X; let W_m^2 denote this event. Thus,

$$\begin{aligned} Pr(W_m^2) &= (1 - U_X^2(\epsilon_2)) Pr(\hat{S}_X = 1 | S_X = 1) \\ &= (1 - U_X^2(\epsilon_2)) \frac{Pr(S_X = 1) - Pr(S_X = 1, \hat{S}_X = 0)}{Pr(S_X = 1)} \\ &= (1 - U_X^2(\epsilon_2))(1 - D_1 \frac{\tau_X + \theta_X}{\tau_X}) \end{aligned} \tag{B.27}$$

and

$$\begin{aligned} D_2 &= E[\frac{1}{w}\sum_{m=1}^{N_w} 1_{[W_m^2]} Z_m] \\ &= \frac{1}{w} E[1_{[W_m^2]} Y_{m+1}] E[N_w] \\ &= \frac{\theta_X}{\tau_X + \theta_X} Pr(W_m^2) \\ &= \frac{\theta_X}{\tau_X + \theta_X}(1 - U_X^2(\epsilon_2))(1 - D_1 \frac{\tau_X + \theta_X}{\tau_X}) \end{aligned} \tag{B.28}$$

To satisfy the distortion criteria we need $D_1 \leq \epsilon_1$ and $D_2 \leq \epsilon_2$. The update rates per changes type I and II, $U_X^1(\epsilon_1)$ and $U_X^2(\epsilon_2)$, then can be written as

$$U_X^1(\epsilon_1) = 1 - \frac{D_1 \frac{\theta_X}{\tau_X}}{\frac{\theta_X}{\tau_X+\theta_X} - D_2} \geq 1 - \frac{\epsilon_1 \frac{\theta_X}{\tau_X}}{\frac{\theta_X}{\tau_X+\theta_X} - \epsilon_2} \tag{B.29}$$

$$U_X^2(\epsilon_2) = 1 - \frac{D_2 \frac{\tau_X}{\theta_X}}{\frac{\tau_X}{\tau_X+\theta_X} - D_1} \geq 1 - \frac{\epsilon_2 \frac{\tau_X}{\theta_X}}{\frac{\tau_X}{\tau_X+\theta_X} - \epsilon_1} \tag{B.30}$$

It can easily be verified that using the lower bounds obtained in equations (B.29) and (B.30) for update rates per each change type will result in distortions $D_1 = \epsilon_1$ and $D_2 = \epsilon_2$, and thus they are the minimum values needed.

Therefore, the total number of updates announced to the control plane divided by the total number of changes is given by

$$U_X(\epsilon_1, \epsilon_2) \quad = \quad U_X^1(\epsilon_1) + U_X^2(\epsilon_2) \tag{B.31}$$

Note that the total rate of type I changes, which is equal to the rate of type II changes in average is given by $\frac{1}{\tau_X+\theta_X}$ changes per second, thus total number of updates per second is given by

$$R_X(\epsilon_1, \epsilon_2) = \frac{U_X(\epsilon_1, \epsilon_2)}{\tau_X + \theta_X} \tag{B.32}$$

Combining equations (B.29-B.32) proves the Lemma. □

B.12 Proof of Lemma 12.3.5

Proof. Recall that $\mathcal{V}_c$ is the set of caches, ρ_i denotes the probability that a specific cache contains item i. Let S_{ij} represent the state of an item i at a node j, which is 1 if cache j contains item i, and 0 otherwise, and let $\hat{S}_{ij}$ denote the corresponding state perceived by the CRS. A request from a user is not served internally (by a cache in second layer) either if no cache contains it:

$$Pr(\forall\ j \in \mathcal{V}_c : S_{ij} = 0) = (1 - \rho_i)^{N_c}, \tag{B.33}$$

or if there are some caches containing it but the CRS is not aware of that:

$$\begin{aligned}
&Pr(\exists\ j \in \mathcal{V}_c :\ \ S_{ij} = 1\ \&\ \hat{S}_{ij} = 0) \\
&= \sum_{k=1}^{N_c} \sum_{1 \leq j_1 < .. < j_k \leq N_c} Pr\left({}^{i \notin \mathcal{V}_c - \{j_1, \ldots, j_k\}\ \&}_{[\hat{S}_{ij_l} = 0,\ S_{ij_l} = 1]_{l=1}^{k}}\right) \\
&= \sum_{k=1}^{N_c} \sum_{1 \leq j_1 < .. < j_k \leq N_c} (1 - \rho_i)^{N_c - k} \Pi_{l=1}^{k} Pr\left({}^{\hat{S}_{ij_l} = 0}_{S_{ij_l} = 1}\right) \\
&= \sum_{k=1}^{N_c} \binom{N_c}{k} (1 - \rho_i)^{N_c - k} D_{1_i}^{k} \\
&= (1 - \rho_i + D_{1_i})^{N_c} - (1 - \rho_i)^{N_c}
\end{aligned} \tag{B.34}$$

where $D_{1_i} \geq 0$ is the probability that i exists in cache j and the CRS does not know about it.

Thus the probability that a request is served externally is $1 - P_i$ which equals

$$(1 - \rho_i)^{N_c} + [(1 - \rho_i + D_{1_i})^{N_c} - (1 - \rho_i)^{N_c}] = (1 - \rho_i + D_{1_i})^{N_c} \tag{B.35}$$

where under the independent cache assumption, the state of an item in a cache is independent of the state in another cache. The probability $D_{1_i} \geq 0$ is always less than the probability of i being in cache j ($D_{1_i} \leq \rho_i$), and if the state updates are done at rate greater than $R_i(\epsilon_1, \epsilon_2)$, it will also be less than ϵ_1. □