# FUNDAMENTAL NETWORK BEHAVIOR OF MOBILE AD HOC NETWORKS

**BY WING HO ANDY YUEN**

**A dissertation submitted to the**

**Graduate School—New Brunswick**

**Rutgers, The State University of New Jersey**

**in partial fulfillment of the requirements**

**for the degree of**

**Doctor of Philosophy**

**Graduate Program in Electrical and Computer Engineering**

**Written under the direction of**

**Professor Roy D. Yates**

**and approved by**

_____

_____

_____

_____

**New Brunswick, New Jersey**

**January, 2004**

## ABSTRACT OF THE DISSERTATION

# Fundamental Network Behavior of Mobile Ad Hoc Networks

**by Wing Ho Andy Yuen**

**Dissertation Director: Professor Roy D. Yates**

This thesis is a collection of research on the fundamental network behaviors that pertain to the subject of mobile ad hoc networks. The first part of this thesis focuses on *mobile infostation networks*, a new class of mobile ad hoc network that exploits node mobility to improve network capacity. We address three important problems, namely the effect of node noncooperation, transmit range and node mobility on network performance.

The issue of node noncooperation is examined in the context of a content distribution application. When two nodes are in proximity, they negotiate for a file exchange in accordance to a social contract. An exchange is warranted only when each node can obtain something it wants from the exchange. Both common interest and dissimilar interest models are examined. The performance of different user strategies are evaluated through analysis and simulations.

The effect of transmit range on network capacity is then examined under a realistic interference model. Four transmission strategies are analysed and we show that a stipulated transmit range improves the capacity compared to the Grossglauser-Tse strategy. The optimal number of neighbors is determined, which is much smaller than the magic number of 6 to 8 neighbors for multihop networks. In addition, the capacity per unit area of the strategies is shown to increase linearly with node density.

We have also examined the effect of node mobility on highway mobile infostation networks via a novel highway model. Using arguments from renewal reward theory, the long run data rate of an observer node can be derived. For node speed that is uniformly distributed, we show that the data rate is independent of observer node speed in reverse traffic. In forward traffic, we show that the data rate increases with observer node mobility.

In the second part of the thesis, we focus on *multihop ad hoc networks*, in which nodes communicate in multihop routing. We have investigated the effect of transmit range on energy efficiency of packet transmissions, and determine a common range for all nodes such that the average energy expenditure per received packet is minimized. Both stationary and mobile networks are considered. The dependence of energy efficiency on various system parameters is investigated.

We also examined the network behavior of a routing algorithm for multihop ad hoc networks. By using an alternate graphical interpretation of simulation results, inter-relationships between performance metrics and system parameters are revealed. These observations often give us insights to the mechanisms that underlie the network behavior.

In the final part of the thesis, we present an online Dutch auction application for mobile cellular network. The price of an item decrements at regular intervals until a buyer place the bid to terminate the auction at the current price. We present a price decrement strategy that take into account of communication costs. This strategy maximizes the expected revenue of the auction host. Significant gain can be obtained compared with a uniform decrement strategy.

# Acknowledgements

First, I would like to express my gratitude to my advisor Professor Roy D. Yates for his guidance and support throughout my Ph.D. studies at WINLAB, Rutgers University. I have benefited tremendously from the his extensive knowledge and technical insight, amusing yet often outrageous, on wireless communications. More importantly, I owe my research skills and style to him, stripping down a research problem to its essence: amenable to analysis yet general enough to yield insightful results to many facets of the problem. Roy also has an interesting and exuberant personality that defies the stereotypes of a university professor. I have the luck to have learnt the art of having fun in doing research from him. It is like collecting gemstones. You know there is something out there. The fun part is to digging it out.

I would like to express my gratitude to my PhD committee members, Prof. Dipankar Raychaudhuri and Prof. Narayan Mandayam. I would also like to thank Prof. David Goodman for serving as the external examiner in the thesis committee. His breadth of knowledge is phenomenal. I also have to thank Prof. Zoran Gajic for serving on the committee of my Ph.D. qualifying examination.

I would like to thank all of my colleagues and friends that I have worked with in my postgraduate years, both in Hong Kong and here in WINLAB. Dr. Chi Wan Sung of the City University of Hong Kong has been a great friend as well as a great colleague. Chapter 5 of this thesis is a joint work with him when I worked under his grant support at Hong Kong in summer 2002. The original idea of Chapter 4 of this thesis was also conceived during that summer. Dr Siun-Chuon Mau of WINLAB is one of my best friends and colleagues in WINLAB. Siun-Chuon and Roy introduced me to the new and exciting paradigm of mobile infostation networks. Joint work with Siun-Chuon yields the results for Chapter 2 of this thesis. I am also also indebted my mentor Dr.

Wing Shing Wong of the Chinese University of Hong Kong. He was my master thesis advisor at the Chinese University of Hong Kong and has introduced me to the exciting field of wireless communications. I am also indebted to him for supporting me with a resesarch grant in summer 2000. Chapter 7 of this thesis is a joint work with Wing Shing during my stint. I also thank Dr. Heung-no Lee of the University of Pittsburgh for hiring me as a summer intern at the HRL Labs in summer 2001. It was a great experience working in a corporate research lab.

Many other friends have made WINLAB an agreeable place for doing research. I have to thank Lang Lin, Henry Wang, Jin Wang, Praveen Gopalakrishnan, Furuzan Atay, Kemal Karakayali, Siamak Sorooshyari, Shrenik Patel and many others. All of you guys help to create a good research atmosphere in WINLAB for us to thrive. I am sure there are more opportunities in the near future that we can have research collaborations. There are even more friends outside WINLAB to thank for. You know who I am talking about. Thanks for the accompaniment, the indulging conversations and providing me for the much needed distractions outside work.

I would like to thank my parents for their encouragement, love and friendship. I am blessed by their support of my decision to pursue a Ph.D. degree in United States. Finally, I thank the Lord for everything. The thesis will not see the light of day without His helping hand. Thanks for providing my home, my friends, and the journey that He has planned for my life.

# Dedication

To my parents

# Table of Contents

# List of Tables

# List of Figures

# List of Abbreviations

**AODV** is for Ad Hoc On-demand Distance Vector routing algorithm

**CBR** is for constant bit rate

**CDF** is for cumulative distribution function

**CSMA** is for carrier sense multiple access

**CTS** is for clear-to-send packets

**DARPA** is for Defence Advanced Research Projects Agency

**DSDV** is for Destination-Sequenced Distance Vector routing algorithm

**DSR** is for Dynamic Source Routing algorithm

**DSSS** is for Direct Sequence Spread Spectrum

**IETF** is for Internet Engineering Task Force

**IID** is for identically and independently distributed

**KKT** is for Karush-Kuhn-Tucker Theorem

**LAN** is for local area network

**MANET** is for mobile ad hoc network

**NAV** is for network allocation vector

**PDF** is for probability distribution function

**RTS** is for request-to-send packets

**SIR** is for signal to interference ratio

**SWIM** is for shared wireless infostation model

**TCP** is for Transport Control Protocol

**TTL** is for time-to-live

**Wi-Fi** is for Wide Fidelity, an acronym for 802.11b products

**UDP** is for User Datagram Protocol

**WAP** is for Wireless Application Protocol

**ZRP** is for Zone Routing Protocol

# Chapter 1

# Introduction

## 1.1 Overview of Wireless Networks

In wireless networking, there are two main classes of communication paradigms, *infrastructure networks* and *ad hoc networks*. Infrastructure networks include cellular networks, wireless LAN's and infostation networks. The network operator deploys a network infrastructure within the coverage area to provide wireless connectivity to the vicinity. The infrastructure is known as *base stations* in cellular networks, *access points* in wireless LAN's and *infostation* in infostation networks, and are connected together to a backbone network by wire. All communications on the wireless medium occurs in one hop between the mobile nodes to the local base station/access point/infostation. A mobile node acts as the source or the sink of a communication circuit. *Ad hoc networks*, on the other hand, preclude the use of a wired infrastructure. These networks are applicable to locations in which a prior deployment of network infrastructure is impossible. Current applications are mostly confined to military and rescue operations for long range outdoor networks, or to indoor network setting such as a conference room with a collection of laptop computers. Mobile nodes are connected together to form a network on the fly. They also have routing capability and may act as the source, sink or a forwarding node to relay packet for other nodes. Multihop networks and mobile infostation networks fall into the category of ad hoc networks.

Cellular networks providing primarily voice service has witnessed the most successful story in infrastructure networking for the last two decades. Typically the base stations are deployed to provide ubiquitous coverage to all mobile nodes at all locations in the network. This can be envisaged in Figure 1.1 when the base station are so close to provide seamless coverage to all areas served by the network operator. Second

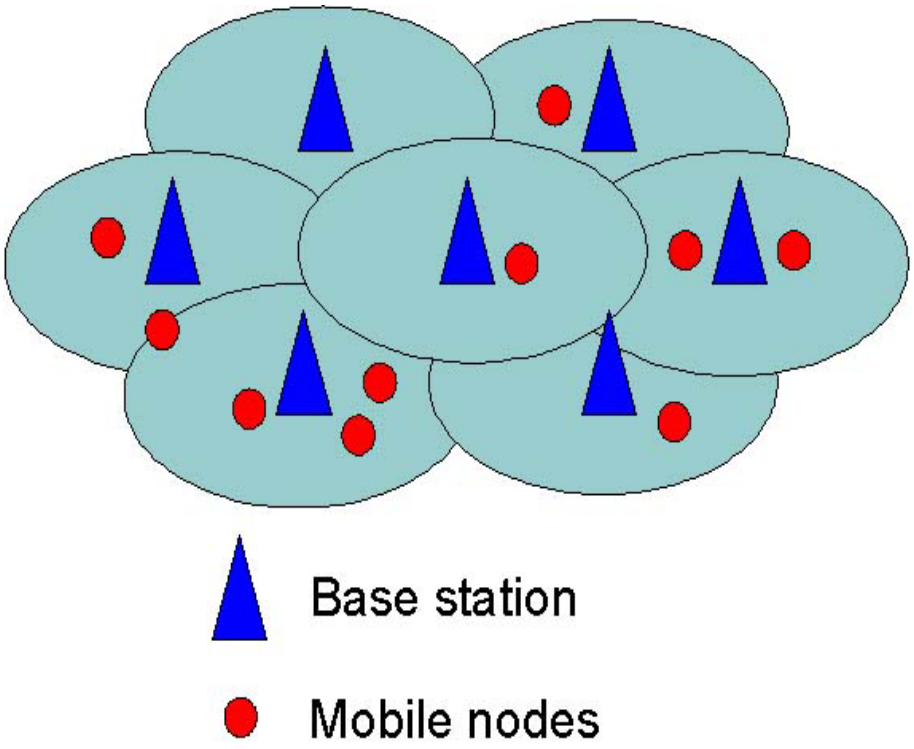



Figure 1.1: Illustration of the infrastructure network model.

generation cellular networks provide voice service predominantly. Third generation cellular networks are being deployed currently and offer heterogeneous voice and data services.

More recently, wireless LAN products have captured the limelight in spite of the doldrums of the telecommunications sector as one of the few market niches enjoying rapid growth, thanks to the proliferation of inexpensive 802.11b products. In contrast to cellular networks, wireless data is the predominant traffic type in a wireless LAN. Data applications are usually bandwidth hungry compared with the traditional low bit rate voice service. The current 802.11a/b/g standards for wireless LAN provides multi-megabit throughput for wireless data. For example, 802.11b offers a nominal rate of 2 Mbps and a peak rate of 11 Mbps. It uses an unlicensed portion of the radio spectrum at 2.4GHz. The availability of low cost hardware [9] contributes to the popularity of the widely successful Wi-Fi products. 802.11a uses OFDM transmission technology, operates at the 5GHz band, and offers a nominal rate of up to 54 Mbps. 802.11g is similar to 802.11a and uses OFDM. It operates at the 2.4GHz band, and is backward

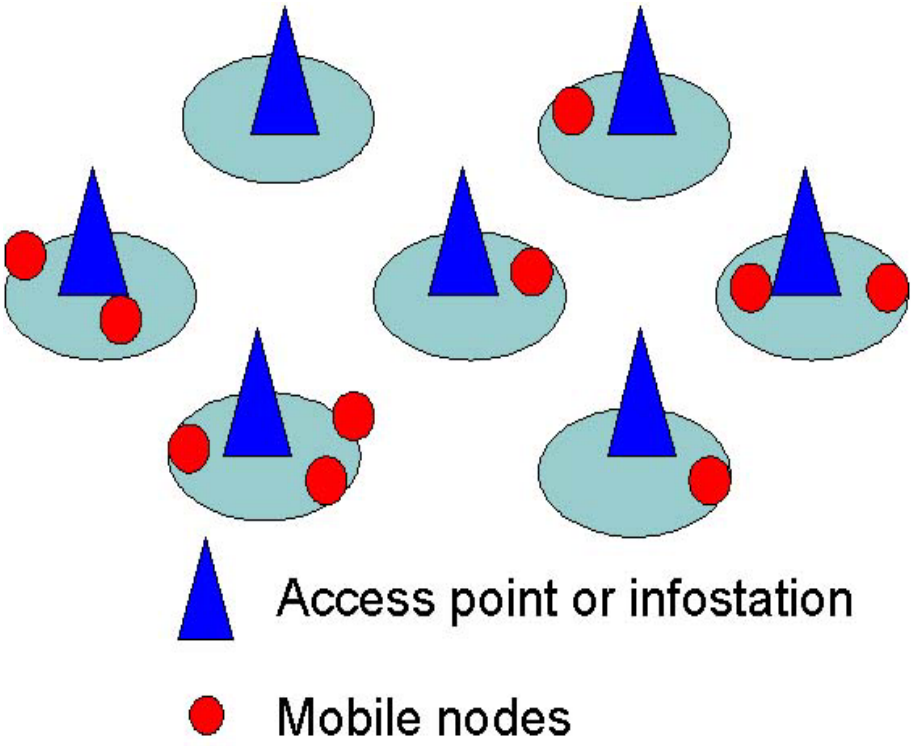



Figure 1.2: Illustration of the infrastructure network model.

compatible to 802.11b. In general, the coverage area of a wireless LAN is very small. As users roam in space they are intermittently connected to a local access point. This corresponds to the scenario in Figure 1.2 when the access points are far apart and the pockets of coverage areas are disjoint.

Infostation networks, pioneered by researchers at WINLAB [16,19,29,97], are a conceptual departure from the ubiquitous (anytime/anywhere) assumption in conventional cellular services. It is motivated by the fact that data services are often connectionless, delay insensitive, and have no specific bit rate requirements [16]. By restricting the transmit range of an infostation to the locality when the channel condition is excellent, the capacity is optimized from an information theoretic perspective [11,18]. For nodes with low mobility, the infostation network is akin to a wireless LAN as depicted in Figure1.2 with high bit rate islands of coverage close to the infostations. However, a wireless LAN typically does not support node mobility. To date, different access points have different ownership and operates autonomously without coordination or cooperation. The sharing of access points for roaming users is largely prohibited. Nevertheless,

an infostation network calls for the explicit co-ordination of infostations as a user moves around. A user may download parts of a large file from different infostations as it roams around the network in due time.

The second paradigm in wireless networking is *mobile ad hoc networks*, which includes multihop ad hoc networks and mobile infostation networks. The concept of multihop networks is not new. In the past two decades there were research in packet radio networks [35,38–40,43,45,48,75,82,92] under the DARPA program, which is in fact multihop networks with a fancy name. On the other hand, the idea of mobile infostation networks very recent, inspired by the infrastructure infostation networks. Although there are no large scale commercial deployment of these two networking paradigms to date, it is undeniable ad hoc networks are becoming one of the most active areas of networking research in these few years. A casual search in the ad hoc network literature reveals that there are very few papers in ad hoc network in the year 1997. Since then, the subject of ad hoc networks has captured the attention of many researchers.

In this thesis, we focus on research issues in mobile ad hoc networks. For pedagogical reasons we will outline multihop ad hoc networks and and then mobile infostation networks in the following.

## 1.2 Multihop Ad Hoc Networks

As shown in Figure 1.3, a multihop ad hoc network consists of mobile nodes which communicate with each other through multi-hop routes. Due to the dynamically changing topology, network routing is an important issue. Recently the Internet Engineering Task Force (IETF) has established a Mobile Ad Hoc Network (MANET) working group, which focuses on unicast and multicast routing protocols that are reactive to dynamic topologies and scale well to large networks. The intense interest in network routing is reflected by the voluminous amount of papers in the routing literature, [13, 22, 25, 30, 33, 47, 50, 54, 60, 62, 63, 73, 79, 84, 93, 94]. A taxonomy of unicast routing protocols could also be found in [73,79]. A number of well known routing protocols such as the destination-sequenced distance vector routing algorithm (DSDV) [62],

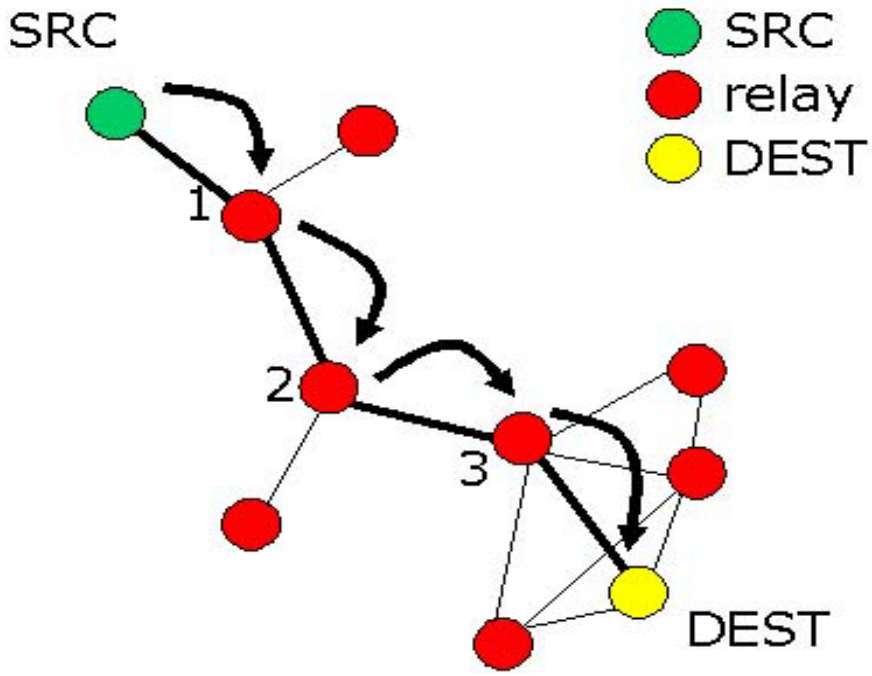



Figure 1.3: Illustration of the multihop ad hoc network model.

dynamic source routing (DSR) protocol [33], the on demand distance vector (AODV) routing protocol [63] and the zone routing protocol (ZRP) [25] are currently under standardization within the MANET working group.

Although many routing algorithms have been proposed in the literature, the achievable capacity in generic ad hoc networks is very low as demonstrated by simulation studies [8, 12, 32]. Recently, adaptive resource allocation techniques such as rate adaptation [27, 69, 99] and power control [6, 14, 49, 72, 84, 100] have been introduced to ad hoc networks as a means to improve network capacity. In the rate adaptation schemes, the transmit power of each node is constant. When the channel information of the receiver is available to the transmitter, the highest transmission rate could be used for a given bit error rate requirement, which maximizes the spectral efficiency. This effectively decreases the packet transmission time and the overall interference seen at each node, which increases the network capacity. Similarly, the judicious use of power control also reduces the cochannel interference of the network. This leads to more efficient frequency reuse and impacts the capacity of the network favorably. These adaptive techniques could be applied to the physical and MAC layer independent to the routing algorithm,

thus preserving the modularity of the various layers in the protocol stack. More generally, network decisions could be made with some knowledge of the channel information from the lower layers [99]. Because rate adaptation and power control schemes will affect some parameters that are being monitored, these measurements are passed to the network layer in the form of a routing metric, which affects route selection.

Power control for ad hoc networks have also been studied using analytical methods in the 80's for general packet radio networks. In [28,44,45,56,91], the objective is to find the transmit power such that the distance advancement towards the destination in one hop, also known as the *forward progress*, is maximized. More recent results are obtained for wideband spread spectrum systems [86,87], more detailed channel models [107] and some alternate optimization objectives other than the forward progress [90]. There are yet some scattered works [23, 61, 64, 65] that study the critical transmit range such that an ad hoc network is connected. Whereas [64] and [65] conjectured the critical range for networks of finite size, [23] addressed the asymptotic connectivity of ad hoc networks when the number of nodes tends to infinity. Using results deriving from percolation theory, [23] gave a proof for the asymptotic critical range. Using measure theoretic arguments, [61] independently discovered a strong law for the longest edge of the minimal spanning tree, which is also the critical range of a network. The strong law holds for nodes that are distributed in a network with an arbitrary density function. In [24], it is further shown that network throughput is near optimal when nodes operate on critical power. Using network simulations, however, we show in [100] that the critical range turns out to be suboptimal in throughput and energy efficiency. The underlying reason of the discrepancy is that the uniform traffic assumption breaks down when the network is critically connected, which is crucial in the proof of [24]. The above works consider only stationary network scenarios. The effect of mobility on the optimum transmit power is neglected. In general, in high mobility scenarios the optimum transmission power increases [78, 100] so that there are fewer link failures at the expense of less efficient frequency reuse.

Although there is intense research activity going on to design more efficient network

protocols, the study of the fundamental physical mechanisms that affect the performance of ad hoc networks has largely been ignored. One goal of this thesis is to investigate the physical and network mechanisms that affect the network performance of multihop ad hoc networks. In chapter 5, we examine the effect of transmit range of energy efficiency and throughput of the network. The optimum transmit range turns out to be much larger than the critical transmit range, and is insensitive to node mobility. In chapter 6, we examine the inter-relationships of various performance metrics and system parameters. This also proves to be a rewarding exercise and we obtained many observations that give us insights to the mechanisms that underlie the network behavior.

## 1.3 Mobile Infostation Network

In a mobile infostation network, any two nodes communicate only when they are in proximity and have a very good radio channel. Under this transmission constraint, any pair of nodes is intermittently connected as mobility shuffles the node locations. The network capacity of mobile infostation networks compares favorably to conventional multihop ad hoc networks [20,24]. In [24] Gupta and Kumar showed that the per node throughput of a multihop network drops to zero at a rate $O(1/\sqrt{n \ln n})$ in the limit of large number of nodes $n$. Thus multihop networks do not always scale with large network size. On the other hand, Grossglauser and Tse showed in [20] that the per node throughput of a mobile infostation networks is $O(1)$, independent of the number of nodes. This capacity is achieved through a two hop relay strategy.

Assume that each node in the network selects a random destination for unicast. We focus on a source node $i$, which has packets to deliver to a destination node $j$, as shown in Figure 1.4. As time evolves, node $i$ moves along a random trajectory and eventually runs into nodes 1 and 2. Although neither nodes 1 nor 2 are the destination of $i$, $i$ still relays the packets to them, with the expectation that when each of the relay nodes reaches the destination $j$, it will complete the second relay on behalf of node $i$. In steady state, each of the other $n-2$ nodes contains packets generated by node $i$ and destined to node $j$. At any network snapshot, it is almost surely that the nearest neighbor

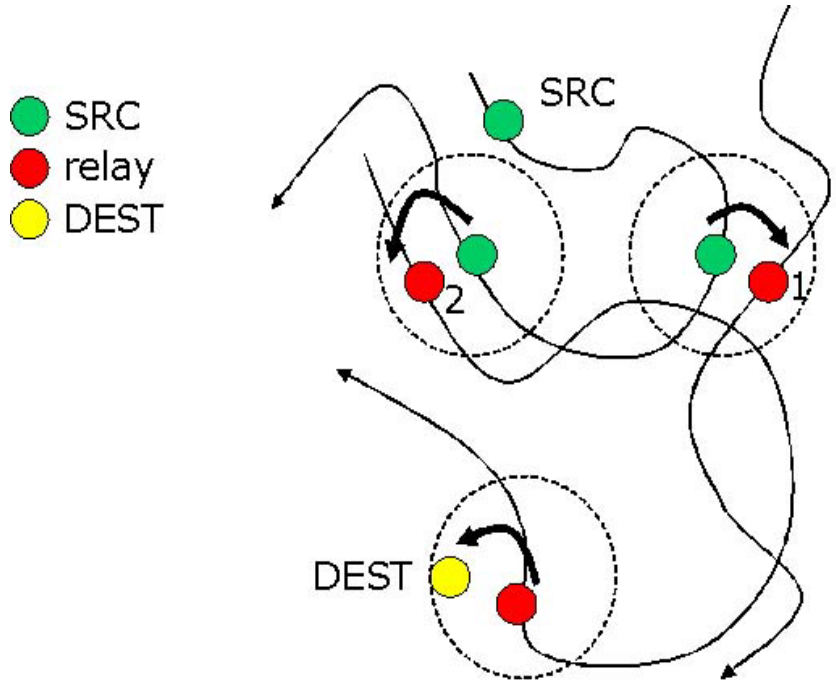



Figure 1.4: Illustration of the mobile infostation network model.

of node $j$ has packets addressed from node $i$ and completes the second relay on the behalf of $i$. That is, the long run per node throughput is constant and is independent of the network size. This capacity improvement comes from the exploitation of node mobility to physically carry the packets around the network, and is independent of the underlying mobility model, as long as the mobility process is ergodic.

Nevertheless, the order of magnitude improvement in network capacity comes at a cost. End-to-end transmissions incur a random delay that is at the same time scale of the mobility process. Thus, a mobile infostation network is applicable to delay tolerant applications with a heavy bandwidth requirement, say, in a content distribution application where all nodes are subscribers to a movie or news content provider. In this type of applications, a user is neither concerned nor aware of the movie download schedules. The application typically runs in the background for a few hours or even a few days as a user commutes to different places in his daily routine. This is consistent with ubiquitous computing environments [96], where computing systems become invisible and fade into the background and work for the users. In this case, we can draw a parallel of *ubiquitous networking environments* since users are not aware of the background

networking in the mobile infostation communication paradigm.

On the other hand, there is also a tradeoff between delay and storage in a mobile infostation network. Since a node transmits the same packets to all the relay nodes, there is heavy redundancy in packet transmissions and storage. This may not present a big challenge to researchers, since hardware storage follows the Moore's law quite well and storage capacity is approximately doubled every year. Moreover, a simple time-to-live (TTL) field can also be appended to each packet such that packets can be dropped when the TTL field has expired. This alleviates the storage requirement in individual nodes at the expense of more delay in packet delivery.

The seminal work [20] by Grossglauser and Tse has set the stage for further research in this new network paradigm. Although research on power control and rate adaptation techniques will push the capacity limit further, it is unlikely that the capacity of multihop networks will increase several orders of magnitude using these techniques. Motivated by the dramatic capacity improvement of mobile infostation networks, there are a number of recent papers that explore the mobile infostation paradigm in different contexts. Whereas [20] focused on unicast, most other papers in the literature focused on multicast. The potential spectrum of applications ranges from biological information acquisition systems used in the habitat monitoring of endangered wildlife species such as whales [85] and zebras [34] on one hand, to mundane movie and news downloading in a content distribution network [103, 104] and location specific information services [58, 59] on the other hand. [58, 59] addressed single hop multicast in mobile infostation networks. Reference [85] describes a new paradigm called the Shared Wireless Infostation Model (SWIM), in which nodes act as infostations and cooperate to forward packets for each other. This is in fact a cooperative mobile infostation network. The delay performance is evaluated via simple analysis and is verified by simulation results.

Multihop networks and mobile infostation networks are the two extreme instantiations of the capacity-delay tradeoff over many possible networking paradigms. In order to expedite data dissemination in a mobile infostation network, multihop forwarding may also be used occasionally, as in [58, 59], if a node has not done so for other nodes for some time. Similarly, node mobility can also be exploited in multihop networks

to improve network performance. For instance, in [21] node mobility is exploited to disseminate co-ordinates of all node locations without incurring any communication overhead. The location information is useful for nodes to make local routing decisions to the destination when geographic routing schemes [31] are used.

Most of the work so far [20, 34, 58, 59, 85] has focused on network scenarios in which nodes cooperate. For some applications such as habitat monitoring of wildlife species, sensor nodes are deployed from a single organization. The cooperation assumption between nodes is valid. On the other hand, in commercial applications each node in the network is autonomous and may act selfishly. A node is not incentivized to relay other people's packets since it is expending its own bandwidth and energy resources in a transmission. We have studied the problem of noncooperation between nodes in [103, 104] in the context of content distribution. The main results are reported in chapter 2 of this thesis. Data of common interest such as a movie is split into small files that are cached at the fixed infostations. Whenever a node comes close to an infostation, files can be downloaded. More generally, when two nodes are in proximity, they can negotiate for a file exchange for their own benefit. It turns out that a new kind of diversity emerges in noncooperative networks, in which we coined *data diversity*. Moreover, user strategies can exploit *multiuser diversity* to further improve the network performance.

In the mobile infostation literature, the concept of physical proximity is not well characterized. [20] assumed that a *candidate transmit node* always transmits to the closest receive node. Although the transmit and receive node pair has the shortest distance, this strategy may not perform well since this distance may be large in some pathological topology realizations. In these links, the benefit of spatial transmission concurrency may be more than offset by a simultaneous increase in total interference power in the network. It may be worthwhile to suppress the transmissions when the channel is less excellent, even though the receive node is the node closest in distance. The resultant decrease in total interference power due to the suppression of transmissions in the less excellent channels may be beneficial to the sum rate of the remaining connections. To ensure that only excellent channels are used, we have imposed an artificial *transmit*

*range* for all nodes in [101, 102, 106]. A candidate transmit node will schedule a transmission only if it sees some receive nodes in its transmit range. The effect of transmit range on the capacity is studied in chapter 3 under a realistic interference model.

In chapter 4, we examine the effect of mobility on highway mobile infostation networks. In [20], mobility provides a mechanism such that numerous instances of excellent channels between different nodes can be exploited. The realization of large network capacity comes from the translation of maximal spatial transmission concurrency in each network snapshot to the long run end-to-end network capacity. The physical implication of mobility in node encounters has been glossed over. In reality, the total connection time of a node over a specific interval depends on the node encounter rate and the connection time in each encounter, both of which depend on the relative mobility of nodes. Although a high node speed results in more node encounters, the connection time in each node encounter also decreases. It is not apparent whether high or low speed results in a larger connection time, and thus, data rate. The simple Markovian mobility model in [103,104] proves to be inadequate for this study. We have proposed a general mobility model for highway networks in [105,106]. The highway scenario proves to be interesting despite its mathematical simplicity. First, forward traffic connection time is much larger than that of backward traffic, but the node encounter rate is also much smaller. It is not apparent which traffic type maximizes the fraction of connection time. Second, the connection time in an encounter depends on the transmit range of the nodes. For both forward and backward traffic, an optimal transmit range exists such that the long run data rate of a node is maximized.

## 1.4 Organization of the Thesis

The first half of the thesis is devoted to mobile infostation networks. Chapter 2 studies the network behavior of a mobile infostation network when nodes are noncooperative. A simple interference model and mobility model is used to facilitate tractable analysis. Chapter 3 examines the effect of transmit range on capacity of a mobile infostation network. A realistic network interference model is used in the study. Chapter 4 examines the physical implications of mobility on highway mobile infostation networks. A new

highway mobility model is proposed and used in the analysis. The second half of the thesis is devoted to the study of multihop ad hoc networks via network simulations on *ns*-2. Chapter 5 studies the effect of transmit range on energy efficiency and throughput of multihop ad hoc networks. The optimal transmit range turns out to be much larger than the critical transmit range, and is insensitive to node mobility. Chapter 6 examines the inter-relationships of performance metrics and system parameters in a multihop network. The final part of the thesis is a stand alone chapter and describes a novel wireless application for mobile cellular networks. A Dutch auctioning strategy is proposed that takes into account of the communication costs of an online Dutch auction application. The revenue of the auction host is maximized. Finally, Chapter 8 provides a summary of the main contributions of this thesis. We then look into a futuristic networking paradigm and show that our foundation work on the exploitation of node mobility in ad hoc communications may have important implications to this paradigm.

# Chapter 2

# Non-cooperative Content Distribution for Mobile Infostation Networks

## 2.1 Introduction

In this chapter we address the issue of noncooperation in the context of a mobile infostation network for movie downloading. All nodes are subscribers to a movie content distribution network. A movie is divided into $K$ files which are then cached in a network of fixed infostations, access points providing pockets of high-speed short-range coverage [17]. When a node comes close to an infostation, files can be downloaded. In an entirely noncooperative network, this would be the only mechanism for file dissemination. It only uses the high-speed channel between an infostation and a node near it, while wasting all the equally excellent channels between closely located nodes. A more efficient system would have any two nodes in proximity to act as mobile infostations to exchange copies of their files. When there are many nodes, a node obtains most of the files from node-to-node file exchanges. Data dissemination is thus distributed to all nodes and all locations in the network.

It is possible to allow file exchanges among mobile nodes while keeping the network essentially noncooperative by stipulating the following *social contract* for all nodes in the network. When two nodes meet, they inspect the file contents of each other. If each node identifies a file that it wants, a bilateral file exchange takes place. Conversely, if either of the nodes cannot find a file it wants, no file exchange takes place since that node has no immediate incentive to transmit a file to the other.

We have shown by analysis and simulations that the networking performance of this file exchange mechanism depends on node mobility and density. More importantly, we find that both fairness and throughput of the network improve as the number of

files in the network increases. We identify this phenomenon as a new form of diversity. Traditional communication diversity techniques exploit the variations of signal strength over temporal, spatial and frequency domains. *Data diversity*, on the other hand, arises due to the enlargement of individuals' preferences of data, and is a consequence of the assumption of noncooperation among the nodes. We conjecture that data diversity has important ramifications in the performance of other networking contexts such as multihop ad hoc networks.

We have also extended the common interest model to the case where each node has dissimilar interest. This is applicable to the contexts in which multiple movies or TV shows are cached in the infostations. When nodes have mutually exclusive or partially overlapping interests, network performance degrades drastically. We have identified two user strategies for the dissimilar interest model. Our simulation results show that network capacity can be significantly improved by exploiting multiuser diversity inherent in mobile infostation networks.

The rest of the chapter is organized as follows. In section 2.2, we describe the system model. Section 2.3 is devoted to performance analysis, and the results are verified by simulations in section 2.4. We describe a new form of diversity - data diversity in section 2.5. In section 2.6, we extend our common interest model to the case where nodes have partially overlapping interests. Simulation results of two user strategies are discussed. The results are interpreted further as a form of multiuser diversity in section 2.7. Finally, conclusions are drawn in section 2.8.

## 2.2 System Model

This work is largely motivated by [20] which employed a signal to interference ratio (SIR) based link quality model to demonstrate the that $N$ nodes in a region could maintain $O(N)$ simultaneous transmissions with acceptable SIR. However, in this work, we look to employ a simpler communication model in order to demonstrate the effect of the social contract on content distribution. As shown in Figure 2.1, the geography consists of $L$ discrete locations in a square grid with an infostation at the center of the

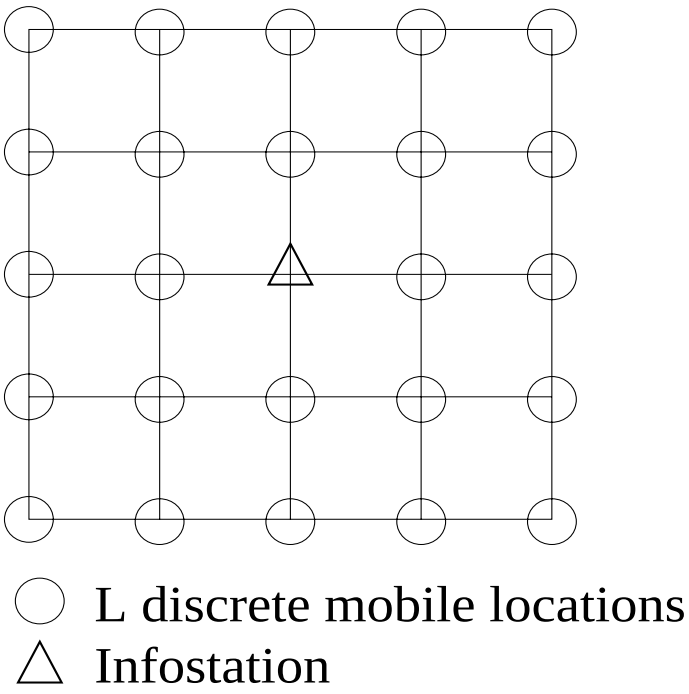



Figure 2.1: Illustration of the network model.

grid. The infostation cache holds the $K$ files of a movie. We assume the geography wraps around at each boundary, effectively creating a toroidal grid. We refer to this $L$ node wraparound grid with one infostation and $L-1$ *regular locations* as a *block*. A block is intended to mimic a typical multi-infostation network in which an infinite grid of infostations populate an infinite plane. The number of locations $L$ relative to the single infostation serves to characterize the density of fixed infostations over the terrain.

The $L$ location grid is populated with $N$ nodes with independent mobility processes. In our simulation experiments, we assume that time is discretized such that at each unit of time, each node randomly and independently moves in one of the four directions with equal probability $q = 0.25$. When two or more nodes are at the same location at the same time, we say those nodes are *neighbors*.

In our communication model, each node either downloads files from an infostation or exchanges files with a neighbor. At the infostation, only file downloading is allowed. At any other locations, file exchanges between mobile nodes are permitted. Given a particular radio bandwidth, the size of a file is chosen such that the time a node

occupies a location allows for either a bilateral file exchange between neighbors at a regular location or for two files to be downloaded from the infostation.

There are two factors that impact data dissemination. First there is a *transmission concurrency constraint* at each location. If there is more than one node at the infostation, contention is resolved by randomly picking one node for downloading. Similarly, when there are more than two neighbors at a location, two of the neighbors are randomly picked to perform a file exchange. Second, the probability of exchange is dictated by the *user strategy* which also consists of two parts. The user strategy must determine first whether to exchange files according to a *social contract*. Specifically, a node may want to exchange for a file because it is genuinely interested in that file. Alternatively, a node may want to exchange for a popular file, which is then used to facilitate future file exchanges. Thus even if a node cannot obtain a file of genuine interest, it may exchange for a file that it does not have. The user strategy then must specify which file should be picked from the other node. In the first part of this chapter, however, there is no distinction between the above models. Since all nodes have common interest in downloading the files of a popular movie, each node is genuinely interested in every file it does not have. In section 2.6, we extend the common interest model to the case where nodes have dissimilar interests that are partially overlapping. In that case, the network performance is dependent on the choice of the above models.

After two neighbors agree to exchange files, each downloads one file from the other. In an encounter in which there are multiple files of interest, a node must decide which file to download. Two strategies are examined in this chapter. For the random strategy, a node randomly selects a file it does not have from the neighbor node. Similarly, a node randomly selects two files that he does not have for downloading at an infostation. For comparison, we also consider a greedy strategy which assumes that each node has full knowledge of the circulation of each file within the network. For an infostation download or a neighbor exchange, a node picks the file that is the least circulated among all files it does not have. This strategy is greedy since it maximizes the probability of exchange $P_E$ between two arbitrary nodes in a static snapshot.

We note that the selection of two arbitrary nodes for file exchange is suboptimal.

Under the social contract the two selected nodes may not perform file exchange. A practical node selection protocol should avoid this by scheduling transmissions only to the node pair with an exchange agreement. The random selection of nodes is used to facilitate performance analysis and provide a lower performance bound to an ideal node selection scheme. On the other hand, the social contract implicitly assumes there are no misbehaved nodes. Each node makes no false claim on the files it possesses and ensures the integrity of all its disseminated files. The social contract provides a framework for studying non-cooperation between nodes. In a practical file exchange protocol, additional security mechanisms must be added to ensure the integrity of the files being exchanged.

The proposed content distribution network admits a number of performance metrics to describe how quickly files are disseminated. We define $T_1$ as the time when 80% of the nodes get all of the files. A network operator is interested in this quantity, which is related to the networking efficiency and the revenue generated from the network. We define $T_2$ as the time when all nodes get 80% of the files. A network subscriber, on the other hand, will be interested in $T_2$, which is related to fairness and perhaps will influence his willingness to pay. We also define $T_3$ as the time for all nodes to get all the files. Finally $T_4$ is defined as the time for an arbitrary node to obtain all files. An analytical expression for $E[T_4]$ is obtained in the next section.

We also evaluate the network performance in terms of *throughput* $C_i$, which characterizes the average rate of file downloading per node. This is defined in terms of the networking time $T_i$ and is given by $C_i \triangleq K/E[T_i]$, for $i = 1, 2, 3, 4$. The units of $C_i$ are files per node per unit time. Note that we can view the distribution to a particular node of movies over time as a renewal process in which the renewal period equals $T_4$, the time required for the node to obtain one movie. Since the node obtains a reward of $K$ files in each renewal period, renewal-reward theory assures that the expected rate at which the node obtains files is precisely $C_4$ [76].

## 2.3 Performance Analysis

When two or more mobile nodes are at the same location, a two-step process determines whether a file exchange occurs. First, the nodes at that location follow a radio access protocol to determine which pair of nodes will attempt a file exchange. We use the term *access* to refer to the event that a node gets to be one of a pair of nodes that examines the files carried by the other. Under some simplifying assumptions, we will see that at a regular location the *access probability* is given by a constant $\beta$, that depends on the number of nodes $N$ and locations $L$ in the block. For a pair of nodes chosen in the access phase, the *exchange probability* $P_E$ denotes the probability that the two nodes can exchange files under the terms of the social contract. The exchange probability will depend on the file contents in each node, which in turn depends on the user strategy.

In this section we provide a simple approximate analysis of $\beta$ and $P_E$. We then develop a simple Markov chain model to obtain the expected networking time $E[T_4]$ and the corresponding throughput $C_4$ for each node. We make the following key assumptions:

- **Memoryless Uniform Mobility** In each time unit, each node is randomly and independently at any of the $L$ locations with probability $p = 1/L$.
- **Independent Uniform Content Distribution** Given that node $i$ has obtained $l_i$ files, all combinations of $l_i$ out of $K$ files are equiprobable, independent of the files held by all other nodes.

It is not hard to see that these assumptions are inconsistent with the system model of section 2.2. In particular, when the number of locations is small and mobility is limited, nodes are likely to be neighbors frequently and have highly correlated content. Nevertheless, our simulation results agrees closely with the analytical results, indicating that these assumptions work well in systems with moderately large number of files $K = 500$ and reasonable mobility $q = 0.25$.

Due to the transmission concurrency constraint, the maximum number of simultaneous transmissions in the block equals $L$, the number of locations. For a given number

of locations, it should be apparent that there is an optimum number of nodes $N$ such that the access probability is maximized. If the number of nodes in the network is small, the spatial transmission concurrency is not fully utilized. Similarly, if there are too many nodes in the block, only a fraction of nodes could schedule transmissions in the $L$ possible locations.

Given a particular node at a given location, memoryless mobility implies that the number of other neighbors at that location is a random variable $J$ with the binomial distribution

$$P[J=j] = \binom{N-1}{j} p^j (1-p)^{N-1-j} \quad j = 0, \ldots, N-1 \tag{2.1}$$

When a given mobile is at the infostation with $J = j$ neighbors, the probability $\beta'$ that the given node is chosen for the infostation download is $1/(j+1)$. Averaged over all $J$, the probability the given node is chosen for the download is

$$\beta' = \sum_{j=0}^{N-1} \frac{1}{j+1} P[J=j] = \frac{1-(1-p)^N}{Np} \tag{2.2}$$

Similarly, when a node is at a regular location with $J = j \geq 1$ other neighbors present, 2 out of $j+1$ nodes are randomly chosen. The conditional access probability that a given node is one of the two chosen nodes is $2/(j+1)$. Thus,

$$\beta = \sum_{j=1}^{N-1} \frac{2}{j+1} P[J=j] \tag{2.3}$$

$$= \frac{2[1-(1-p)^N - Np(1-p)^{N-1}]}{Np} \tag{2.4}$$

Based on (2.4), the optimal $N$ is around $2L$. Below, in equation (2.12), a more careful optimization of $\beta(N)$ in the limit of large $N, L$ with fixed density $\rho \triangleq N/L$, reveals that $\rho_{\text{opt}} \simeq 1.8$. One can use this result to determine the optimal spatial density of fixed infostations based on the anticipated spatial density of mobile subscribers.

When nodes $i$ and $j$ have the opportunity to exchange files, the probability of exchange $P_E$ depends on the files each node is holding. Suppose nodes $i$ and $j$ have $l_i$ and $l_j$ files in their caches. An exchange between the nodes will occur *unless* one node has a collection of files that is subset of the other's collection. Assuming, without loss

of generality, that $l_i \leq l_j$, an exchange failure occurs if node $i$ chooses its subset of $l_i$ files out of the $l_j$ files of node $j$. Since there are $\binom{K}{l_i}$ total ways for node $i$ to choose its files, the probability of exchange is

$$P_E(l_i, l_j) = 1 \; - \; \frac{\binom{l_j}{l_i}}{\binom{K}{l_i}} \qquad 0 \leq l_i \leq l_j \leq K \tag{2.5}$$

From (2.5), we can derive a tight upper bound for the probability $P_{E^c} \triangleq 1 - P_E$ of no file exchange between neighbor nodes with $l_i$ and $l_j$ files such that $aK \leq l_i \leq l_j \leq (1-a)K$ and $0 < a < 1/2$. When $K$ is large such that $aK$, $(1-a)K$, and $(1-2a)K$ are all much greater than 1, an asymptotic upper bound $\tilde{P}_{E^c}$ for $P_{E^c}$ coincides with the Stirling's approximation for $P_{E^c}$ and is given by

$$\ln \tilde{P}_{E^c} = \Big[2\,(1-a)\ln(1-a) \; - \; (1-2a)\ln(1-2a)\Big]K \tag{2.6}$$

As the multiplier of $K$ is negative for $0 < a < 1/2$, we deduce that when $0 < a < 1/2$,

$$\lim_{K\to\infty} P_E(l_i, l_j) = 1, \qquad aK \leq l_i \leq l_j \leq (1-a)\,K \tag{2.7}$$

That is, if each node has a non-vanishing fraction of all $K$ files, a file exchange almost certainly will occur when the number of files in the system is large.

To find an upper bound for $P_{E^c}$ that is valid for most values of $l_i$ and $l_j$, we observe that the small $x$ approximation $\ln(1+x) \simeq x$ implies

$$\ln \tilde{P}_{E^c} \simeq -2a^2K, \tag{2.8}$$

implying that $P_{E^c}$ can be made arbitrarily close to zero by choosing $a > O\Big(1/\sqrt{K}\Big)$. When the number of files in the system is large, file exchange almost always happens among neighbors during most of the file dissemination process. In practice, we can regard $P_E = 1$ when $K \geq 1000$. We will come back to this point when we discuss our simulation results in Figure 2.3.

In the following, we derive the expected networking time $E[T_4]$ for a node to obtain all files and the associated throughput $C_4$. We assume that $K$ is large such that (2.7) holds and we model the dynamics of movie downloading by the discrete time Markov

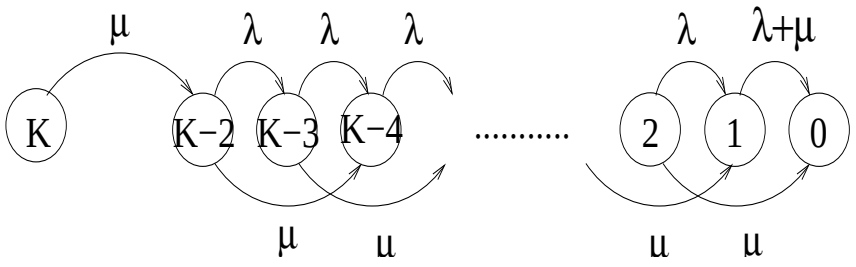



Figure 2.2: Illustration of the Markov chain model. The shown values denote the state transition rates. Note that the depiction of self transitions is omitted.

chain illustrated in Figure 2.2. Denote the state as the number of files remaining to be downloaded to a node. Initially a node is at state $K$. Since the first two files must be obtained from an infostation, the next state is $K-2$. Subsequently, in states $k \in \{1, \ldots, K-2\}$, each unit of time allows the following possibilities:

- With probability $p$, the node encounters the infostation and then with probability $\beta'$ downloads two files. The state goes from $k$ to $k-2$ with probability $\mu = p\beta'$.
- With probability $1-p$, the node is at a regular location and then with probability $\beta$ participates in a file exchange. The state goes from $k$ to $k-1$ with probability $\lambda = (1-p)\beta$.
- With probability $1-\lambda-\mu$, no new files are obtained and the state stays the same.

Denote the expected first passage time from state $i$ to state 0 as $g_i$, where $(2 \leq i \leq K-2)$. Conditioning on the next state transition and rearranging yields the difference equation,

$$g_i = \frac{1}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu} g_{i-1} + \frac{\mu}{\lambda+\mu} g_{i-2} \tag{2.9}$$

where the boundary conditions are given by $g_0 = 0$ and $g_1 = 1/(\lambda+\mu)$. Using z-transforms, we solve (2.9) to obtain

$$g_i = \frac{i(\lambda+2\mu) + \left(1 - \left(\frac{-\mu}{\lambda+\mu}\right)^i\right)\mu}{(\lambda+2\mu)^2} \tag{2.10}$$

It is obvious that $E[T_4] = 1/\mu + g_{K-2}$, where $1/\mu$ is the expected time until a node first encounters the infostation and obtains the first two files.

For a network with a single infostation supporting $N$ nodes over $L$ locations, we consider the large-system and many-files regime in which $N, L, K \gg 1$ while the spatial density of nodes $\rho \triangleq N/L$ is held constant. In this regime, (2.2) and (2.4) imply that the infostation download probability and the conditional access probability converge to

$$\beta'(\rho) \sim \frac{1-e^{-\rho}}{\rho} \tag{2.11}$$

$$\beta(\rho) \sim \frac{2}{\rho}\Big(1-(\rho+1)e^{-\rho}\Big) \tag{2.12}$$

Furthermore, $\lambda \sim \beta(\rho)$ and $\mu \sim \beta'(\rho)/L$ and the asymptote of the expected time for an arbitrary node to collect all $K$ files is

$$E[T_4] \sim \frac{K}{\beta(\rho)} + \frac{L}{\beta'(\rho)} \tag{2.13}$$

Here, the second term is equal to $1/\mu$ to account for the time for a node to fetch the first two files in an infostation encounter. The first term is an approximation to $g_{K-2}$ by assuming all remaining files are obtained from node to node file exchanges when infostation density is low, i.e. $L \gg 1$. If we further allow $K$ to grow large relative to both $N$ and $L$, the corresponding throughput $C_4$ of a node is

$$C_4 = \frac{K}{E[T_4]} \sim \beta(\rho), \qquad \frac{K}{N}, \frac{K}{L} \to \infty \tag{2.14}$$

We observe that the node density $\rho$ that maximizes $\beta$ also minimizes the expected networking time $E[T_4]$ and maximizes the throughput $C_4$.

To appreciate the extent to which social contract improves the rate of file dissemination of a completely noncooperative network, in which the only mechanism for file distribution is direct downloading from fixed infostations, we consider the Markov chain model for the latter. The corresponding difference equation for the first passage time from state $i$ to 0 is $g_i = 1/\mu + g_{i-2}$ for $i \le K-2$, yielding $E[T_4] = g_K = KL/2\beta'$ and

$$C_4 = \frac{2\beta'(\rho)}{L} \tag{2.15}$$

Hence, the social contract provides an $O(L)$, or equivalently $O(N)$ since $L$ and $N$ are of the same order, improvement to the individual file collection rate. The key ingredient in this improvement is the increase from $O(1)$ file deliveries per unit time made by

an infostation to $O(N)$ peer-to-peer file exchanges per unit time. With more complex models for radio communication and user mobility, in particular those employed in [20], the ability to support $O(N)$ communication links in a population of $N$ mobile nodes should yield similar improvements.

The social contract also leads to a similar improvement to the dissemination rate considered in our simulations, defined as the rate at which files are collected by nodes through either downloading from fixed infostations or file exchanges. Since the individual file collection rate $C_4$ is $\beta$, file dissemination rate with social contract is $N\beta$ during most of the dissemination process. On the other hand, the file downloading rate at an infostation is 2 if a node is present there, thus file dissemination rate without social contract is slightly less than 2. Therefore, the improvement offered by the social contract is of the order $N$.

## 2.4 Simulation Results

In this section, we examine the impact of the number of nodes $N$ and number of files $K$ in the system on the network performance, evaluated in terms of the expected networking time $E[T_i]$ and throughput $C_i$. In our simulations, the network size is kept constant at $L = 25$ nodes. A node moves to one of the neighbor locations w.p. $q = 0.25$ at each unit time. The performance metrics are obtained from ensemble averaging over 100 simulations.

For performance evaluation, we define the *dissemination rate* as the total number of files obtained, either by download from the infostation or by file exchange, per unit time over all mobile nodes. Figure 2.3 shows the dissemination rate averaged over 100 simulations runs. The number of nodes is held constant at $N = 50$ and the number of files is varied ($K = 50, 100, 500, 1000$). In all cases, the differences between the random and the greedy strategies were found to be very small. Thus, the random strategy is a good alternative to the greedy strategy for practical implementation.

From Figure 2.3, the y-intercept is slightly less than 2. Since the node density is high, it is probable to find at least a node at an infostation location and download 2 files

at $t = 0$. The file dissemination process has three distinct phases. In the first phase, the infostation seeds the mobile nodes with files and the dissemination rate increases rapidly as nodes obtain the ability to exchange files. Once most nodes have visited the infostation, $P_E \simeq 1$ and the dissemination rate remains steady at a peak rate that is a function of the access probability $\beta(\rho)$. In particular, each node will exchange one file with probability $P_E\beta(\rho) \simeq \beta(\rho)$. Over all $N$ nodes, the dissemination rate is $N\beta(\rho)$. Once a node has acquired all $K$ files, the social contract dictates that the node refrain from file exchanges. As the number of nodes with all $K$ files becomes significant, we enter the third phase in which the dissemination rate declines to zero as time evolves. The remaining nodes must download their files directly from an infostation, prolonging the time to download the entire movie. For all values of $K$, our simulations exhibit a significant tail associated with this final phase of dissemination.

As mentioned in the last section, in the absence of node to node file exchanges, the rate of file downloading shown in Figure 2.3 would have been constantly the y-intercept value of about 2, as opposed to $N\beta(\rho)$ most of the time. The simulation results are consistent with the analysis in the last section. As $P_E \simeq 1$ for large $K$, in each unit of time, each node will obtain one file with probability $\beta(\rho)$. With $N$ nodes in total, the average dissemination rate in the middle phase is $N\beta(\rho)$. In Figure 2.3, $N = 50$, $L = 25$, yields $\rho = N/L = 2$ and the middle phase dissemination rate is very close to $N\beta(2) \simeq 30$ files per unit time. The ratio of this rate to that of the completely noncooperative network is about 15—a dramatic improvement. Incidentally, we can interpret Figure 2.3 as a scaled version of $P_E$ as a function of $t$. When $t \to 0$, most nodes have nothing in their caches, thus $P_E(t) \simeq 0$. Similarly, $P_E(t) \simeq 0$ when $t$ is large since most of the nodes have finished downloading everything.

Lastly, for a finite population of nodes, we can mark the boundaries of the middle phase by the times about which all nodes have $O(\sqrt{K})$ and $O(K - \sqrt{K})$ files, based on the discussion of the upper bound of $P_{E^c}$ after (2.8). We hence observe that the first and third phases require $O(L\sqrt{K})$ time, roughly on the order of the time required for each node to acquire $\sqrt{K}$ file solely by visiting the infostation. On the other hand, in the middle phase, the system must deliver $O(NK)$ files in total at a dissemination rate

of $N\beta(\rho)$ files per unit time, and this requires $O(K)$ time. As $K$ increases (with $N, L$ fixed although not small), this middle phase comes to dominate the total dissemination time. Hence, for large $K$, the average dissemination rate is effectively the same as the peak dissemination rate of the middle phase. In short, as $K \to \infty$, the curve of Figure 2.3 converges to a rectangle with a constant file dissemination rate of $N\beta(\rho)$ files per unit time for a duration of $K/\beta(\rho)$ time units. This conclusion is consistent with the observation that the peak dissemination rate $N\beta(\rho)$ is simply $N$ times the average per node capacity $C_4$. We note that as $K \to \infty$, the transmission of each channel is only limited by contention, indicating the noncooperation strategy achieves almost optimum resource utilization.

In Figure 2.4, the networking time $T_i, i = 1, 2, 3$ are plotted against the number of nodes $N$. The number of files is kept constant at $K = 200$. From (2.2), it is easily verified that $\beta(\rho)$ is maximized at $\beta = 1.7933$ users/location, or $N_{\text{opt}} = 45$ users over $L = 25$ locations. This agrees with our observation in Figure 2.4(a), confirming that $N \simeq 45$ also minimizes $E[T_1]$. When $N$ increases past $N_{\text{opt}}$, $E[T_1]$ increases due to the increased contention at each location; however, the increase is partially offset by the increased opportunity for exchanges; hence, $E[T_1]$ is fairly insensitive to $N$ when $N \geq N_{\text{opt}}$. When $N < N_{opt}$, $E[T_1]$ increases quickly for decreasing $N$. When $N$ is small and node density is low, the system performance is hampered by the limited availability of file exchanges. In this case, $E[T_1]$ is very sensitive to $N$ since a small increase in $N$ significantly increases the rate of file exchange.

In Figure 2.4(b),(c), the optimum number of nodes that minimizes the networking time $T_2$ and $T_3$ are respectively $N_{\text{opt}} = 20$ and $N_{\text{opt}} = 10$ nodes, rather than $N = 45$ nodes. This disparity arises from the observation in Figure 2.3(a),(b) that when $K$ is not large, the total download time depends strongly on the duration of phase three which has a long tail. The tail length depends largely on the rate at which mobile nodes can download from the infostation. The tail decreases as $N$ decreases because fewer nodes results in each node having better access to the infostation. On the other hand, $T_1$ is unaffected by the long tail. A plausible reason is that networking is unfair; 80% of the nodes finish downloading all files well before hitting the long tail regime.

With reference to Figure 2.5, the networking time is plotted against the number of files $K$ cached in an infostation. It is obvious that the networking time $T_i, i = 1, 2, 3$ could be fitted to an asymptote as $K \to \infty$. The variance for $E[T_1]$ is small, indicating that the networking effect due to node mobility is deterministic. The slope of the asymptote is found to be around 1.63, which is equal to $1/\beta(N)$. $E[T_2]$ and $E[T_3]$, on the other hand, exhibit larger variances. The slope of the asymptotes for $E[T_2]$ and $E[T_3]$ are 1.1 and 1.6. When $K \leq 500$, we observe that $E[T_2]$ is larger than $E[T_1]$. Beyond $K = 500$, $E[T_2]$ is smaller than $E[T_1]$. This demonstrates that as $K$ increases, the networking between the nodes is more fair. That is, all nodes have approximately the same file downloading time. A plausible reason is that $P_E \to 1$ as $K$ increases. The downloading rate is no longer influenced by individual file content, but depends primarily on mobility and contention. For large $K \geq 500$, the downloading time is long compared with the time scale of mobility ergodicity. Each node therefore has a downloading time that is almost the same, such that $E[T_1] > E[T_2]$.

## 2.5 Data Diversity

In Figure 2.5, we showed that the networking time $E[T_i]$, $i = 1, 2, 3$ can be fitted nicely to an asymptote as $K$ increases. The corresponding throughputs are plotted in Figure 2.6 versus $K$. We observe that the throughput is an increasing function of $K$. It is instructive to find the asymptotic value of throughput $C_i^\infty$ as $K \to \infty$. To do this, we use the intuition captured in (2.13) and approximate the asymptote of $T_i$ by

$$T_i^\infty = m_i K + c_i \tag{2.16}$$

where $m_i$ is the slope and $c_i$ is the vertical intercept. Since the asymptote $T_i^\infty$ approaches $E[T_i]$ arbitrarily close when $K \to \infty$, we compute the asymptotic capacity as

$$C_i^\infty = \lim_{k\to\infty} \frac{K}{T_i} = \lim_{k\to\infty} \frac{K}{T_i^\infty} = \frac{1}{m_i} \tag{2.17}$$

Recall that $m_3 = 1.63$ as read from Figure 2.5(c). Thus $C_3 = 0.613$ files per node per unit time, or 30.65 files per unit time in our network where $N = 50$. This agrees

with our result in Figure 2.3(d). When $P_E \simeq 1$, the rate for data dissemination is around 30 files per unit time. Incidentally, we observe that

$$\lim_{K\to\infty} C_3 = \lim_{K\to\infty} C_4 \tag{2.18}$$

When $K \to \infty$, networking is fair and each node has the same throughput asymptotically. Thus, our simulation results are consistent with our simplified analysis.

The apparent increase in throughput can be understood using the concept of *data diversity*. In wireless communications diversity refers to the exploitation of variations in signal strength due to multipath fading. Since multipath fading exhibits signal variations over spatial, time and frequency domains, diversity techniques can be applied to select the strongest signal component over the respective domains. Diversity can also be exploited in a more general sense. In multiuser diversity, for instance, a receiver exploits the variability of received signal strength over different mobile nodes, and selects the node with the best channel for transmission.

Whereas the above techniques belong to the category of communication diversity, we argue that a new form of diversity, coined *data diversity*, is exhibited in noncooperative content distribution. When nodes are not cooperating, each node effectively has a preference list of files that evolves with time. If the number of disseminated files is large, there are more selections from a node's perspective. (2.5) and (2.7) dictate that file dissemination under the social contract is more efficient when there are more selections available for each node. There are, however, some differences between receiver diversity and data diversity. We note that receiver diversity is the result of a passive environment and we can exert no influence to the outcome. Data diversity, on the other hand, is the consequence of our social contract, over which we have complete control. Nevertheless, the social contract provides a general framework to study non-cooperation content distribution in mobile infostation networks. We have shown that data diversity is relevant to noncooperative data dissemination, which is gaining more attention in the networking community. Data diversity may also have implications to other peer to peer networks other than mobile infostation networks such as content distribution on the wired Internet.

Consider the possibility that several content providers use the mobile infostation infrastructure to disseminate their content (that are not highly overlapping) to a common group of subscribers. If a subscriber has files from content provider A and he encounters another subscriber with files from content provider B, these files generally would not be inter-exchangeable since they originated from different content providers. However, our results point out that content distribution for each provider would be more efficient, in terms of both throughput and fairness, if there were mutual agreements between content providers such that all files are inter-exchangeable, effectively increasing the content size $K$.

On the other hand, even if the content providers do not collude in data dissemination, data diversity can still be useful, say, in the dissemination of a single movie of a movie distribution network. Consider the scenario when a DVD quality movie is disseminated in a highway infostation network populated with fast vehicular subscribers. A typical drive-through infostation has a coverage radius of 20m [16]. A vehicle at a speed 20m/s therefore has a connection time of 2 seconds when it is in the coverage area of an infostation. Similarly, for two vehicles moving in opposite direction, the connection time is only 1 second. Suppose the infostation radios operate at a modest data rate of 160Mbit/s (which still substantially outperform the state of the art 54Mbit/s 802.11a access points available today). In order to facilitate the file exchange of two data files in the worst case of a head-on mobile to mobile encounter, the file size should be no more than 10MByte. On the other hand, the typical size of a DVD quality movie is roughly 5GByte. Thus, a movie should be split into $K = 500$ files and cached in fixed infostations for dissemination. Our simulation results in Figure 2.6(c) have shown that with a modest content size of $K = 500$ files, the achievable per node capacity $C_3$ is 80% of the theoretical per node capacity $\lim_{K\to\infty} C_4$ for asymptotically large $K$. Thus, without even relying on the collusion between the content providers, we can enjoy the benefits of data diversity in the dissemination of a single movie.

## 2.6 Dissimilar Interests

In our basic model, we assume all nodes have a common interest in $K$ files. In this section, we extend the common interest model to the case where each node has interest in only a subset of the $K$ files cached in the infostation. Depending on the type of content, the interests of the nodes can be *mutually exclusive* or *partially overlapping*. For instance, suppose multiple movies, say $1/\alpha$ movies are cached in the infostations, where $0 < \alpha \leq 1$. Each movie has the same length and is divided into $\alpha K$ files. If each node is interested in one movie only, then any two nodes will have interests that are either exactly the same or mutually exclusive. More generally, the interests of all nodes are partially overlapping. Consider the case where multiple TV shows are cached in the infostations. Without loss of generality we assume each TV show is stored as one file. Each node is interested in $\alpha K$ TV shows or files that is randomly selected from all $K$ cached files.

Recall in section 2.2 that a user strategy consists of two parts. Suppose two nodes seize the local channel successfully. First the two nodes must determine whether to exchange files. Second, upon an agreement of performing a file exchange, each node determines what to exchange as specified by the *random* or *greedy* strategy. In the common interest model, each node is interested in every file cached in the infostations. A node therefore is genuinely interested in every file that it does not have. In the dissimilar interest model, however, the above assumption is no longer valid. We can differentiate two user strategies in which neighbor nodes determine whether to exchange files. In **user strategy I**, neighbor nodes $A$ and $B$ perform a file exchange only if both nodes discover a file of genuine interest on inspection of each other's caches. In **user strategy II**, nodes $A$ and $B$ are obliged to exchange files if each node has a file that the other node does not have, whether or not those files are of genuine interest.

Once the nodes agree on a file exchange, either the *random* or *greedy* downloading strategy can be used in both user strategies. Nevertheless, we have demonstrated through analysis and simulations in earlier sections that the random and greedy downloading algorithms have almost identical performance. Hereafter, we consider only the

random downloading strategy when we compare the performance of user strategy I and II in the simulation studies.

We have performed simulations to study the network performance for the multiple movies model, where node interests are either exactly the same or mutually exclusive. The network performance is evaluated in terms of $\alpha$, which characterizes the extent of overlapping interest with other nodes. When $\alpha$ is very small, each node is interested in a small fraction of all files. The interests of any two nodes are likely to be mutually exclusive. As $\alpha$ increases, more nodes are interested in the same files. It is therefore more probable for a node to run into another node that has the same interest. When $\alpha = 1$, all nodes are interested in all $K$ files and our model reduces to the common interest model.

In our simulations, we assume the number of nodes in each infostation block is $N = 40$ and the total number of files is $K = 1000$. We consider the multiple movies model in which $1/\alpha = 1, 2, 4, 5, 10, 20, 40$ movies are distributed at the infostations. Each movie is split into $\alpha K$ files, and the corresponding values of $\alpha$ are $1, 0.5, 0.25, 0.2, 0.1, 0.05, 0.025$. In the case of 40 movies, each node is interested in different movies and have mutually exclusive interest. The number of nodes having the same interest increases with $\alpha$. When $\alpha = 1$, all nodes have a common interest for the same movie. Denote $E[T_i^{\alpha,j}]$, $i = 1, 2, 3$ as the expected networking time of user strategy $j$, where $j = 1, 2$. We are interested in finding the expected networking time for both user strategies.

Referring to Figure 2.7, the networking time of both user strategies is plotted versus $\alpha$. We observe that even when $\alpha$ is very small, the downloading time of user strategy I is quite large. In particular, when $\alpha = 0.025$, the number of files wanted by each node is only $\alpha K = 25$. The corresponding expected networking time $E[T_i^{\alpha}]$, $i = 1, 2, 3$ for both user strategies is approximately 700, 750, and 850 units. At $\alpha = 0.025$, each file is desired by one node. This is easily seen since by symmetry, each file is desired by $\alpha N = (0.025)(40) = 1$ node. Suppose all nodes observe user strategy I. It is obvious there is no file exchange between nodes since each node keeps only files that is wanted by that particular node only. On the other hand, when user strategy II is used, file exchanges between nodes are allowed. Nevertheless, a node never fetches a

file and benefits from a file exchange since all nodes have mutually exclusive interest. For both user strategies, each node has to download every desired file directly from an infostation. The absence of concurrent file exchanges in conjunction to infostation downloading explains the long and identical networking time.

Referring to Figure 2.7 again, it is obvious that $E[T_i^{\alpha,1}]$ and $E[T_i^{\alpha,2}]$ are increasing with $\alpha$ for $i = 1, 3$. This is plausible since in general, more time is needed for a fraction of nodes to finish file downloading as the number of desired files increases. An interesting (although not statistically significant) exception is observed for $E[T_2^{\alpha,1}]$, and might be explained by the following. When the number of files $\alpha K$ to be downloaded is small, a node usually runs into other nodes that have mutually exclusive interests. The node therefore has to download most of the files directly from the infostations, unable to enjoy the benefit of spatially concurrent file exchanges. As a result, these nodes have a large networking time. As $\alpha$ increases further, most, if not all, of the nodes participate in beneficial file exchanges due to the presence of nodes with the same interests. Since $E[T_2^{\alpha,1}]$ is dominated by the nodes without file exchanges when $\alpha$ is small, this explains the peak at $\alpha = 0.2$.

In order to explain the increasing trend of networking time with $\alpha$, and to characterize the performance difference for both user strategies, we examine the mechanism of the data dissemination in the following. As $\alpha$ increases from $\alpha = 0.025$, there is more nodes with the same interests. Each file is desired by $\alpha N$ users on average. Consider user strategy I. Approximately $\alpha N$ nodes are willing to act as the *networking agents* for each file and possibly carry the file in their cache as these nodes roams around the network. When $\alpha$ gets larger, the number of networking agents for each file increases. Since the circulation of a particular file is constrained by the number of networking agents for that file, increasing $\alpha$ effectively promotes the circulation of each file. This impacts the number of node-to-node file exchanges favorably, allowing more simultaneous file exchanges to take place. Consequently, the networking time $E[T_1^{\alpha,1}]$ and $E[T_3^{\alpha,1}]$ flatten quickly as $\alpha$ is increased.

For user strategy II, the networking time is consistently smaller than that of user strategy I as $\alpha$ increases from 0.025. Although nodes have little overlap of common

interests when $\alpha$ is small, user strategy II dictates that a file exchange ensues whenever each node can retrieve a file that it does not have on inspection of the cache of the other node. Thus, all $N$ nodes are willing to act as the networking agents for all files. The circulation of each file is not constrained by the particular interests of each node. Since nodes are more admissable and willing to carry files in user strategy II, the networking time is consistently smaller.

In the case $\alpha = 1$, our dissimilar interest model reduces back to the common interest model. Both user strategies I and II have identical networking time $E[T_i^{\alpha}]$, $i = 1, 2, 3$, that agrees to the corresponding values $E[T_i], i = 1, 2, 3$ for the common interest network model. When $K$ is reasonably large (in our case $K = 1000$), data diversity dictates that $P_E \rightarrow 1$ and the networking time is then only constrained by the contention probability $\beta$ given by (2.13).

## 2.7 Multiuser Diversity

In Figure 2.7, we showed that the networking time $E[T_i^{\alpha}]$, $i = 1, 2, 3$ for user strategy II is always less than that of user strategy I. The corresponding network capacity is plotted versus $\alpha$ in Figure 2.8. Again, x-axis denotes the fraction $\alpha$ of files that each node is interested in, where $\alpha$ takes the values of 0.025, 0.05, 0.1, 0.2, 0.25, 0.5,1. We observe that for both user strategies, the network capacity $C_i^{\alpha}$, $i = 1, 2, 3$ is strictly increasing with $\alpha$. The capacity of user strategy II is consistently larger than that of user strategy I when nodes have dissimilar interests ($\frac{1}{N} < \alpha < 1$). The capacity of both strategies coincide when $\alpha \leq \frac{1}{N}$ and $\alpha = 1$. When $\alpha \leq \frac{1}{N}$, all nodes have mutually exclusive interests. Even though user strategy II allows node-to-node file exchanges, there is no corresponding gain in network capacity. Similarly, when $\alpha = 1$, our model reduces back to the common interest model. Thus both user strategies I and II have almost identical capacities.

The increasing trend of network capacity with $\alpha$ can be understood using the concept of *multiuser diversity* inherent to mobile infostation networks. The efficiency of dissemination of this file is dependent on the willingness of the mobile nodes to carry it

across the network. If a node is willing to carry a particular file, then the node is effectively acting as a *networking agent* for that file. For user strategy I, each file is wanted by approximately $\alpha N$ nodes, who are willing to act as the networking agents for the file. For strategy II, each node is obliged to carry every file even if the file is not wanted by the node. The number of networking agents is then equal to the number of nodes $N$ irrespective of $\alpha$. We argue that the performance improvement of user strategy II is an exploitation of multiuser diversity, where the number of nodes willing to act as networking agents for each file is increased. Since the circulation of a particular file is equal or less than the number of networking agents for that file, the actual circulation of each file improves as the number of networking agents increases. As a consequence of improved file circulation, the efficiency of file exchanges improves as stipulated by data diversity, allowing multiple spatially concurrent file exchanges to take place.

From the above argument, we expect the two user strategies have the greatest performance disparity when $\alpha$ is small. Figure 2.8, however, shows that the percentage performance disparity is maximum when $\alpha$ is about 0.5. We note that the increase of the number of networking agents indeed leads to a proportional increase in the number of files in circulation. However, when $\alpha$ is small, each file is of genuine interest to only a few nodes and most file exchanges involve files that are of no interest to either node. Thus even if the circulation of all files is increased significantly, the corresponding increase in the number of file exchanges is not beneficial.

There are two opposing factors that impact the performance of user strategy II. For small $\alpha$, the number of networking agents for user strategy II is increased dramatically by a factor of $1/\alpha$. However, most of the file exchanges are not beneficial since node interests are largely non-overlapping. For large $\alpha$, there is only a nominal increase in the number of networking agents. However, since most nodes have very similar interests, each node gets many desired files and benefits from file exchanges. Our simulation results show that for $\alpha = 0.5$, we achieve an attractive, and perhaps optimum, tradeoff in terms of capacity gain. The corresponding capacity $C_i^{\alpha,2}$, $i = 1, 2, 3$ improvement of user strategy II over user strategy I is above 66% for all three cases.

Consider a movie distribution network in which 20 movies are cached in the infostations, making a total of $K = 1000$ cached files. Suppose each node is interested in only one movie of 50 files. This is equivalent to our multiple movies model with $\alpha = 0.05$. If all nodes observe user strategy I, the networking time $E[T_i^{\alpha,1}]$ is respectively $1100, 1200$ and 1300 units. On the other hand, if all nodes observe user strategy II, the networking time $E[T_i^{\alpha,2}]$ is $825, 825$ and 1000 units, roughly 70% of the original time. In content distribution, usually each node wants to minimize the networking time for files of genuine interest. Our simulation results point out that if a node acts as a networking agent for files he is not interested in, it actually expedites the file downloading process, reducing the networking time while enjoying a network capacity gain as warranted by multiuser diversity. This is an interesting result because it implies each node has an incentive to act as a networking agent and assist in data dissemination without having an explicit node cooperation model.

Although the exploitation of multiuser diversity in user strategy II yields better network capacity, it comes at a cost of increased energy consumption due to more frequent file exchanges. Thus there is a tradeoff between energy consumption and network capacity. If the network nodes have plentiful energy reserves, say infostations on vehicles, they should adopt user strategy II to tradeoff energy consumption for better throughput capacity. On the other hand, for nodes having scanty energy supply, they can cut down the energy consumption by sacrificing some throughput. Moreover, nodes do not need to adopt the same user strategy in a network. Each node can independently decide what user strategy to adopt based on its current level of residual energy.

We note that in user strategy II, there is implicit cooperation between nodes. Each node is obliged to act as the networking agent for files that it is not interested in, That is, each node caches and disseminates personally uninteresting files for other nodes as it roams the network. The performance gain of user strategy II over strategy I agrees with the intuition that more cooperation usually leads to better system performance. Although user strategy II requires implicit cooperation between nodes, there is no corresponding control overhead due to user cooperation. We do not assume the exchange

of files of genuine interest to neighbor nodes takes priority over other types of file exchanges. In our implementation, when there are multiple neighbor nodes at the same location, the first two nodes that broadcast control messages to request a file exchange seize the channel. This rule is equivalent to randomly picking two nodes from all neighbor nodes with no signaling overhead and is completely determined by contention. Note that giving priority to exchanges of files of genuine interest may improve overall system performance if one can develop an efficient protocol between multiple neighbor nodes to determine the optimal node pair to exchange files.

## 2.8 Conclusion and Further Work

We have addressed the issue of noncooperation among nodes in the context of content distribution in mobile infostation networks. In the first part, we assume all nodes have a common interest of $K$ files cached in the infostations. We have shown that it is possible to drastically increase the rate of file dissemination of a completely noncooperative network by requiring the absolute minimal cooperation among users in the form of a social contract. A random and a greedy file downloading algorithms are examined and shown to have similar performance. We show that there exists some optimal node density in these networks such that the access probability of a node is maximized and the networking time is minimized. More importantly, we show that the total number of files cached in the infostations impacts the networking fairness and throughput. We identify this phenomenon as data diversity that is distinct from conventional communication diversity. When nodes are noncooperative and have individual preference on data, the network exhibits data diversity and the throughput of each node increases with increasing content variety. In the second part, we extend the common interest model to the case where nodes have partially overlapping but dissimilar interests. Two user strategies are considered for this model. We show in our simulations that a file exchange strategy that takes advantage of the multiuser diversity inherent in mobile infostations results in enhanced network performance. We conclude that both data diversity and multiuser diversity can be exploited in the mobile infostation architecture even if nodes are noncooperative.

In the present work, simple mobility and interference models are used to facilitate analysis. This approach has been fruitful, leading to the observations of two diversity phenomena in noncooperative content distribution. Nevertheless, a thorough examination of the implications of mobility and interference to the network performance of mobile infostations is called for. As a first step, the issue of interference modeling is addressed in a recent paper [100]. The effect of transmit range on network capacity is examined. We found out a stipulated transmit range improves the capacity of a mobile infostation network further. An optimal number of neighbors exists for mobile infostation networks that is distinct from the well known 6-8 magic number [28, 44, 91] for multihop ad hoc networks. Moreover, the network capacity is linearly increasing with node density. Thus mobile infostation is an attractive alternative to multihop networking in future pervasive computing environments, where high node density dooms the throughput of multihop networks. On the other hand, the effect of mobility on mobile infostations is currently being studied. The connection time in each node to node encounter obviously depends on node mobility and needs to be quantified. To this end we have proposed a sophisticated mobility model for highway mobile infostation networks that allows for performance analyses based on renewal and queuing theories. We conjecture that the performance of mobile infostation networks are robust to mobility.

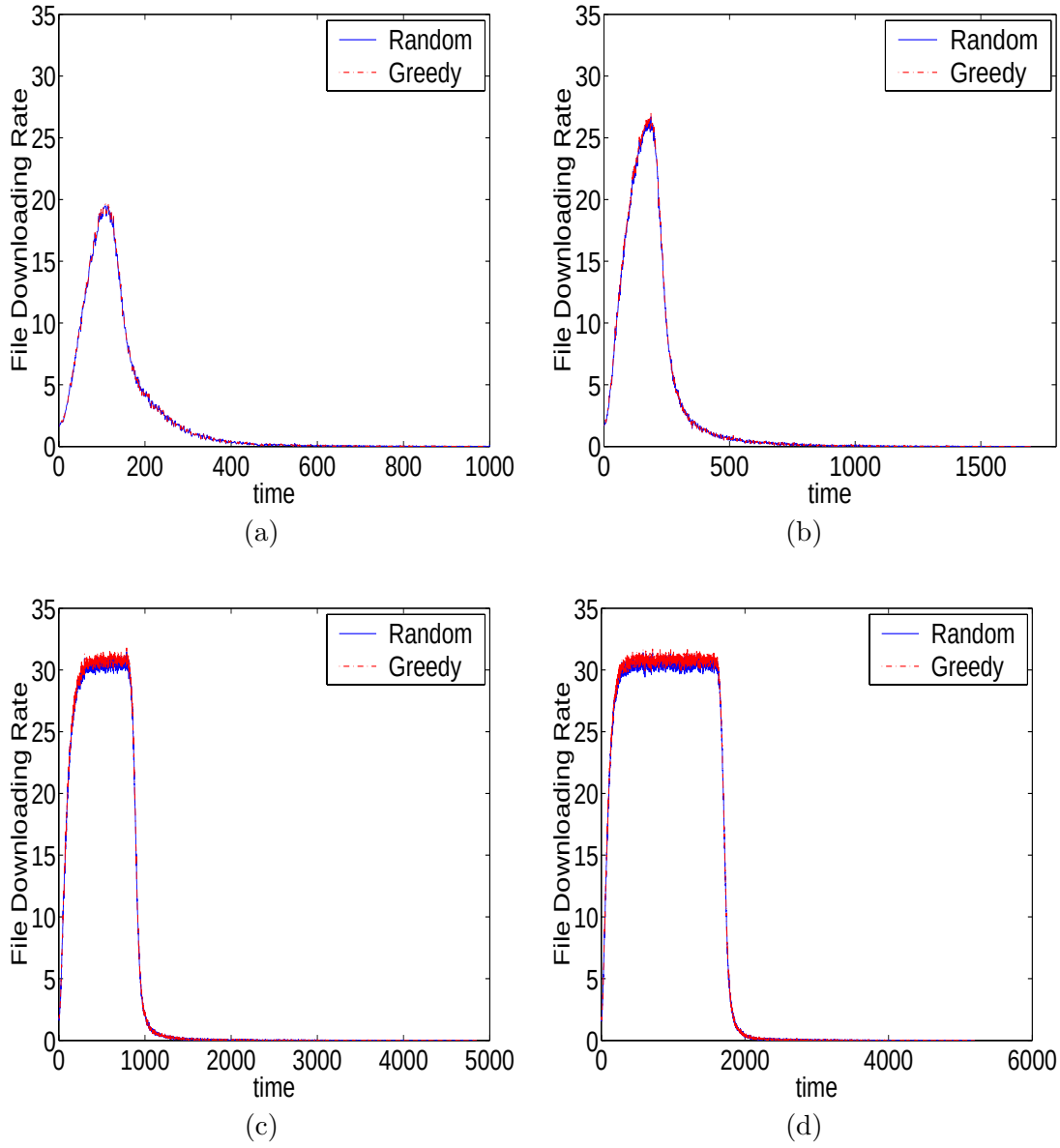



Figure 2.3: Average number of files obtained at each unit time over 100 simulations. (a) K=50, (b) K=100, (c) K=500, (d) K=1000.

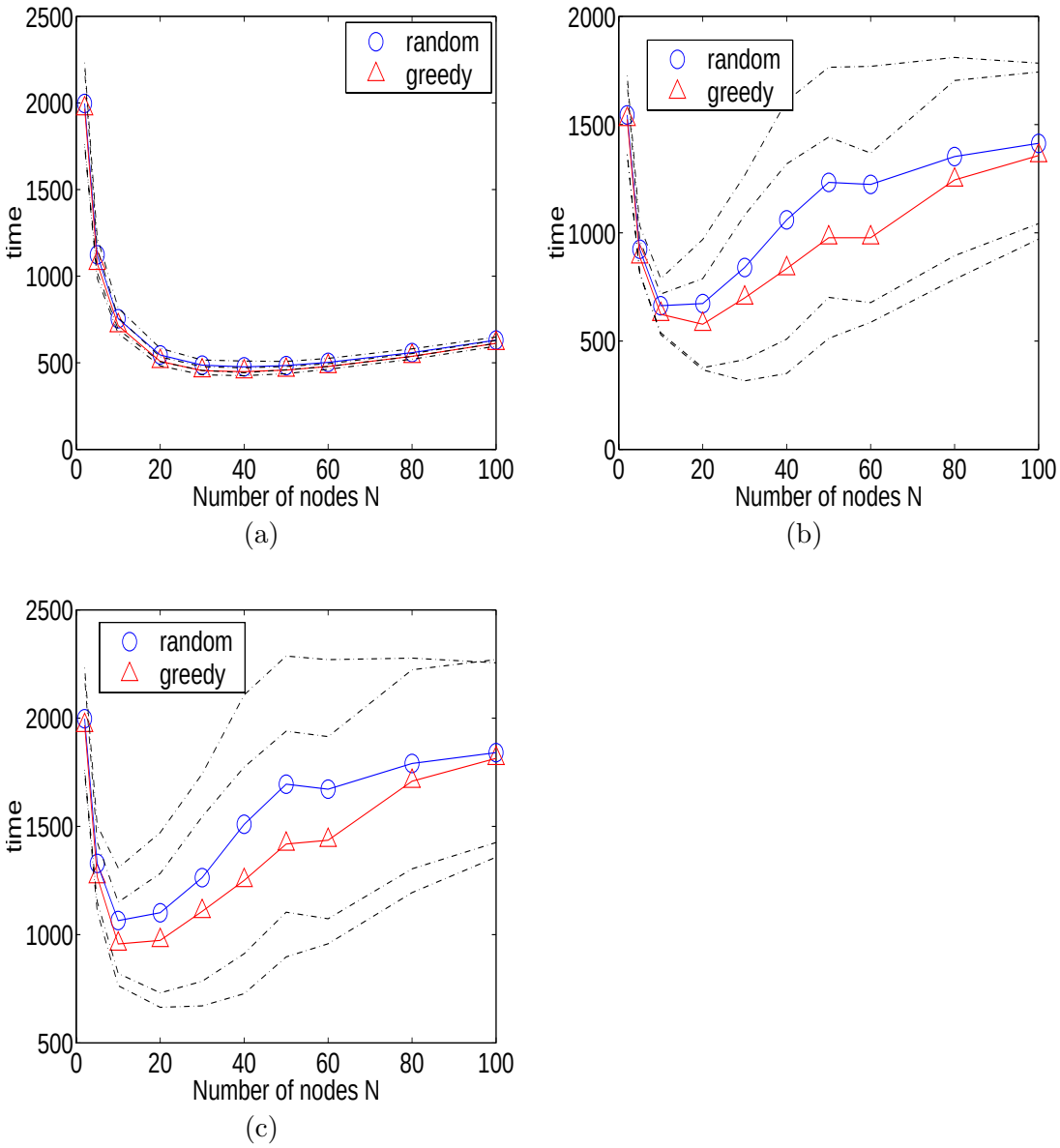



Figure 2.4: Average networking time vs. the number of nodes $N$. (a) $E[T_1]$ when 80% of all nodes obtain all files, (b) $E[T_2]$ when all nodes obtain 80% of all files, (c) $E[T_3]$ when all nodes obtain all files.

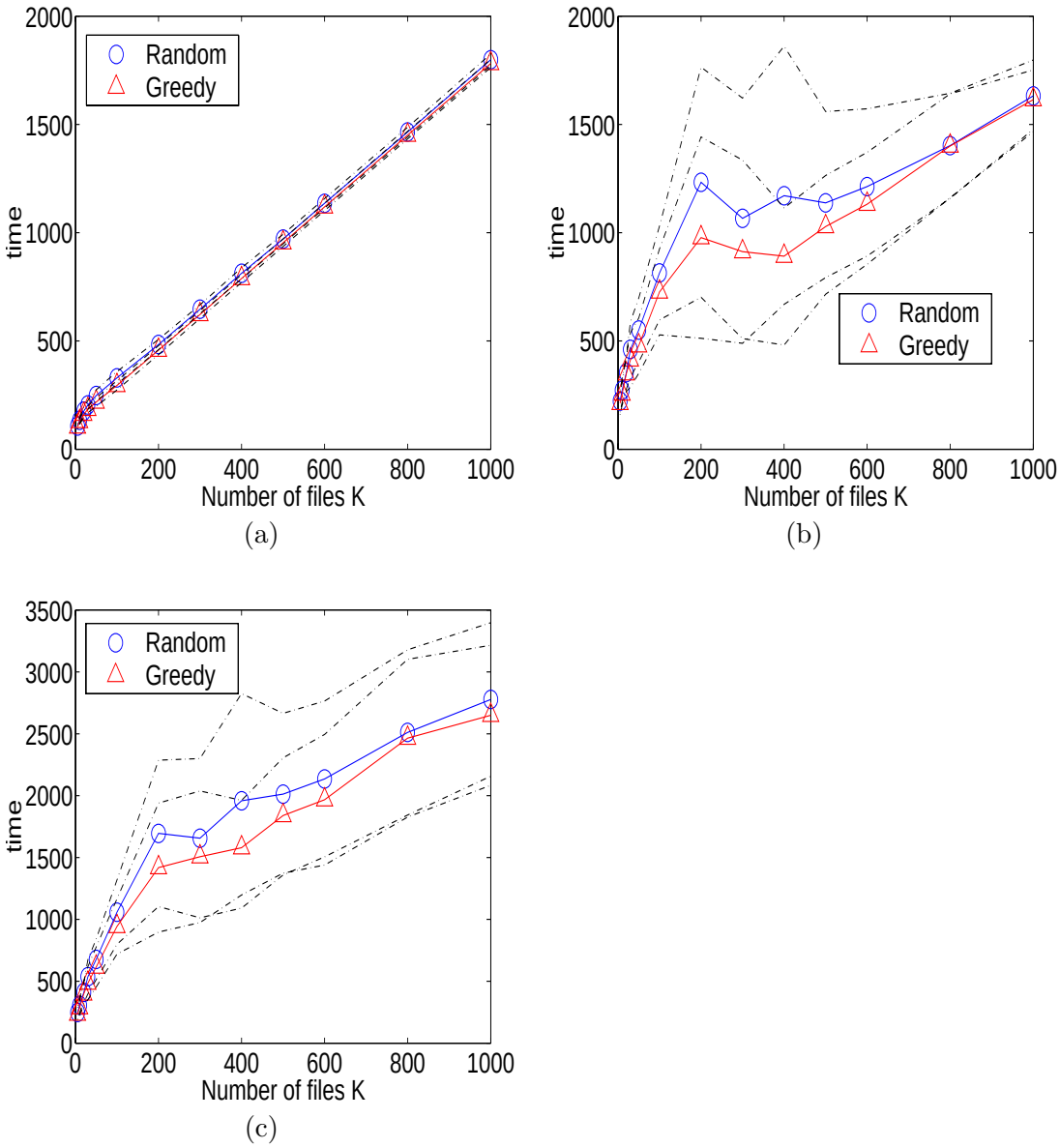



Figure 2.5: Average networking time vs. the number of cached files $K$. (a) $E[T_1]$ when 80% of all nodes obtain all files, (b) $E[T_2]$ when all nodes obtain 80% of all files, (c) $E[T_3]$ when all nodes obtain all files. The dashed lines denote the 1 standard deviation upper and lower bounds from the mean value.

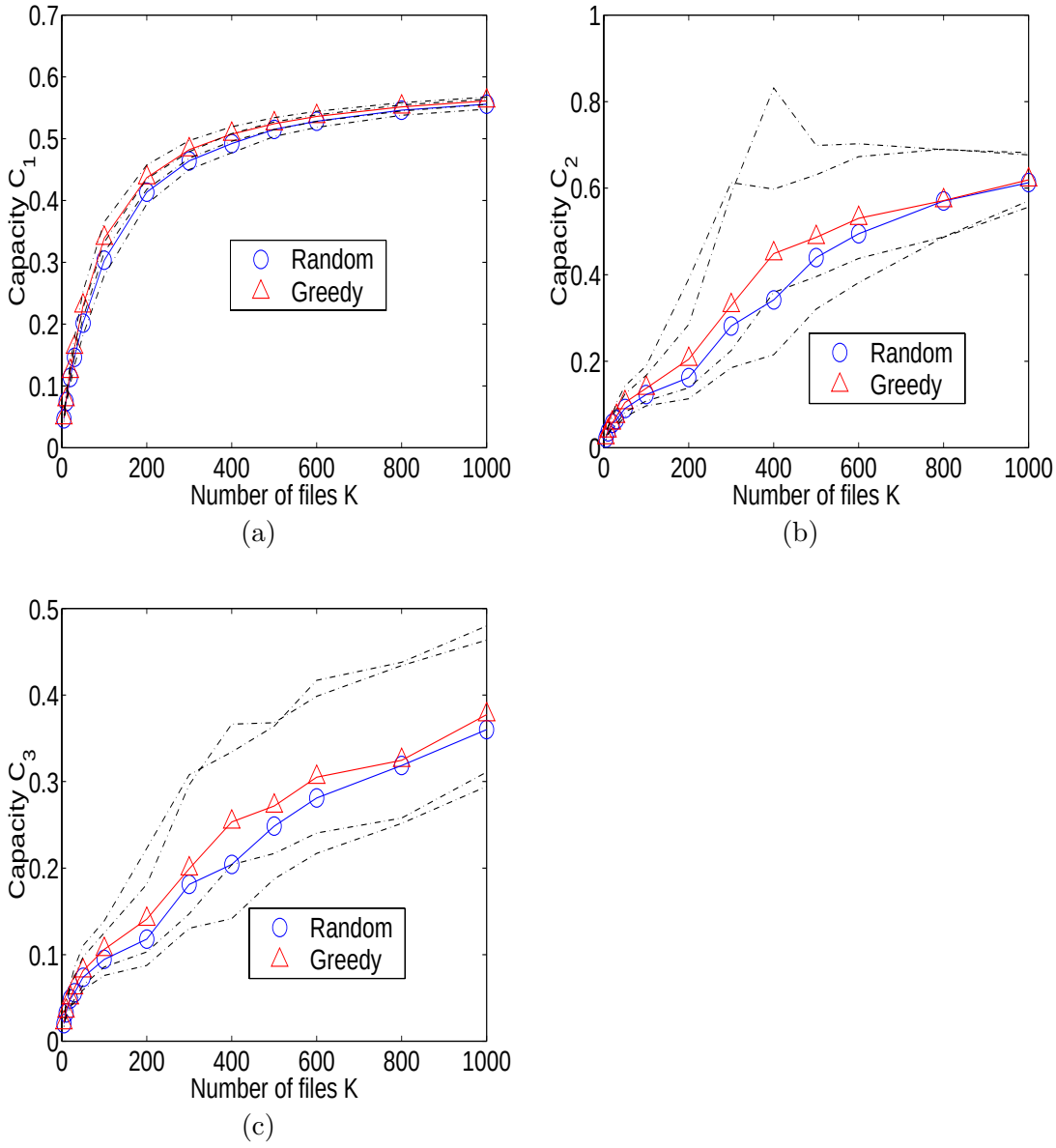



Figure 2.6: Throughput capacity vs. the number of cached files $K$. (a) $C_1$ when 80% of all nodes obtain all files, (b) $C_2$ when all nodes obtain 80% of all files, (c) $C_3$ when all nodes obtain all files. The dashed lines denote the 1 standard deviation upper and lower bounds from the mean value.

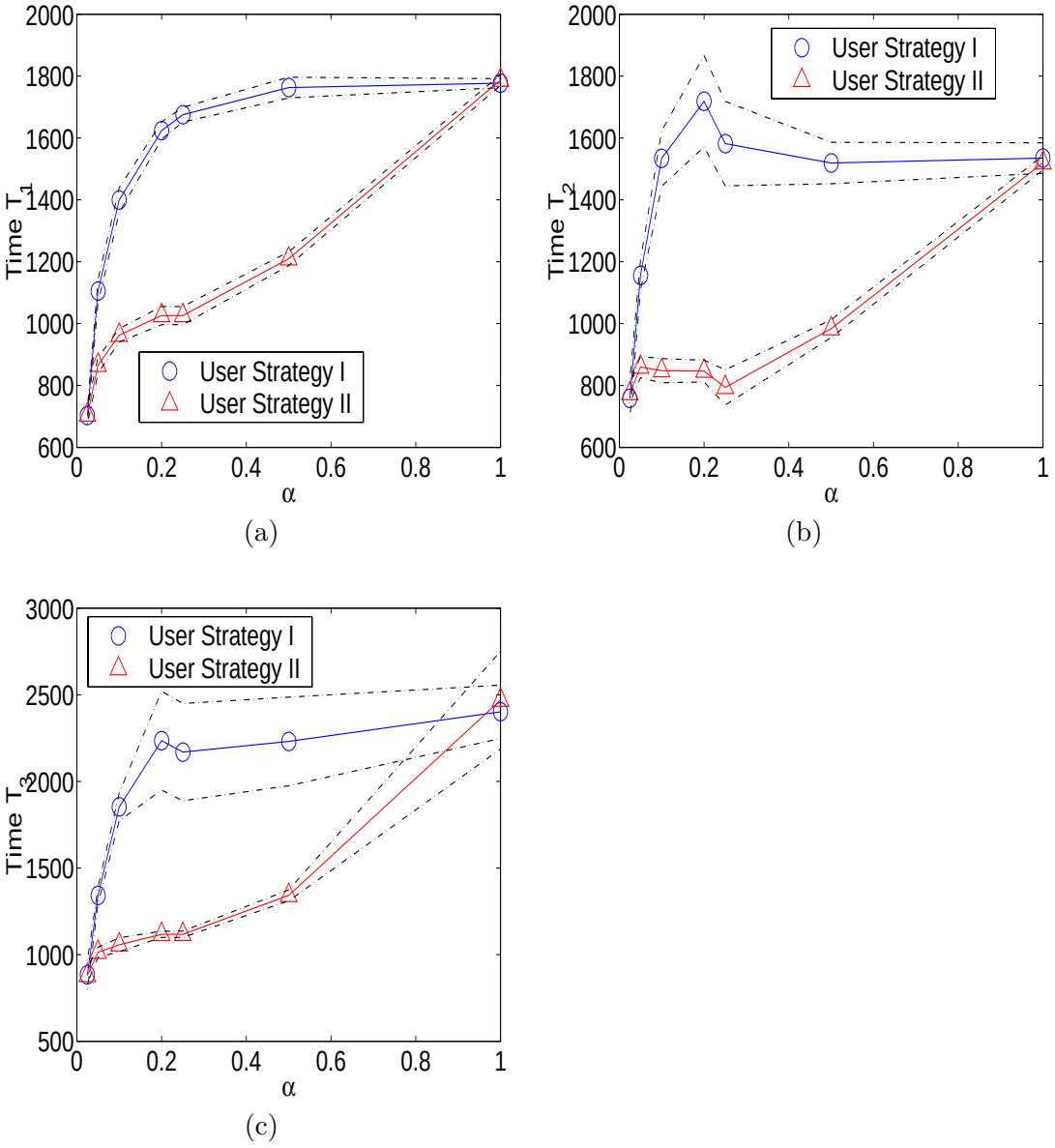



Figure 2.7: Average networking time vs. the fraction of interested files $\alpha$. (a) $E[T_1]$ when 80% of all nodes obtain all files, (b) $E[T_2]$ when all nodes obtain 80% of all files, (c) $E[T_3]$ when all nodes obtain all files. The dashed lines denote the 1 standard deviation upper and lower bounds from the mean value.

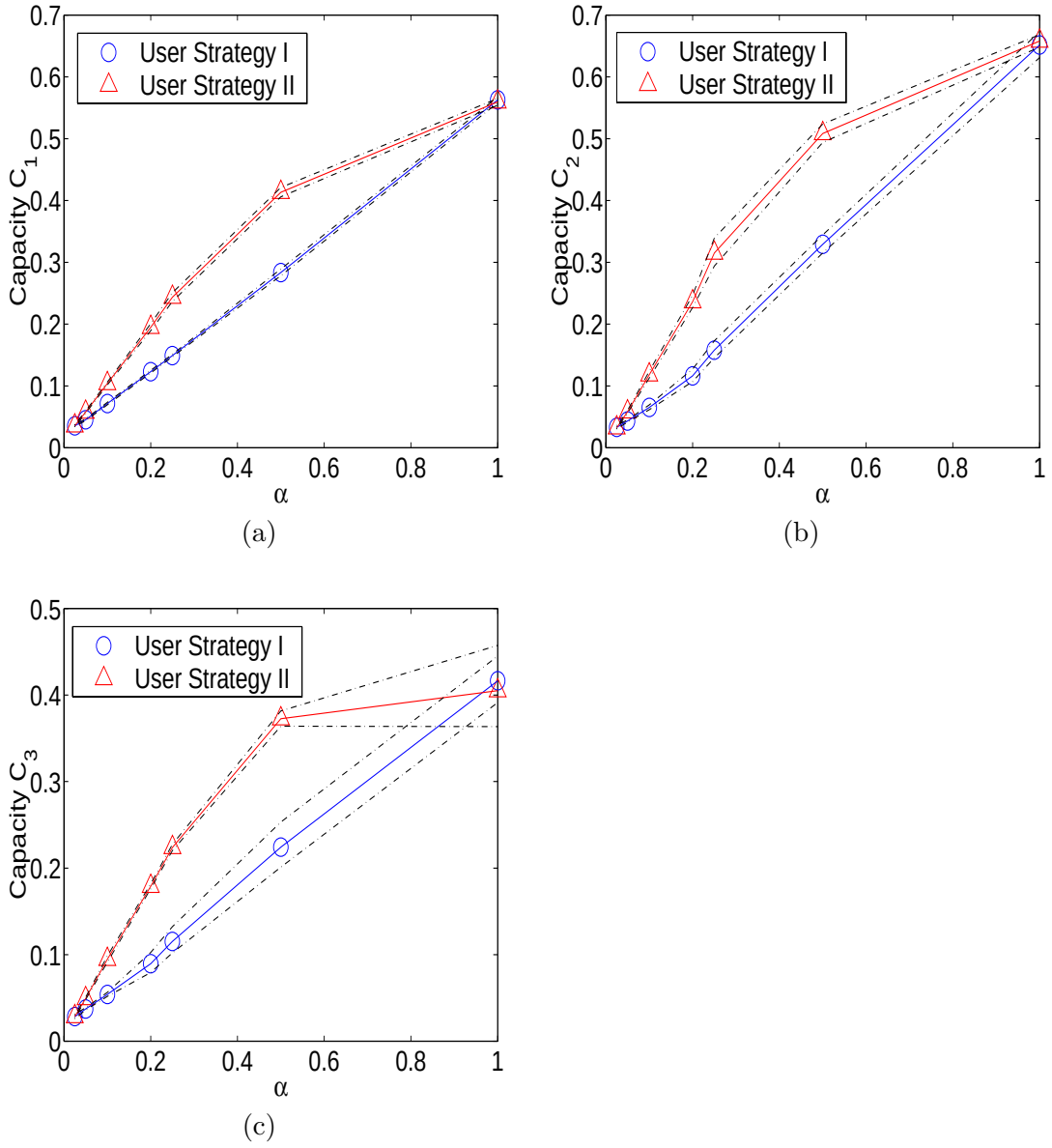



Figure 2.8: Throughput capacity vs. the fraction of interested files $\alpha$. (a) $C_1$ when 80% of all nodes obtain all files, (b) $C_2$ when all nodes obtain 80% of all files, (c) $C_3$ when all nodes obtain all files. The dashed lines denote the 1 standard deviation upper and lower bounds from the mean value.

# Chapter 3

# Optimum Transmit Range and Capacity of Mobile Infostation Networks

## 3.1 Introduction

In the mobile infostation literature, the concept of physical proximity is not well characterized. In [103], it assumed that the planar network consists of discrete locations, in which any two collocated nodes can participate a file exchange. Physical proximity is defined in terms of a hypothetical grid of discrete points, leading to an overly simplified mobility and interference model. On the other hand, [20] assumed that a *candidate transmit node* always transmits to the closest receive node. Although the transmit and receive node pair has the shortest distance, this strategy may not perform well since this distance may be large in some pathological topology realizations. In these links, the benefit of spatial transmission concurrency may be more than offset by a simultaneous increase in total interference power in the network. It may be worthwhile to suppress the transmissions when the channel is less excellent, even though the receive node is the node closest in distance. The resultant decrease in total interference power due to the suppression of transmissions in the less excellent channels may be beneficial to the sum rate of the remaining connections. To ensure that only excellent channels are used, a natural strategy will be imposing an artificial *transmit range* for all nodes. A candidate transmit node will schedule a transmission only if it sees some receive nodes in its transmit range. We note that this definition of transmit range is different from the meaning in the usual sense. In cellular networks, a transmit range of $r_0$ usually refers to the fact that the SIR $\gamma$ at the transmit range boundary marginally meet an outage requirement. That is, the transmit range is a constraint imposed by the physical environment and a myriad of communication technologies. In our context

of mobile infostations, a transmit node may well see many receive nodes beyond the transmit range due to the physical proximity of nodes. However, we impose this artificial *transmit range* and block all these potential transmissions, though the channels are perfectly fine. Here we explicitly trade spatial transmission concurrency for greater spectral efficiency of the remaining connections in the network. As far as the transmit node is concerned, all nodes within the transmit range are its *neighbors*. It is desirable to see if the stipulation of an artificial transmit range will further improve the network capacity.

The rest of the chapter is organized as follows. In section 3.2, we describe the system model, the four strategies and the performance metric. In section 3.4.1, four transmission strategies are compared on the basis of capacity maximization. We identify the scaling invariance property of the network in section 3.4.2 and compare the optimal parameters of the four transmission strategies in section 3.4.3. Finally, we discuss the implications of our results and wrap up in section 3.6.

## 3.2 System Model

We assume nodes populate a planar region according to a homogeneous spatial Poisson process with constant intensity $\lambda$, otherwise known as the node density. Time is divided into slots. In each slot, a fraction $\theta$ of all mobile nodes are randomly selected as *candidate transmit nodes*. This ensures that the point processes for the candidate transmit nodes and receive nodes are Poisson, with average node density $\lambda\theta$ and $\lambda(1-\theta)$ respectively [42]. The Poisson assumption of the transmit nodes is needed to facilitate the computation of interference statistics.

We consider a sender-centric transmission model for the nodes. A candidate transmit node transmits when there are receive nodes within a ring of radius $r_0$. Referring to the example of Figure 3.1, three candidate transmit nodes (T1 to T3) have receive nodes in their transmit range and therefore proceed with transmission. The remaining candidate transmit nodes (T4 to T7) cannot find any receive node and remain silent in the time slot. If there are more than one receive node in range, say T3, it may select

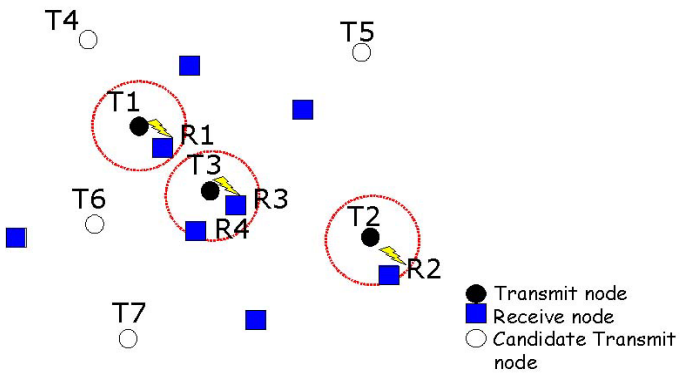



Figure 3.1: A network populated with candidate transmit nodes and receive nodes. A candidate transmit node attempts a transmission if there are receive nodes in its transmit range.

a receive node randomly, or the closest receive node in range R3, and initiate data transmission. It may happen two transmit nodes select the same receive node simultaneously, which is not a problem for receivers that can capture more than one packet. However, in a receiver-centric model two receive nodes may select the same transmit node to initiate data transmission. A conflict resolution mechanism is needed that preclude performance analysis. It was shown in [20] that a receiver-centric transmission model yields a slightly higher SIR stochastically. For the sake of tractable analysis however, we employ the sender-centric transmission model.

We assume all nodes transmit at the same power. The network is interference limited and background noise at a receive node is neglected. In the absence of noise, the SIR at a receive node is independent of the transmit power. In our subsequent analysis, we will therefore assume without loss of generality that each node has a normalized power of 1. The path gain $g(r)$ of a signal is solely determined by the distance $r$ between a transmitter and receiver. Second order effects such as shadowing and multipath fading [89] are ignored. We assume that interference are non-coherently combined at the receive node and treat the total interference power as the sum of the interference power of a Poisson field of interferers. Denote the distance of node $i$ to its intended receive node $j$ as $r_i$. The SIR at the receive node $j$ is thus

$$\gamma_j = \frac{g(r_i)}{Y} = \frac{g(r_i)}{\sum_{k \neq i} g(r_k)}, \tag{3.1}$$

where $Y$ is the summation of interference power contributions from all interference transmitters. Moreover, each point in the plane sees the same interference statistics due to the spatial invariance of homogeneous Poisson process. Hence we drop the index $j$ in the SIR $\gamma$ in subsequent analysis to emphasize that the receive SIR at any arbitrary receive node is the same.

We have looked into four transmission strategies and compare them in the metric of *expected capacity per unit area per unit bandwidth* $E[C]$, in the unit $bit/s/Hz/m^2$. Here the notion of capacity is defined in a loose sense. The theoretical capacity of the strategies are computed under the assumption of single-user receiver decoding. The capacity represents an upper bound performance of a particular transmission and reception strategy and should not be confused with the maximum network capacity over all possible networking and decoding strategies. The capacity of a particular strategy can be contrasted to the packet success rate of practical systems, which is discussed in section 3.5.

Prior work [20,24] assumed that the network area is fixed while the number of nodes is varied. The network capacity is well defined. However, here we assume an infinite size network scattered with a Poisson field of nodes. It is therefore more appropriate to discuss the capacity per unit area instead. On the other hand, the effect of bandwidth scaling is not investigated in this chapter. Without loss of generality we assume our system operates on unit bandwidth. Our performance metric then becomes *expected capacity per unit area*, or *expected spectral efficiency per unit area*, which is used interchangeably in this work. Mathematically, the capacity per unit area is written as

$$E[C] = E[\lambda_t \log_2(1 + \frac{g(R)}{Y})], \tag{3.2}$$

where $\lambda_t$ is the node density of the transmit nodes, and $\log_2(1 + \frac{g(R)}{Y})$ is the capacity of a link with SIR $g(R)/Y$ at unit bandwidth. Note that the expectation is taken over the random variables $R$ the distance of the communication node pair and total interference power $Y$. Our aim is to determine the optimum transmit range $r_0$ and the fraction of candidate transmit nodes $\theta$ based on the objective $E[C]$.

We investigate four transmission strategies in this chapter: a non-adaptive strategy,

a random node in range strategy, a closest node in range strategy and the closest node strategy. In the non-adaptive strategy, the transmission rate is determined by the SIR at the transmit range boundary, i.e.

$$\gamma(r_0) = \frac{g(r_0)}{Y}. \tag{3.3}$$

Even if the SIR is higher when two nodes are closer than distance $r_0$, the additional link capacity warranted by the higher SIR is not exploited. We denote the performance metric of the non-adaptive strategy as $E[\underline{C}]$ to allude that this strategy provides a lower performance bound to the four strategies.

Both the random node in range and the closest node in range strategies operate on the assumption of adaptive transmission. While the transmit power of all nodes is kept constant, the transmission rate is varied so that the link capacity is fully utilized. As the name implies, in the random node in range strategy a candidate transmit node randomly selects a receive node for transmission when multiple receive nodes are within its range. In the closest node in range strategy, the best channel is exploited and the closest node in range is selected. In the case there are no receive nodes in the range of a candidate transmit node, as are all the transparent nodes in Figure 3.1, no transmission is scheduled. It is obvious the latter strategy has superior performance since the candidate transmit node always selects the receive node with the best SIR and link capacity. We denote the performance metric as $E[\overline{C}]$ to emphasize that this strategy provides an upper performance bound of all the four strategies. The corresponding metric for the random node in range strategy is denoted as $E[C_{rand}]$.

We also examine a reference strategy with an unconstrained transmit range, i.e. $r_0 \to \infty$. A candidate transmit node always transmit to the closest receive node even though it may be far away in some pathological topology realizations. This strategy is similar to the strategy in [20], though there is no consideration of rate adaptation in that paper. For the sake of fair comparison, however, we assume the reference strategy is rate adaptive in this chapter. Hereafter, we refer to this strategy as the Grossglauser-Tse (GT) strategy. The corresponding capacity per unit area is denoted as $E[C_{GT}]$. Since there is no transmit range for this strategy, we optimize $E[C_{GT}]$ over $\theta$.

### 3.3 Interference Modeling

In order to compute $E[C]$ we need to derive the PDF of the interference power $Y$ and connection distance $R$ of the node pair. We assume that interference is non-coherently combined at the receiver, and treat the interference of each node as white noise. The interference power of all transmit nodes adds up to a total interference power $Y$. Instead of working with $Y$ directly we consider $Y_a$ the interference power at an arbitrary receive node from the transmit nodes within a radius $a$. $Y_a$ is the total interference power of a random number of interferers $N(a)$ with random interference $g(R_k)$. Mathematically we write

$$Y_a = \sum_{k=1}^{N(a)} g(R_k). \tag{3.4}$$

A standard approach to deal with a random sum of random variables involves manipulations in the transform domain using moment generating or characteristic functions [15]. Since it is possible that the total interference power of an infinite number of nodes to be unbounded when $a$ goes to infinity, characteristic function must be used to avoid the problem of divergence. This technique was used by [87] to determine the PDF of the interference power in a Poisson field of interferers.

Our derivation of the interference statistics closely parallels that in [87], with node density $\lambda$ replaced by the transmit node density $\lambda_t$ to denote the point process of the transmit nodes. Suppose the transmit range of all nodes is $r_0$. A candidate transmit node transmits if the number of receive nodes in its range $N(r_0)$ is non zero. Thus, the transmit node density $\lambda_t$ is

$$\lambda_t = \lambda\theta Pr[N(r_0) > 0] \tag{3.5}$$

$$= \lambda\theta(1 - e^{-\lambda(1-\theta)\pi r_0^2}). \tag{3.6}$$

Assume that the transmit nodes are Poisson, the number of nodes in the area $N(a)$ is also Poisson with node density $\lambda$. $Y_a$ is then evaluated by conditioning on $N(a)$, the Poisson number of interferers in the area in the transform domain. After some integrations and algebraic manipulations, the total interference power in the Poisson field of interferers is obtained by taking the limit $a \to \infty$.

In this work, we only consider the two ray ground reflection model. The path gain is given by $g(r) = r^{-4}$. The exact derivation of random sum in (3.4) is outlined in [87] and is not repeated here. The PDF is only dependent on the transmit node density, given by

$$f_Y(y) = \frac{\pi}{2}\lambda_t y^{-3/2} e^{-\frac{\pi^3\lambda_t^2}{4y}}. \tag{3.7}$$

The corresponding CDF is

$$F_Y(y) = \text{erfc}\Big(\frac{\pi^{3/2}\lambda_t}{2\sqrt{y}}\Big), \tag{3.8}$$

where $\text{erfc}(x)$ is the complementary error function commonly used in probability of error calculations in digital communications systems [68], given by

$$\text{erfc}(x) = 1 - \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,dt. \tag{3.9}$$

Although both candidate transmit nodes and receive nodes are Poisson distributed, the transmit nodes are not Poisson distributed in general. Suppose two candidate transmit nodes $i$ and $j$ are close to each other at points $dA_i$ and $dA_j$. If $i$ is a transmit node, it implies there is at least one receive node, say $k$ is in its range. Since node $j$ is close to $i$, it is likely that $j$ also finds the same receive node $k$ in its range. Therefore, the event that $j$ is a transmit node is not independent of the event that $i$ is a transmit node. This violates the independent increment property of Poisson process since the two points $dA_i$ and $dA_j$ are not overlapping. Nevertheless, for small $r_0$ and $\theta$, the mean distance between two closest candidate transmit nodes is larger than $2r_0$. This reduces the instances of overlapping coverage area between candidate transmit nodes. We have run simple experiments to confirm that the Poisson assumption of transmit nodes applies to the values of $r_0$ and $\theta$ of our interest. We will therefore assume that the transmit nodes are Poisson such that (3.7) applies.

Note that the interference power $Y$ is a random variable with infinite mean and variance. The PDF is solely dependent on the system parameter $\lambda_t$, which in turn depends on $r_0$ and $\theta$. As $r_0$ increases, the probability that a node transmits increases. In the limit $r_0 \to \infty$ when there is no constraint on the transmit range, every candidate

transmit node transmits. Thus the total interference power $Y$ of the system strictly increases with $r_0$. On the other hand, no transmissions are possible when $\theta = 0$ or $\theta = 1$ due to the absence of either transmit or receive nodes. An optimal $\theta$ exists such that the density of the transmit nodes is maximized. This can be readily seen by twice differentiating $\lambda_t$ w.r.t. $\theta$.

The transmit node density describes only one dimension of the optimization problem. It is not always desirable for a network to be operated at maximum spatial concurrency to allow every candidate transmit node to transmit. By confining all transmissions to node pairs that have a communication distance less than a stipulated transmit range, the reduction of spatial transmission concurrency can be traded off for more spectral efficiency in individual links. We are now confronting the problem of jointly optimizing $r_0$ and $\theta$ for the maximization of the expected capacity per unit area $E[C]$.

## 3.4 Performance Analysis

### 3.4.1 Capacity Maximization

In the non-adaptive strategy, the SIR $\gamma$ is a function of random interference power only. The expected capacity per unit area is therefore obtained by conditioning on the total interference power $Y$.

$$E[\underline{C}] = E[\lambda_t \log_2(1+\underline{\gamma})] \tag{3.10}$$

$$= \frac{\lambda_t}{\ln 2}\int_0^\infty \ln\Big(1+\frac{g(r_0)}{y}\Big) f_Y(y)dy. \tag{3.11}$$

Evaluating the integral (3.11) yields

$$E[\underline{C}] = -\frac{\lambda_t}{\ln 2}\Bigg(\frac{\pi^3\lambda_t^2 r_0^4}{2}\ {}_2F_2\Big(1,1;\frac{3}{2},2;\frac{\pi^3 r_0^4\lambda_t^2}{4}\Big) + b - \pi\ \mathrm{erfi}\Big(\frac{\pi^{3/2}r_0^2\lambda_t}{2}\Big) + \ln(\pi^3 r_0^4\lambda_t^2)\Bigg), \tag{3.12}$$

where

$$b = \lim_{n\to\infty}\Big(\sum_{k=1}^{n}\frac{1}{k} - \ln n\Big) \approx 0.5772 \tag{3.13}$$

is the Euler's constant and erfi($x$) is the imaginary error function given by

$$\text{erfi}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{t^2} dt. \tag{3.14}$$

${}_pF_q(a_1, ..., a_p; b_1, ..., b_q; x)$ is the generalized Hypergeometric function, a series of the form

$${}_pF_q(a_1, ..., a_p; b_1, ..., b_q; x) = \sum_{k=0}^{\infty} c_k x^k \tag{3.15}$$

for which the ratio of successive terms can be written

$$\frac{c_{k+1}}{c_k} = \frac{(k+a_1)(k+a_2)...(k+a_p)}{(k+b_1)(k+b_2)...(k+b_q)(k+1)} \tag{3.16}$$

and $c_0 = 1$. We observe that both the generalized Hypergeometric function and the imaginary error function diverge as $x$ increases. However, the difference of these two functions is always finite.

The capacity formula $E[\underline{C}]$ involves special functions that does not yield tractable analytical expressions on differentiation w.r.t. the optimization variables $r_0$ and $\theta$. As shown in Figure 3.2, $E[\underline{C}]$ is plotted at the node densities $\lambda = 1, 5, 10, 20$ nodes/$m^2$. Although $E[\underline{C}]$ is not convex, it is fortunate that simple gradient algorithms can still be used to determine the optimal transmit range $r_0$ and fraction of candidate transmit nodes $\theta$ for each value of node density $\lambda$. As shown in Figure 3.4, the optimal transmit range, fraction of candidate transmit nodes, the expected number of nodes within the transmit range and the expected capacity per $m^2$ are plotted versus node density.

When the transmission strategy is rate adaptive, the SIR $\gamma$ is dependent on both the interference power $Y$ and the distance of the receive node from the candidate transmitter $R$. The expected capacity per unit area is obtained by conditioning on both the interference power $Y$ and communication distance $R$. Given there exists a non-zero number of nodes $N(r_0)$ in the coverage radius, we define $R$ as the distance to the receive node to which we communicate. We denote the PDF of the connection distance $R$ as $f_R(r|n)$. It is implicitly understood that the number of receive nodes in the transmit range $N(r_0)$ is non zero when a transmission is attempted. On the other hand, the PDF may be dependent on the number of receive node $n$ in the transmit range.

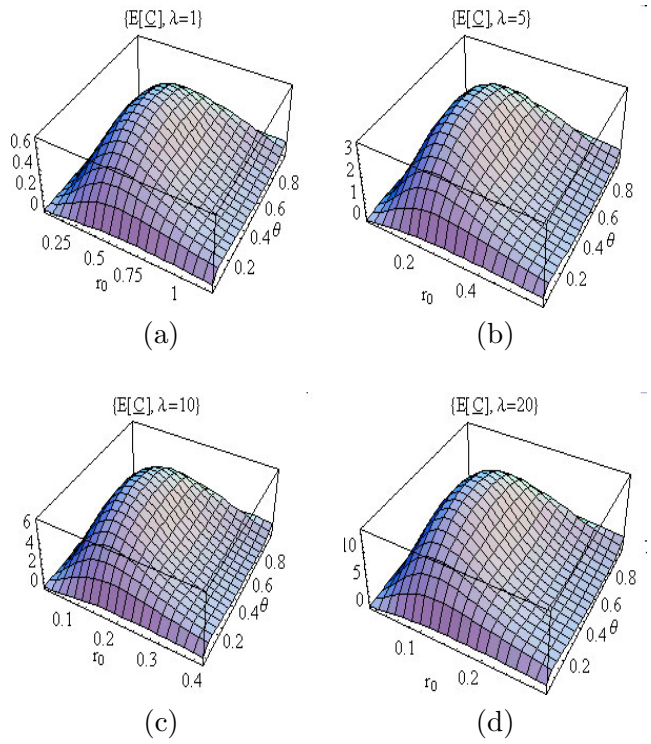



Figure 3.2: $E[C]$ vs. transmit range $r_0$ and fraction of candidate transmit nodes $\theta$. (a) $\lambda = 1/m^2$, (b) $\lambda = 5/m^2$, (c) $\lambda = 10/m^2$, (d) $\lambda = 20/m^2$.

Since the receive nodes are Poisson distributed with intensity

$$\lambda_r = \lambda(1-\theta), \tag{3.17}$$

each receive node within the range is uniformly located in the area $\pi r_0^2$. In the random node in range strategy, the distance between the random receive node and the transmit node therefore has a PDF

$$f_R(r) = \begin{cases} 2r/r_0^2 & 0 \le r \le r_0 \\ 0 & \text{o.w.} \end{cases} \tag{3.18}$$

independent of $n$. For the closest node in range strategy, the distance is the minimum of among the receive node distances. It is straightforward to deduce

$$f_{R|N(r_0)=n}(r|n) = \frac{2nr}{r_0^2}\left(1 - \left(\frac{r}{r_0}\right)^2\right)^{n-1}. \tag{3.19}$$

The PDF of the distance to the closest receive node is then computed by conditioning

on $n$ the number of receive nodes in range

$$f_R(r) = \sum_{n=1}^{\infty} f_{R|N(r_0)=n}(r|n) Pr[N(r_0)=n] \tag{3.20}$$

$$= \frac{2\lambda_r \pi r e^{-\lambda_r \pi r^2}}{1-e^{-\pi\lambda_r r_0^2}} \qquad 0 \le r \le r_0. \tag{3.21}$$

In the GT strategy, a candidate transmit node always transmits. Taking the limit $r_0 \to \infty$ to (3.21), the PDF of the connection distance $f_R(r)$ with an unconstrained transmit range is

$$f_R(r) = 2\pi r \lambda_r e^{-\lambda_r \pi r^2} \qquad 0 \le r < \infty. \tag{3.22}$$

For the above adaptive strategies, the expected sum rate per unit area $E[C]$ is then computed as

$$E[E[\lambda_t \log_2(1+\gamma(R,Y))]] \tag{3.23}$$

$$= \frac{\lambda_t}{\ln 2} \int_0^{r_0} \int_0^{\infty} \ln\Big(1+\frac{g(r)}{y}\Big) f_Y(y) dy f_R(r) dr \tag{3.24}$$

$$= \int_0^{r_0} -\frac{\lambda_t}{2\ln 2}\Bigg( (\pi^3 \lambda_t^2 r^4) \ {}_2F_2\Big(1,1;\frac{3}{2},2;\frac{\pi^3 r^4 \lambda_t^2}{4}\Big) + 2\Big(b - \pi \ \mathrm{erfi}\Big(\frac{\pi^{3/2} r^2 \lambda_t}{2}\Big) + \ln(\pi^3 r^4 \lambda_t^2)\Big)\Bigg) f_R(r) dr, \tag{3.25}$$

where $f_R(r)$ assumes the form of (3.18), (3.21), (3.22) for the three adaptive strategies. In the random node in range strategy, $E[C]$ can be evaluated as

$$E[C] = \frac{1}{2r_0^2 \ln 2}\Bigg[ 2\pi r_0^2 \lambda_t \mathrm{erfi}\Big(\frac{\pi^{3/2} r_0^2 \lambda_t^2}{2}\Big) - \frac{\pi^3 r_0^6 \lambda_t^3}{3} \ {}_2F_2\Big(1,1;2,\frac{5}{2};\frac{\pi^3 r_0^4 \lambda_t^2}{4}\Big) - \frac{2}{\pi}\Big( -2(1-e^{\frac{\pi^3 r_0^4 \lambda_t^2}{4}}) + \pi r_0^2 \lambda (b-2) + \lambda_t \pi r_0^2 \ln(\pi^3 r_0^4 \lambda_t^2)\Big)\Bigg]. \tag{3.26}$$

With reference to Figure 3.3, although $E[C_{rand}]$ is not generally convex w.r.t. $r_0$ and $\theta$, it is straightforward to use simple gradient algorithms to optimize the system parameters. For the closest node within range and the GT strategy, (3.25) cannot be evaluated analytically and numerical integration must be used. Nevertheless, the

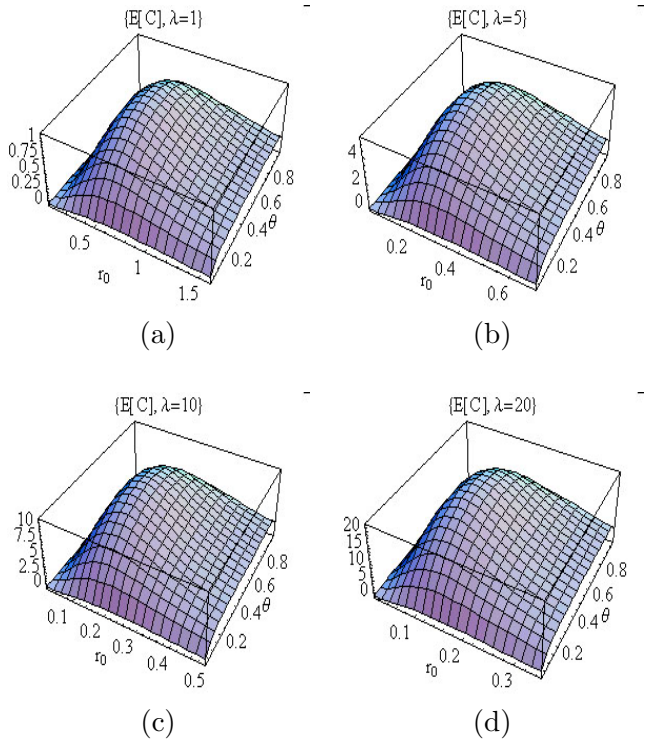



Figure 3.3: $E[C_{rand}]$ vs. transmit range $r_0$ and fraction of candidate transmit nodes $\theta$. (a) $\lambda = 1/m^2$, (b) $\lambda = 5/m^2$, (c) $\lambda = 10/m^2$, (d) $\lambda = 20/m^2$.

optimum system parameters that maximize $E[\overline{C}]$ and $E[C_{GT}]$ can be determined by gradient algorithms.

### 3.4.2 Optimum Transmit Range and Scaling Invariance

The existence of an optimal range for capacity maximization is intuitively obvious. When the transmit range is too large, a transmit node may connect to a receive node that is not close. Although there are more simultaneous transmissions over an area, the increase in the mutual interference reduces the achievable rate for each transmit receive node pair considerably. On the other hand, when the transmit range is too small, only node pairs in close proximity transmits. High spectral efficiency of individual links can be obtained due to the reduction of interference power. Few candidate transmit nodes actually transmits, however, since very few receive nodes are very close to the candidate transmit nodes. Thus, the potential spatial transmission concurrency is not

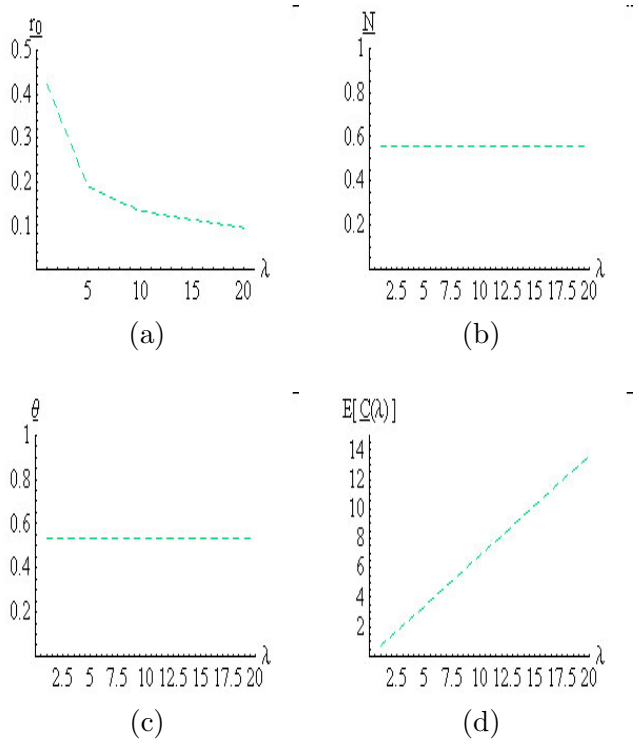



Figure 3.4: Optimized non-adaptive strategy at different node density $\lambda$. (a) transmit range $r_0$ vs. node density $\lambda$, (b) expected number of nodes in range $N$ vs. node density $\lambda$, (c) fraction of candidate transmit nodes $\theta$ vs. node density $\lambda$, (d) expected capacity per unit area vs. node density $\lambda$.

fully utilized, leading to a poor capacity per unit area usage.

A couple of interesting observations can be made in Figure 3.4. First, the optimal range $r_0$ shrinks as node density increases. As node density increases, it is more likely for a transmit node to find receive nodes at a smaller range. A decrease in the transmit range does not adversely affect the number of simultaneous transmissions in the network. Moreover, the optimal range $r_0$ shrinks in a way such that the expected number of neighbors of a candidate transmit node $N$ is constant, as shown in Figure 3.4(b). Similarly, Figure 3.4(c) shows that the optimal fraction of transmit nodes $\theta$ is also invariant to node density. Finally, the expected capacity per unit area is linearly increasing with node density. These observations are inter-related and can be explained using the *rescaling* argument drawn from continuum percolation theory [51].

A percolation model is characterized by a point process and a connectivity function.

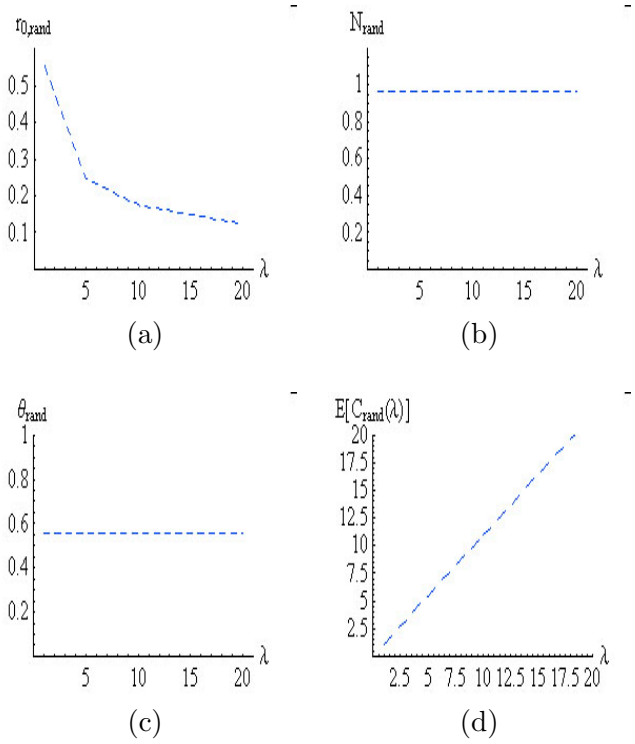



Figure 3.5: Optimized Random Node in Range Strategy at different node density $\lambda$. (a) transmit range $r_0$ vs. node density $\lambda$, (b) expected number of nodes in range $N$ vs. node density $\lambda$, (c) fraction of candidate transmit nodes $\theta$ vs. node density $\lambda$, (d) expected capacity per unit area vs. node density $\lambda$.

In our context of a homogeneous spatial Poisson process, the point process is completely characterized by the node density $\lambda$. A connectivity function, on the other hand, specifies the probability that a link exists between two nodes as a function of distance $r$ between them. Here we are using the on-off random connection model, in which two nodes are connected w.p. 1 when their distance is less than $r$, which is the same as our artificial transmit range $r_0$. We denote our percolation model as $\Pi(\lambda, r)$. Any network topology with node density $\lambda$ and transmit range $r$ is therefore a realization of the percolation model $\Pi(\lambda, r)$.

With reference to Figure 3.8, realizations of two percolation models $\Pi(\lambda_1(\theta_1), r_1)$ and $\Pi(\lambda_2(\theta_2), r_2)$ are drawn. The two realizations are *coupled* in the sense that the second realization is exactly identical to the first except for the distance scaling in the

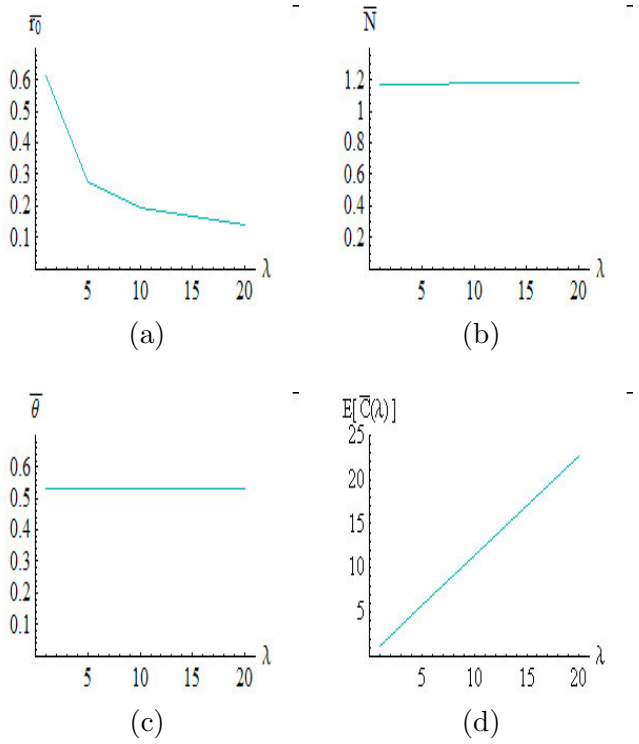



Figure 3.6: Optimized Closest Node in Range Strategy at different node density $\lambda$. (a) transmit range $r_0$ vs. node density $\lambda$, (b) expected number of nodes in range $N$ vs. node density $\lambda$, (c) fraction of candidate transmit nodes $\theta$ vs. node density $\lambda$, (d) expected capacity per unit area vs. node density $\lambda$.

2-dimensional space. Accordingly, the following rules must be satisfied.

$$\theta_1 = \theta_2 \tag{3.27}$$

$$\lambda_1 A_1 = \lambda_2 A_2 \tag{3.28}$$

$$\lambda_1 r_1^2 = \lambda_2 r_2^2 \tag{3.29}$$

Equation (3.27),(3.28) and (3.29) express the conservation of the fraction of transmit nodes, number of nodes in the network area, and number of neighbors of an arbitrary node $N$. These rules must be observed if the two realizations are really scaled version of each other. Note that the two topology realizations have exactly the same connectivity structure. The SIR of an arbitrary link in realization 1, and the associated link capacity, must be identical to that of the corresponding link in realization 2. Since the capacity of a link depends only on the SIR at the receive node, the equivalence of link SIR in two coupled realizations implies that both realizations have same sum capacity.

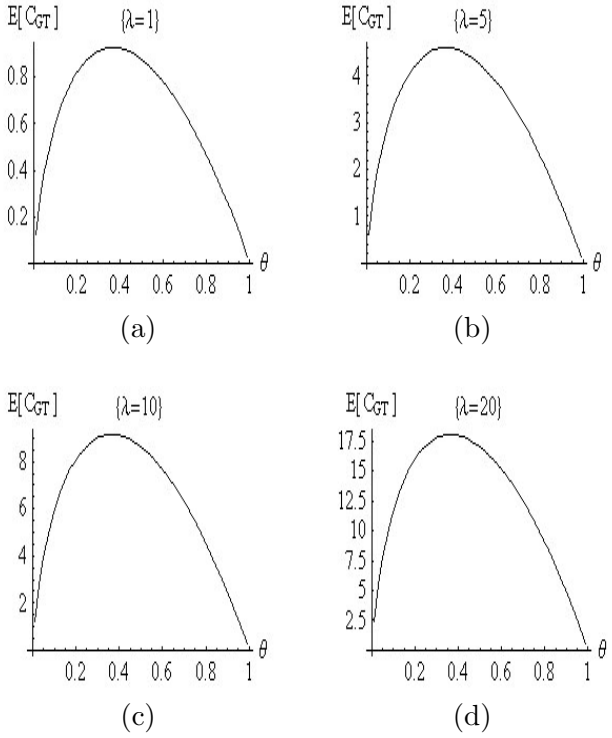



Figure 3.7: $E[C_{GT}]$ vs. the fraction of transmit nodes $\theta$. (a) $\lambda = 1/m^2$, (b) $\lambda = 5/m^2$, (c) $\lambda = 10/m^2$, (d) $\lambda = 20/m^2$.

Denote $c(A_i), i = 1, 2$ as the sum capacity of realization 1 and 2, where $A_i$ is the network size of realization $i$. Using the technique of coupling, for each realization of one percolation model $\Pi(\lambda_1, r_1)$, we can always find an equivalent realization in the other percolation model $\Pi(\lambda_2, r_2)$ that is a rescaled version of the first. Taking the expectation over all realizations, we deduce that

$$E[c(A_1)] = E[c(A_2)] \tag{3.30}$$

Suppose $\theta_1$ and $r_1$ jointly maximize $E[c(A_1)]$. From rescaling we know that the optimal $\theta_2$ and $r_2$ maximizes $E[c(A_2)]$ must satisfy $\theta_1 = \theta_2$ and $\lambda_1 r_1^2 = \lambda_2 r_2^2$. That is, the number of neighbors of a node $N$, and the fraction of transmit nodes $\theta$ are constant.

We just observe that the expected sum capacity over all realizations of two scaled percolation models is the same. It is instructive to determine analytically if the expected SIR of an arbitrary link of length $r_0$ is also invariant to node density, with expectation taken over all realizations of the Poisson field of interferers. The expected SIR at a

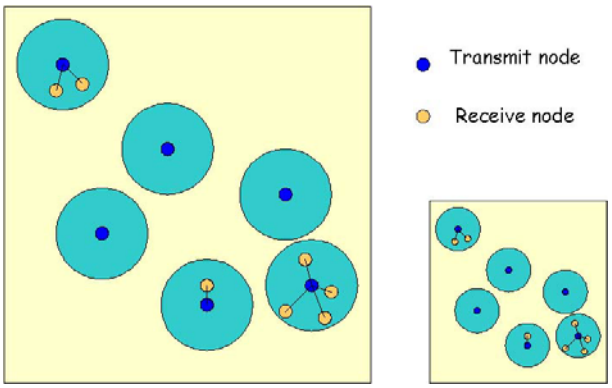



Figure 3.8: Illustration of rescaling of two coupled percolation models.

receive node at the transmit range boundary is

$$E[\gamma(r_0)] = \int_0^\infty \frac{g(r_0)}{y} f_Y(y)dy \tag{3.31}$$

$$= \frac{2}{\pi^3 r_0^4 \lambda_t^2} \tag{3.32}$$

$$= \frac{2}{(1-e^{-N(1-\theta)})^2 N^2 \pi \theta^2}, \tag{3.33}$$

where $N = \lambda \pi r_0^2$ is the average number of neighbors of a node.

More generally, if the receive node is at a distance $\alpha r_0, 0 < \alpha < 1$ from the transmit node, the expected SIR is

$$E[\gamma(\alpha r_0)] = \int_0^\infty \frac{g(\alpha r_0)}{y} f_Y(y)dy \tag{3.34}$$

$$= \frac{2}{(1-e^{-N(1-\theta)})^2 N^2 \pi \alpha^4 \theta^2}. \tag{3.35}$$

Thus, if the distance $r$ of any transmit and receive node pair relative to the transmit range $r_0$ is constant, i.e. $r/r_0 = \alpha$, it follows from (3.35) that the expected SIR of of the connection is invariant to node density.

The linear increase in expected capacity per unit area is a direct consequence of rescaling in percolation models. The corresponding capacity per unit area for percolation model 1 and 2 are $c(A_1)/A_1$ and $c(A_2)/A_2$. Taking expectations over all coupled realizations, we have

$$E[C_2] = E[C_1]\frac{A_1}{A_2} = E[C_1]\frac{\lambda_2}{\lambda_1}. \tag{3.36}$$

Let $\lambda_1 = 1$, we obtain

$$E[C_2] = \lambda_2 E[C_1]. \tag{3.37}$$

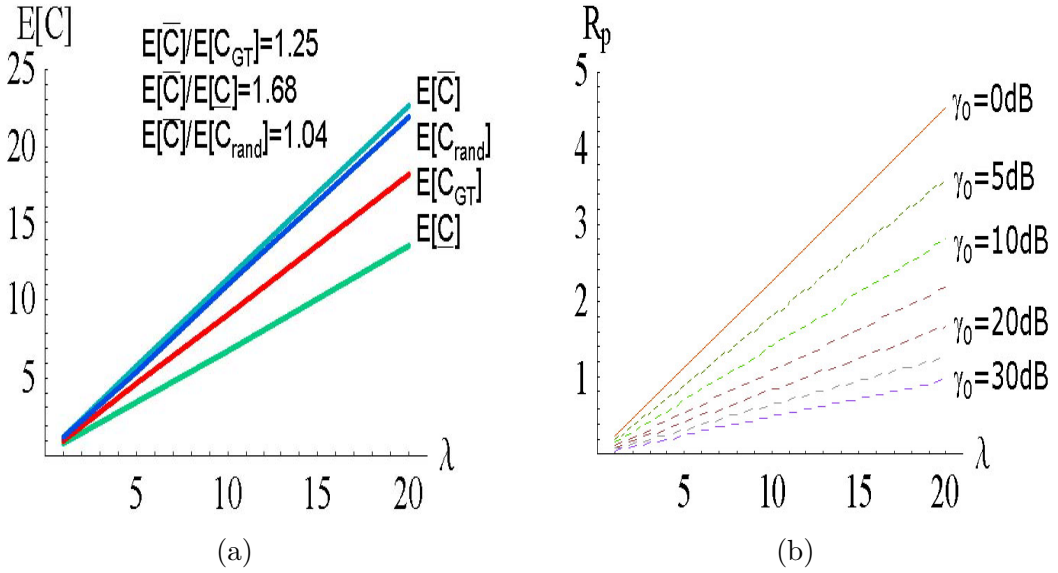



Figure 3.9: The expected sum rate per unit area for theoretical and practical systems as a function of node density $\lambda$. (a) Theoretical capacity per unit area for four strategies. (b) Packet success rate per unit area for practical systems with different SIR threshold $\gamma_0$.

That is, the expected capacity per unit area is linearly increasing with node density. The slope corresponds to the expected capacity per unit area when node density is 1.

For the adaptive strategies, similar observations to the non-adaptive strategies can be made. With reference to Figure 3.5 and Figure 3.6, the optimal range shrinks for both the random node and closest node in range strategies. The expected number of nodes in range, fraction of candidate transmit nodes are also invariant with node density. Similarly, in the GT strategy, the optimal fraction of candidate transmit nodes seems to be invariant to node density as shown in Figure 3.7. Moreover, these capacity curves agree with those obtained in [20]. The optimal value occurs around $\theta = 0.36$. The rescaling argument applies to all the adaptive strategies as well. Thus, $\theta$ and $N$ are invariant with node density whereas the expected capacity per unit area is linear with with node density.

In Figure 3.9(a), we compare the capacity of the four strategies. The non-adaptive strategy has the worst performance as expected, with $E[\overline{C}]/E[\underline{C}] = 1.68$. The closest node within range strategy outperforms the random node within range strategy by a

small margin ($E[\overline{C}]/E[C_{rand}] = 1.04$). At the optimal range, the average number of nodes within the transmit range is between 0.6 to 1.2 for the four strategies. Thus, most of the time a random node is exactly the same as the closest node. This explains the close performance of the two strategies. The GT strategy, however, has a capacity that is almost halfway between the closest neighbor in range strategy and the non-adaptive strategies, with $E[\overline{C}]/E[C_{GT}] = 1.25$. Although the GT strategy is rate adaptive, an unconstrained transmit range allows connection to a distant receive node in some pathological cases. By stipulating a transmit range that excludes transmissions to distant nodes, only good channels are exploited and network interference is reduced.

### 3.4.3 Optimum Point of Network Operation

It is also instructive to compare the optimal values of the fraction of transmit nodes $\theta$, number of neighbors $N$ of a node and the probability of a node transmission for all the strategies. A node is selected as a candidate transmitter with probability $\theta$. For the GT strategy, a candidate node always transmits to the closest neighbor. Thus the probability that an arbitrary node transmits is $\theta$. For the random and closest node in range strategies, the candidate node transmits with probability $(1-e^{-(1-\theta)N})$ when the transmitter sees some receive nodes is in range. Thus an arbitrary node transmits with probability $\theta(1-e^{-(1-\theta)N})$ The values for the four strategies are summarized in Table 3.1. We observe that the optimal value of $\theta$ is close to 0.5 in all strategies. A connection

| Capacity | $E[\underline{C}]$ | $E[C_{rand}]$ | $E[\overline{C}]$ | $E[C_{GT}]$ |
|---|---|---|---|---|
| $\theta$ | 0.533 | 0.555 | 0.531 | 0.364 |
| $N$ | 0.558 | 0.964 | 1.17 | n/a |
| $\theta(1-e^{-(1-\theta)N})$ | 0.1223 | 0.1936 | 0.2243 | $\theta = 0.364$ |

Table 3.1: Optimized parameters for the four strategies.

is made up by a transmit and receive node pair. If either kind of nodes are dominant in the network, the scarcity of the other kind of nodes adversely affect the number of transmit and receive node pairs in proximity. A fraction $\theta$ close to 0.5 conforms to our intuition and enables a nice mix of transmit and receive nodes over space for creating

numerous excellent channels. The observation that $\theta$ is slightly larger than 0.5 for all strategies indicates that the transmit nodes has a slightly more influential role in the creation of connections as hinted by the sender-centric approach.

The optimal number of neighbors $N$ increases from 0.56 to 1.17 as we move from the non-adaptive to the closest node in range strategy. The non-adaptive strategy should be operated at a small range, since the link capacity at any point inside the transmit range boundary is not fully utilized. For the adaptive strategies, the random node in range strategy should be operated at a smaller range, to minimize the opportunity cost in case the random node is not the closest receive node. The closest node strategy is not penalized for having a larger transmit range compared to the other two strategies. The shortest link to a receive node is always chosen for connection.

The probability of transmission $\theta(1-e^{-(1-\theta)N})$ also increases from 0.1123 to 0.2243 as we move from the non-adaptive to the closest node in range strategy. Since $\theta$ is similar in the strategies, the probability of transmission is dictated by $N$. The non-adaptive strategy is penalized severely for having a large transmit range. A transmission is attempted when a receive node is close by, at a transmission probability of 0.1223. Thus a sacrifice of spatial transmission concurrency is traded off for more spectral efficiency of individual links. On the other hand, the GT strategy has a large transmission probability, more than 50% larger than the closest node in range strategy. Since all candidate transmit nodes transmit in the GT strategy, maximum spatial concurrency is attained at the expense of increased network interference and decreased spectral efficiency in individual links. The comparison of the four strategies in Figure 3.9(a) shows that both the non-adaptive and the GT strategy have inferior performance versus the adaptive strategies with a stipulated transmit range. This suggests that an optimal tradeoff exists between spectral efficiency and spatial concurrency such that the overall capacity per unit area is maximized.

## 3.5 Packet Success Rate Maximization

The capacity per unit area provides an upper bound on the number of bits per second per unit area for the transmission strategies. Nevertheless practical modulation and error correction schemes do not yield performance that matches the capacity limit. In this section, we investigate the optimal transmit range and fraction of candidate transmitters in practical systems and contrast the results with the performance upper bound.

In practical systems, different multilevel modulation schemes are used in conjunction with error correction codes. Assuming the total interference power at a receiver is constant in a time slot. Each symbol sees the same amount of interference and thus have equal symbol error rate $P_M(\gamma)$ that is a function of the SIR $\gamma$. The symbol may be mapped to binary codewords through simple mapping, such as Gray coding in square constellation MQAM systems [89], where adjacent symbols in the signal space differs by only one bit in the binary codeword. Interleavers may also be used so that the output bits have decorrelated bit error in adjacent bits. Moreover, a $t$-error correcting block code of length may be used. Under these assumptions, the packet success probability $s(\gamma)$ can be computed and is a smooth curve with a slope in the transition region that depends on the error correction capability of the code. For long and good error codes, the packet success probability can be approximated by a unit step function w.r.t. the SIR $\gamma$ [68], with the transition denoted as the SIR threshold $\gamma_0$. This simple abstraction allow for a closed form computation of expected packet success probability for practical systems. Thus, in this chapter we assume the packet success probability $s(\gamma)$ is given by

$$s(\gamma) = s(g(r)/y) \quad = \quad \begin{cases} 1 & 0 \leq y \leq g(r)/\gamma_0 \\ 0 & \text{o.w.} \end{cases} . \tag{3.38}$$

We denote our performance metric as $R_p$, the *expected number of packets delivered per time slot per unit area*, or simply *expected packet success rate per unit area*. This is given by

$$R_p = \lambda_t E[s(\gamma)]. \tag{3.39}$$

where $E[s(\gamma)]$ is the expected packet success probability per node per time slot and $\lambda_t$ is the node density of the transmit nodes.

The determination of $R_p$ for the four strategies is straightforward. For the non-adaptive strategy,

$$\underline{R_p} = \lambda_t \int_0^{\frac{g(r_0)}{\gamma_0}} f_Y(y)dy \tag{3.40}$$

$$= \lambda_t \text{erfc}\Big(\frac{\pi^{3/2} r_0^2 \sqrt{\gamma_0}\lambda_t}{2}\Big). \tag{3.41}$$

For the adaptive strategies, $R_p$ can be evaluated by conditioning on both the interference power $Y$ and distance of the receive node from the transmit node $R$.

$$R_p = \lambda_t \int_0^{r_0} \int_0^{\frac{g(r)}{\gamma_0}} f_Y(y)dy f_R(r)dr. \tag{3.42}$$

Recall that in the random node in range strategy, $f_R(r)$ is uniformly distributed in the coverage area and is given by (3.18). Substitute (3.18) to (3.42), we have

$$R_{p,rand} = \frac{2(1 - e^{-\frac{\pi^3 r_0^4 \gamma_0 \lambda_t^2}{4}})}{\sqrt{\gamma_0}\pi^2 r_0^2} + \lambda_t \text{erfc}\Big(\frac{\pi^{3/2} r_0^2 \sqrt{\gamma_0}\lambda_t}{2}\Big). \tag{3.43}$$

It is obvious that the adaptive random node strategy outperforms the non-adaptive strategy by comparing (3.43) and (3.41). The performance disparity increases as $r_0$ increases. In the closest node in range strategy, $f_R(r)$ is given by (3.21). Substitute (3.21) to (3.42), $\overline{R_p}$ is evaluated as

$$\frac{\lambda_t}{1 - e^{-\pi r_0^2 \lambda_r}} \Bigg[ 1 - e^{-\pi r_0^2 \lambda_r} \text{erfc}\Big(\frac{\pi^{3/2} r_0^2 \sqrt{\gamma_0}\lambda_t}{2}\Big) + e^{\frac{1}{\pi\gamma_0}\left(\frac{\lambda_r}{\lambda_t}\right)^2} \Bigg( \text{erfc}\Big(\frac{1}{\sqrt{\pi\gamma_0}}\left(\frac{\lambda_r}{\lambda_t}\right) + \frac{\pi^{3/2} r_0^2 \sqrt{\gamma_0}\lambda_t}{2}\Big) - \text{erfc}\Big(\frac{1}{\sqrt{\pi\gamma_0}}\left(\frac{\lambda_r}{\lambda_t}\right)\Big)\Bigg)\Bigg]. \tag{3.44}$$

In the GT strategy, $f_R(r)$ is given by (3.22). Thus,

$$R_{p,GT} = \lambda_t \Bigg( 1 - e^{\frac{1}{\pi\gamma_0}\left(\frac{\lambda_r}{\lambda_t}\right)} \text{erfc}\Big(\frac{1}{\sqrt{\pi\gamma_0}}\left(\frac{\lambda_r}{\lambda_t}\right)\Big)\Bigg). \tag{3.45}$$

As one would expect, the closest and random node strategies outperforms the GT strategy and the non-adaptive strategies by a margin. Instead of plotting the curves of all strategies here, here we consider the random node in range strategy only. Our

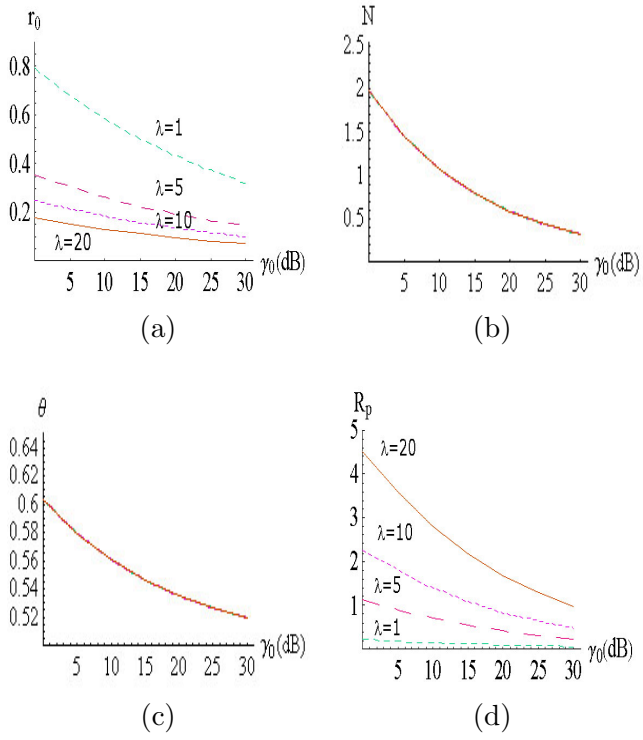



Figure 3.10: Optimized random node within range strategy at different node density $\lambda$ in a practical system with SIR threshold $\gamma_0$. (a) transmit range $r_0$ vs. $\gamma_0$, (b) expected number of nodes in range $N$ vs. $\gamma_0$, (c) fraction of candidate transmit nodes $\theta$ vs. $\gamma_0$, (d) expected packet success rate per unit area vs. $\gamma_0$.

focus is on the sensitivity of these parameters to the SIR threshold $\gamma_0$ determined by the modulation and error control scheme. As shown in Figure 3.10, the optimal range, fraction of candidate transmit nodes, number of neighbors and the packet success rate per unit area are plotted as a function of the SIR threshold $\gamma_0$ for node density $\lambda = 1, 5, 10, 20$ nodes/$m^2$. The invariance of the number of neighbors $N$ and the fraction of candidate transmit nodes $\theta$ to node density also applies to a practical system, readily seen as the curves for different node densities overlap in Figure 3.10(b) and (c). We also vary the SIR threshold $\gamma_0$ from 0 to 30dB, which is a factor of 1000 in linear scale. We observe that the optimal fraction of candidate transmit nodes is insensitive to the variation of $\gamma_0$, hovering between 0.5 and 0.6 and decreases gradually as the SIR threshold is raised. As discussed in the previous section, a value of $\theta$ close to 0.5 ensures a nice mix of transmit and receive nodes that give rise to numerous excellent

channels. The slight decrease of $\theta$ at high SIR threshold indicates an optimally operated system, nodes are more conservative and show less willingness to transmit. This ensures the packet success rate is not adversely affected at the expense of some loss of spatial concurrency.

Similarly, the optimal transmit range also shrinks when the SIR target is raised to ensure that only the good channels capable of meeting the SIR target are used. The decrease in spatial concurrency is inconsequential, since the signal strength of any connection outside the transmit range may be too weak to meet the SIR target and any received packet from a node outside the range may be discarded in any case. The optimal number of neighbors $N$ is weakly dependent on the SIR threshold, ranging from $N \approx 2$ at low SIR threshold $\gamma_0 = 0$dB and $N \approx 0.5$ when the SIR threshold is raised to 30dB. This should be contrasted to the optimal number of neighbors of around 1 in a theoretical system. Although there is a tradeoff of spatial concurrency for more spectral efficiency at high SIR threshold, the deviation of optimal $N$ in practical systems with very different SIR requirements is small.

In Figure 3.10(d), the packet success rate per unit area is also decreasing as the SIR threshold increases. However, this comparison is far from fair. When the SIR threshold is high, it usually alludes that the modulation scheme has a very high spectral efficiency. If modulations of high order constellation size are used, multiple bits can be received per symbol, which in turns raise the bit rate. A more general metric must be considered to allow for a more meaningful comparison of modulation schemes with different spectral efficiency. On the other hand, it is also interesting to know if the packet success rate per unit area $R_p$ is linearly increasing with node density. To see this we plot $R_p$ versus node density in Figure 3.9(b). It is obvious that the linear dependence of $R_p$ to node density also applies to practical schemes for all values of $\gamma_0$. This is expected since our rescaling argument from percolation theory is only dependent on the topology. The underlying communication model is irrelevant to the application of the rescaling argument.

## 3.6 Discussion

We have examined four transmission strategies in this chapter, and showed that adaptive strategies with a stipulated transmit range perform substantially better than the GT strategy with an unconstrained transmit range. Our results imply there is a tradeoff between the spatial transmission concurrency and the spectral efficiency of each transmission. In order to maximize the capacity per unit area, it is necessary to limit the number of simultaneous transmissions to reduce the network interference power such that the SIR and the spectral efficiency of other connections are improved. Moreover, the random node within range strategy has a performance that is close to the closest node in range strategy. Thus a designer of multiple access protocols only needs to focus on contention of local channel when several receive nodes are in proximity. There is no need of a scheduling algorithm for prioritized transmissions based on distance or received power.

The results shown in Table 3.1 show that the optimum range of our strategies is between 0.6 to 1.2 neighbors independent of node density. These results can be contrasted to the results in [28, 44, 91], which suggested that a magic number of 6 to 8 neighbors or a scaled version of it to account for the processing gain in spread spectrum systems [87] and second order effects of the channel [107] is optimum. In these works, a hypothetical line is drawn from a source to the destination node. The transmit range is chosen such that the the expected distance advance in one transmission projected to this line is maximized. This performance metric is called the *forward progress* in the literature. The concept of forward progress is predicated on the assumption that mobile nodes communicate using multihop routing. What we have shown in this chapter suggests that capacity per unit area of one network snapshot can be fully utilized only if each transmit node sees one neighbor node on the average. Our results demonstrate that the mobile infostation network is a paradigm that fits into this optimization criterion.

To appreciate the potential improvement in *link capacity* over the multihop paradigm, we plot the expected SIR $\gamma(N)$ at the transmit range boundary as a function of number of neighbors (3.33) of a node in Figure 3.11. It is interesting to note that as the number

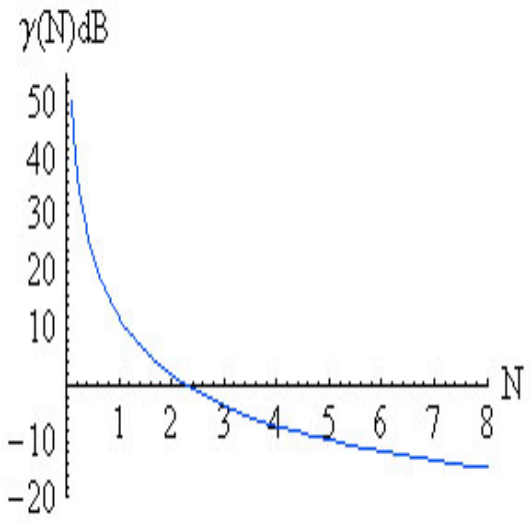



Figure 3.11: Illustration of SIR $\gamma$ as a function of number of neighbors $N$.

of neighbors $N$ increase from 1 to 8, the SIR at the range boundary drops from 15dB to -15dB, a factor of 1000. Substitute the value of the SIR to the capacity equation, we see that the link capacity of a mobile infostation connection is 111.93 times over a multihop forwarding connection. The dramatic improvement in link capacity, together with [20] which explicitly show that the sum capacity in each network snapshot is sustainable in the long run, convince us that a much larger end-to-end throughput capacity is realizable for mobile infostation networks.

Recall that [20] showed the mobile infostation paradigm allows a network throughput that is scalable to the number of nodes. We have obtained exact capacity per unit area expressions as a function of transmit range, the fraction of candidate transmit nodes and node density. It turns out that the mobile infostation paradigm not only improves the spectral efficiency of a link over the multihop paradigm. It is somewhat surprising to find out that the spectral efficiency per unit area is linearly increasing with node density in mobile infostation networks. This is counter-intuitive since an increase in the node density is often accompanied by a corresponding increase of network interference. However, a mobile infostation also shrinks the transmit range such that the number of nodes within the transmit range remains constant. Thus, a mobile infostation also exploits the increase in physical proximity of the receive nodes as node density increases. The contrasting effects of increasing signal strength and increasing interference power at

high node density work together that brings to the independence of link SIR's to node density. At high node density, the same sum capacity can be achieved at a smaller area, leading to an increase in capacity per unit area. This result has far reaching implications for the feasibility of future pervasive computing environments. The proliferation of mobile devices makes the deployment of dense node networks in the future almost a certainty. Unfortunately multihop networks suffers from the curse of node density. The excessive need of multihop forwarding in high node density environments drives the achievable per-node throughput to zero. In contrast, node density is a blessing in mobile infostation networks. The increase in interference power due to increased node density is counter-balanced by the improved channel due to the proximity of receive nodes at high node density. Since nodes are packed closer in high node density scenarios, better spatial concurrency is achieved, leading to an increase in capacity per unit area. Our results show that the capacity per unit area for mobile infostations actually goes to infinity as node density increases.

We have also examined the optimal transmit range and fraction of candidate transmit nodes in practical systems where the modulation scheme and error correction scheme is prespecified. A simple abstraction was made for practical systems using the notion of SIR threshold $\gamma_0$. When good long error correction codes are used, the packet success probability as a function of SIR $\gamma$ is well approximated by a unit step function at the transition $\gamma_0$. It is interesting to note that when the SIR threshold $\gamma_0$ is high, the probability of packet success decreases. However, a high $\gamma_0$ also implies that spectrally efficient modulation schemes can be used. A tradeoff therefore exists between spectral efficiency and packet success rate. A low SIR threshold raises the packet success rate at the expense of lower spectral efficiency and vice versa. Suppose a particular modulation scheme is use in a practical system. An optimal constellation size therefore exists that optimally tradeoff the spectral efficiency and packet success rate to maximize the bit rate of a link. In general, spectral efficiency, spatial concurrency and packet success rate are inter-related. It will be interesting to determine a constellation size and the transmit range such that the throughput capacity per unit area is optimized. However, due to the size limit of this chapter, the issue of optimal constellation

size for different modulation and error correcting schemes are not included.

In retrospect, we have looked into the optimal transmit range and the theoretical and practical achievable rate per unit area of mobile infostation networks. The concept of transmit range is novel in the paradigm of mobile infostations. Capacity equations are derived for four strategies and we show that a stipulated transmit range improves capacity. Though it is not obvious in the problem formulation, the optimal number of neighbors of a node, and the fraction of nodes as candidate transmit nodes is invariant to node density. Comparisons have been made to the well known magic number of 6 to 8 neighbors, reflecting the contrasting optimization criteria for the multihop networking and mobile infostations paradigm. Another finding is that the capacity per unit area is linearly increasing with node density. This can be explained by a rescaling argument drawn from percolation theory. This has implications in the design of ad hoc networks in future pervasive networking environments with high node density. We also extend our results to practical systems characterized by a SIR threshold. The invariance of $N$ and $\theta$ to node density continues to hold and the packet success rate per unit area is linearly increasing with node density. When a system has high SIR requirements, $N$ and $\theta$ decreases slightly. In particular, when the SIR requirement increases by a factor of 1000 from 0 to 30dB, the optimal number of neighbors slightly decreases from 2 to 0.5 only.

# Chapter 4

# Effect of Node Mobility on Highway Mobile Infostation Networks

## 4.1 Introduction

In this chapter, we examine the effect of node mobility in mobile infostation networks. In [20], mobility provides a mechanism such that numerous instances of excellent channels between different nodes can be exploited. The realization of large network capacity comes from the translation of maximal spatial transmission concurrency in each network snapshot to the long run end-to-end network capacity. The physical implication of mobility in node encounters has been glossed over. In reality, the total connection time of a node over a specific interval depends on the node encounter rate and the connection time in each encounter, both of which depend on the relative mobility of nodes. Although a high node speed results in more node encounters, the connection time in each node encounter also decreases. It is not apparent whether high or low speed results in a larger connection time, and thus, data rate. To this end we propose a new mobility model for highway networks. The highway scenario proves to be interesting despite its mathematical simplicity. Consider the *forward traffic* scenario, that is traffic travelling at the same direction as the user of interest, or the *observer node*. The connection time in one node encounter is much larger than that of reverse direction traffic, but the node encounter rate is also much smaller. In the *reverse traffic* scenario, on the other hand, the connection time in a node encounter is typically small, since nodes are travelling in opposite directions. Nevertheless, node encounter rate is also much higher in the reverse traffic scenario. It is not immediately apparent which traffic type offers the greater fraction of connection time, or number of connections in queueing terminology. Second, the connection time in an encounter depends on the transmit

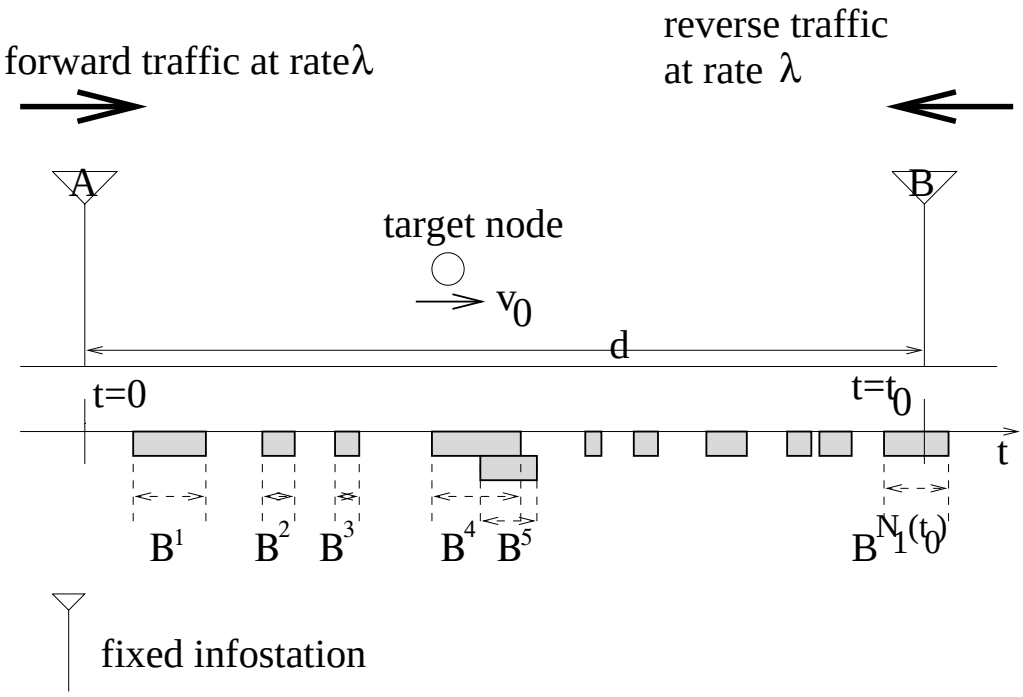



Figure 4.1: Illustration of the highway mobile infostation network model.

range of the nodes. For both forward and reverse traffic, an optimal transmit range exists such that the long run data rate of a node is maximized.

The rest of the chapter is organized as follows. In section 4.2, we describe the system model. Section 4.3 is devoted to performance analysis for arbitrary speed distribution. The special case of uniform speed distribution is considered in section 4.4 and numerical results are obtained in section 4.5. Finally, we discuss the implications of our results in section 4.6.

## 4.2 System Model

We consider a highway network in which fixed infostations are placed regularly at a distance $d$ from each other. We assume that all nodes are subscribers of a content provider, say a movie distribution network. Movies are split into many files and are cached in the infostations at various locations. Besides downloading directly from an infostation, a node participates in data exchanges whenever there is another node in proximity. We assume data exchanges between two proximate nodes always take place without further negotiation. The amount of data exchanged is proportional to the connection time in

an encounter and the data transmission rate. It was shown in [104] that in a large network, peer-to-peer node exchanges account for most of the data transmissions. As the network size increases, the importance of fixed infostations in data dissemination dwindles. Thus, in this chapter we focus on peer-to-peer connections between proximate mobile nodes in node encounters only. Connections to fixed infostations on the highway are ignored.

In our analysis, we focus on an arbitrary highway segment between infostations $A$ and $B$, as shown in Figure 4.1. On each highway segment, a node moves at a speed $V$, an iid random variable drawn from a known but arbitrary distribution $G$. Since nodes have different speeds, a node may overtake other nodes or be overtaken as it traverses the highway segment. We make all our observations at a specific node, called the *observer node*. Two types of traffic are considered here. For *forward traffic*, nodes are injected into the highway segment at a Poisson rate $\lambda$ from infostation $A$. The Poisson arrival assumption of mobile nodes is valid if the speed of individual nodes is independent and does not interact. That is, we assume there is no delay incurred in a node encounter, in which a platoon of nodes forms behind a node that moves slowly. This is plausible in a wide highway with multiple lanes and moderate traffic, where nodes overtake others at different lanes. The injected nodes move at the same direction as the observer node. This is called the wide motorway model in [42]. Similarly, for *reverse traffic* nodes are injected into the highway segment at a Poisson rate $\lambda$ from infostation $B$. The injected nodes move in the opposite direction of the observer node. More generally, a node changes speed as time evolves. We assume each node still moves at a constant speed in a highway segment. Whenever a node traverses a new highway segment, we stipulate that each node selects a new speed from the distribution $G$, independent of the previous speed.

Suppose the observer node moves at a speed $V = v_0$ on a highway segment from infostation $A$ to $B$. We denote the time for the node to traverse a highway segment as the *cycle duration*, given by $T = d/V$, with a corresponding distribution $F$. $F$ and $G$ are obviously related, given by $\overline{F}(t) = G(d/t)$, where $\overline{F}(t) = 1 - F(t)$ denotes the complementary distribution function. In this chapter, we describe mobility of the

observer node in terms of cycle duration rather than node speed for convenience, since the performance metrics are closely related to $t_0$.

Given the observer node cycle duration $t_0 = d/v_0$ in a highway segment, we denote $N_1(t_0)$ and $N_2(t_0)$ as the number of *node encounters* in forward traffic and reverse traffic scenarios, where a node encounter occurs when two nodes are approaching to within a transmit range $r$ from each other, and the subscripts 1 and 2 denotes a connection with forward and reverse traffic respectively. The *connection time* in each node encounter is defined as the duration when both nodes are within the transmit range $r$ from each other. Obviously, the connection times in forward traffic $Y_1^i(t_0)$ and reverse traffic $Y_2^i(t_0)$ at the $i$-th node encounter are random variables dependent on the relative speed of the nodes and the common transmit range of all nodes $r$. For many speed distributions, two nodes having a similar speed may have a connection time with unbounded mean. However, each node only has a finite amount of data for dissemination to another node. To model this we specify a *connection time limit* parameter $c$ to limit the actual connection time in a node encounter, given by $B_1^i(t_0) = \min(Y_1^i(t_0), c)$ and $B_2^i(t_0) = \min(Y_2^i(t_0), c)$. We also denote the total connection time of the observer node in a highway segment as $Z_1(t_0)$ and $Z_2(t_0)$. Obviously,

$$Z_1(t_0) = \sum_{i=1}^{N_1(t_0)} B_1^i(t_0) \tag{4.1}$$

$$Z_2(t_0) = \sum_{i=1}^{N_2(t_0)} B_2^i(t_0) \tag{4.2}$$

When speed changes are incorporated to our mobility model, the long run average fraction of connection time and data rate are the appropriate metrics. It turns out that a simple characterization of these metrics is possible by drawing results from renewal reward theory [77]. Let $M(t), t \geq 0$ be a counting process to denote the number of highway segments traversed by the observer node. At the $n$th highway segment, the observer node selects an iid random speed $V_n$ independent of the speed $V_{n-1}$ at the previous highway segment $n-1$. The corresponding cycle durations $T_n$ are iid random variables. Since $M(t)$ is a counting process with iid interarrival times, $M(t)$ is a renewal process. Moreover, we denote $R_n$ as the reward earned in the $n$th cycle, or renewal

period. If we let

$$R(t) = \sum_{n=1}^{M(t)} R_n, \tag{4.3}$$

then $R(t)$ is the total reward earned by time $t$. Let $E[R] = E[R_n]$ and $E[T] = E[T_n]$, the renewal reward theorem [77] states that if $E[R] < \infty$ and $E[T] < \infty$, then with probability 1,

$$\lim_{t\to\infty} \frac{R(t)}{t} = \frac{E[R]}{E[T]} \tag{4.4}$$

That is, the rate of earning reward in the long run is just the ratio of the expected reward in a cycle and the expected cycle duration.

Accordingly, if we define a reward of 1 unit is earned every time the observer node encounters another node, the reward accrued in highway segment $n$ is $R_n = N_1(T_n)$ for forward traffic and $R_n = N_2(T_n)$ for reverse traffic. The long run node encounter rate of the observer node is simply

$$\mathcal{N}_1 = \lim_{t\to\infty} \frac{R(t)}{t} = \frac{E[N_1(T)]}{E[T]}. \tag{4.5}$$

in forward traffic scenario and

$$\mathcal{N}_2 = \lim_{t\to\infty} \frac{R(t)}{t} = \frac{E[N_2(T)]}{E[T]}. \tag{4.6}$$

in reverse traffic scenario. Similarly, suppose a reward equivalent to the connection time $B_1(t_0)$ is earned each time the observer node encounters another node. Let the observer node mobility at the $n$-th highway segment be $T_n = t_0$. The accrued reward $R_n$ is the sum of the connection times of all node encounters in the highway segment, i.e.

$$R_n = Z_1(t_0) = \sum_{i=1}^{N_1(t_0)} B_1^i(t_0). \tag{4.7}$$

in forward traffic scenarios. In this case, the long run rate of earning reward is given by

$$\mathcal{Z}_1 = \lim_{t\to\infty} \frac{R(t)}{t} = \frac{E[Z_1(T)]}{E[T]}. \tag{4.8}$$

Similarly in reverse traffic scenarios we have

$$R_n = Z_2(t_0) = \sum_{i=1}^{N_2(t_0)} B_2^i(t_0) \tag{4.9}$$

and

$$\mathcal{Z}_2 = \lim_{t\to\infty} \frac{R(t)}{t} = \frac{E[Z_2(T)]}{E[T]}. \tag{4.10}$$

Finally, suppose a reward equivalent to the amount of data sent and received is earned each time the observer node encounters another node. Assuming non-adaptive radios are used, the data rate is the Shannon rate at the transmit range boundary $r$, given by

$$C(r) = \ln(1 + 1/r^4), \tag{4.11}$$

where we have assumed a path gain exponent of 4 and ignored the effect of mutual interference. Let the cycle duration of the observer node at the $n$-th highway segment be $T_n = t_0$. The accrued reward $R_n$ is the total amount of data transmitted or received by the observer node in the highway segment, denoted as $W_1(t_0)$. The average rate of earning reward in the long run should be interpreted as the long run data rate, given by

$$\mathcal{W}_1 = \frac{E[W_1(T,r)]}{E[T]} = C(r)\frac{E[Z_1(T,r)]}{E[T]} = C(r)\mathcal{Z}_1 \tag{4.12}$$

in forward traffic scenarios and

$$\mathcal{W}_2 = \frac{E[W_2(T,r)]}{E[T]} = C(r)\frac{E[Z_2(T,r)]}{E[T]} = C(r)\mathcal{Z}_2, \tag{4.13}$$

in reverse traffic scenarios. We emphasize both connection time $Z$ and the amount of delivered data $W$ are dependent on the transmit range $r$. It is intuitive that $\mathcal{W}_1 = 0$ and $\mathcal{W}_2 = 0$ when the transmit range is either zero or very large. An optimal transmit range $r$ exists for both traffic types such that $\mathcal{W}_1$ and $\mathcal{W}_2$ are maximized respectively.

In this chapter, we consider the special case when each node selects an arbitrary speed upon entrance to the highway. However, each node moves with the *same* speed in different highway segments. Since the cycle duration $T_n$ is still iid, the renewal arguments continues to apply in the constant speed case. Moreover, the long run fraction of connection time $\mathcal{Z}_1$ and $\mathcal{Z}_2$ simplifies to

$$\mathcal{Z}_1 = \eta_1(t_0) = \frac{E[Z_1(t_0)]}{t_0} \tag{4.14}$$

and

$$\mathcal{Z}_2 = \eta_2(t_0) = \frac{E[Z_2(t_0)]}{t_0}, \tag{4.15}$$

which can be interpreted as the *expected fraction of connection time* $\eta_1(t_0)$ and $\eta_2(t_0)$ as a function of observer node mobility $t_0$. In general, since a node can simultaneously maintain more than one connection, $\eta_1(t_0)$ and $\eta_2(t_0)$ can be larger than 1. In queueing terminology, the observer node is a server and the connection time in a node encounter corresponds to the service time. Although an optimum transmit range exists in both forward and reverse traffic scenarios, we will not pursue this idea further in this chapter. Bear in mind that when the transmit range is conditionally given, the fraction of connection time $\mathcal{Z}$ is linearly proportional to the long run data rate $\mathcal{W}$.

## 4.3 Performance Analysis

Consider the forward traffic scenario. Suppose the observer node enters the highway segment at time $s$ and departs at time $s + t_0$. We denote an event occurs at time $t \in [0, \infty)$ if a node enters the highway segment at infostation $A$. Since the node travels with random speed $V = d/T$, this node leaves the highway segment at time $t + T$. We define $p_1(t)$ as the probability that a forward entrant at time $t$ has an encounter to the observer node at the highway segment. It is straightforward to show that for $t < s$, an encounter occurs if $t + T > s + t_0$ when the observer node overtakes the encounter node. That is,

$$p_1(t) = P[T + t > s + t_0] = \overline{F}(s + t_0 - t). \tag{4.16}$$

Similarly, for $s < t < s + t_0$, an encounter occurs if $t + T < s + t_0$ when the observer node is overtaken by the encounter node. This occurs with probability

$$p_1(t) = P[T + t < s + t_0] = F(s + t_0 - t). \tag{4.17}$$

Finally, for $t > s + t_0$, a node encounter will not occur in the highway segment, i.e. $p_1(t) = 0$. Combining the three cases together, we have

$$p_1(t) = \begin{cases} \overline{F}(s + t_0 - t) & t < s \\ F(s + t_0 - t) & s < t < s + t_0 \\ 0 & t > s + t_0 \end{cases}. \tag{4.18}$$

Assuming the network has been operated for a long time $s \to \infty$ before we observe the observer node enters the highway segment. The total number of node encounters is also

a Poisson process and in steady state $s \to \infty$, it is given by

$$\lim_{s\to\infty} E[N_1(t_0)] = \lim_{s\to\infty} \lambda \int_0^\infty p_1(t)dt \tag{4.19}$$

$$= \lambda \left( \int_0^{t_0} F(t)dt + \int_{t_0}^\infty \overline{F}(t)dt \right). \tag{4.20}$$

It can be shown $E[N_1(t_0)]$ attains a global minimum when the observer node cycle duration $t_0$ is the median of the distribution $F$ By twice differentiating (4.20) [77]. This agrees with our intuition that there are few node encounters if the observer node moves at a speed that goes along with the majority.

For reverse traffic, we define an event occurs at time $t$ if a node enters the highway segment from infostation $B$. For an event at time $t$, it is marked with probability $p_2(t)$ if there is a node encounter with the observer node at the highway segment. For $t > s + t_0$, the reverse entrant node enter the highway segment after the observer node has left, the encounter probability is therefore $p_2(t) = 0$. For $s < t < s + t_0$, the reverse entrant node enters the highway segment after observer node, but before the observer node has left. Thus the encounter probability is $p_2(t) = 1$. Finally, when $t < s$, a node encounter occurs if the reverse entrant node leave after the observer node arrives at the highway segment. This happens with probability

$$p_2(t) = P[T + t > s] = \overline{F}(s - t). \tag{4.21}$$

Combining the three cases, we have

$$p_2(t) = \begin{cases} 0 & t > s + t_0 \\ 1 & s < t < s + t_0 \\ \overline{F}(s - t) & t < s \end{cases}. \tag{4.22}$$

The total number of node encounters in steady state is

$$\lim_{s\to\infty} E[N_2(t_0)] = \lim_{s\to\infty} \lambda \int_0^\infty p_2(t)dt \tag{4.23}$$

$$= \lambda(t_0 + E[T]), \tag{4.24}$$

where $E[T]$ is the expected cycle duration given by

$$E[T] = \int_0^\infty \overline{F}(t)dt. \tag{4.25}$$

The long run node encounter rate for both traffic types can be obtained by averaging over the speed distribution. Thus we have

$$\begin{aligned} E[N_1(T)] &= \int_0^\infty E[N_1(t_0)]\, dF(t_0) && (4.26)\\ &= \lambda \int_0^\infty \int_0^{t_0} F(t)dt\, dF(t_0) && (4.27)\\ &+ \lambda \int_0^\infty \int_{t_0}^\infty \overline{F}(t)dt\, dF(t_0), && (4.28) \end{aligned}$$

which yields

$$E[N_1(T)] = 2\lambda \int_0^\infty \overline{F}(t)F(t)dt \qquad (4.29)$$

upon simplification using integration by parts. Similarly, we have

$$\begin{aligned} E[N_2(T)] &= \int_0^\infty E[N_2(t_0)]\, dF(t_0) && (4.30)\\ &= 2\lambda E[T]. && (4.31) \end{aligned}$$

(4.29) and (4.31) suggest that the expected node encounter rate for reverse traffic is always larger than that for forward traffic, which is obviously true. Moreover, (4.31) shows that the expected node encounter rate is completely characterized by the traffic intensity $\lambda$ and the first moment of distribution $F$.

To compute the expected connection time in one encounter for forward traffic $E[B_1(t_0)]$, we note that

$$\begin{aligned} E[B_1(t_0)] &= \int_0^c P[Y_1(t_0) > t]dt && (4.32)\\ &= \int_0^c P\left[\frac{2r}{|v_0 - V|} > t\right] dt && (4.33)\\ &= \int_0^c G\left(\frac{2r}{t} + \frac{d}{t_0}\right) - G\left(\frac{d}{t_0} - \frac{2r}{t}\right) dt. && (4.34) \end{aligned}$$

Similarly, in reverse traffic we have

$$\begin{aligned} E[B_2(t_0)] &= \int_0^c P[Y_2(t_0) > t]dt && (4.35)\\ &= \int_0^c G\left(\frac{2r}{t} - \frac{d}{t_0}\right) dt. && (4.36) \end{aligned}$$

Refer to Figure 4.1 again, the total connection time for forward traffic is obtained by summing all individual connection time $B_1^i(t_0)$, $i \in [1, N_1(t_0)]$ over the cycle. In

the event that the connection time of the encounter $N_1(t_0)$ overshoots the end of the cycle, the observer node undergoes a renewal and selects a new speed. This in turn modifies the connection time $B_1^{N_1(t_0)}$. Nevertheless, the boundary effect of an overshoot connection time is minimal when either $N_1(t_0)$ is large, or when $B_1(t_0) \leq c \ll t_0 = d/v_0$. The former assumption is valid when the traffic intensity $\lambda$ is moderate, such that $N_1(t_0) \gg 1$. The latter assumption is valid when the distance between fixed infostations $d$ is large, which is likely in an initial deployment of a fixed infostation network. Ignoring the boundary effect of $B_1^{N_1(t_0)}(t_0)$, we have

$$Z_1(t_0) \quad = \quad \sum_{i=1}^{N_1(t_0)} B_1^i(t_0). \tag{4.37}$$

It can be shown that $B_1^i(t_0)$ are iid random variables and $N(t_0)$ is Poisson. However, $N_1(t_0)$ and $B_1^i(t_0)$ are in general not independent. In fact, when node mobility is high, $N_1(t_0)$ is large and the corresponding $B_1(t_0)$ is small. Thus $Z_1(t_0)$ is not a compound Poisson process. Nevertheless, we note that $N_1(t_0)$ is a stopping time w.r.t. the sequence $B_1^i(t_0)$ since the stopping rule $\{N_1(t_0) = n\}$ is completely determined by the information up to time $n$, and is unrelated to $B_1^{n+1}(t_0)$, $B_1^{n+2}(t_0)$ and so on. Thus, Wald's equality [15] can be applied to (4.37) to yield

$$E[Z_1(t_0)] \quad = \quad E[N_1(t_0)]E[B_1(t_0)]. \tag{4.38}$$

Similarly, in reverse traffic we have

$$E[Z_2(t_0)] \quad = \quad E[N_2(t_0)]E[B_2(t_0)]. \tag{4.39}$$

The long run fraction of connection time, or number of connections of the observer node for both traffic types can be obtained by conditioning on distribution $F$, given by,

$$\mathcal{Z}_1 \quad = \quad \frac{E[Z_1(T)]}{E[T]} = \frac{\int_0^\infty E[Z_1(t_0)]dF(t_0)}{E[T]} \tag{4.40}$$

and

$$\mathcal{Z}_2 \quad = \quad \frac{E[Z_2(T)]}{E[T]} = \frac{\int_0^\infty E[Z_2(t_0)]dF(t_0)}{E[T]}. \tag{4.41}$$

Given the transmit range $r$, $\mathcal{Z}_1$ and $\mathcal{Z}_2$ are linearly related to the long run average data rate $\mathcal{W}_1$ and $\mathcal{W}_2$. Since we do not focus on finding an optimal transmit range that

maximizes the long run average data rate in this chapter, we do not discuss $\mathcal{W}_1$ and $\mathcal{W}_2$ further.

## 4.4 Uniform Speed Distribution

We consider the case when node speed is uniformly distributed according to (4.42), given by

$$G(v) = \begin{cases} 0 & 0 \le v \le v_a \\ \frac{v-v_a}{v_b-v_a} & v_a \le v \le v_b. \\ 1 & v \ge v_b \end{cases} \tag{4.42}$$

The corresponding distribution of the cycle duration $T = d/V$ is

$$F(t) = \begin{cases} 0 & 0 \le t \le d/v_b \\ \frac{v_b-d/t}{v_b-v_a} & d/v_b \le t \le d/v_a. \\ 1 & t \ge d/v_a \end{cases} \tag{4.43}$$

Our objective here is twofold. First, we consider the case when each node selects its speed from distribution $F$ and then moves at constant speed at all highway segments. We will examine the effect of observer node mobility $t_0$ on its fraction of connection time, or expected number of connections for both forward and reverse traffic scenarios. Second, we incorporate the extended mobility model, where a node selects a new speed at each highway segment. We will examine the long run average number of connections and data rate of a random node in both forward and reverse traffic scenarios.

Substituting (4.43) into (4.20), (4.29), (4.34), $E[N_1(t_0)]$, $E[N_1(T)]$ and $E[B_1(t_0)]$ can be readily computed as

$$E[N_1(t_0)] = \frac{\lambda}{v_b - v_a}\left((v_a + v_b)t_0 + d\ln\frac{d^2}{t_0^2 e^2 v_a v_b}\right) \tag{4.44}$$

$$E[N_1(T)] = \frac{2d\lambda}{(v_b - v_a)^2}\left((v_a + v_b)\ln\frac{v_b}{v_a} - 2(v_b - v_a)\right), \tag{4.45}$$

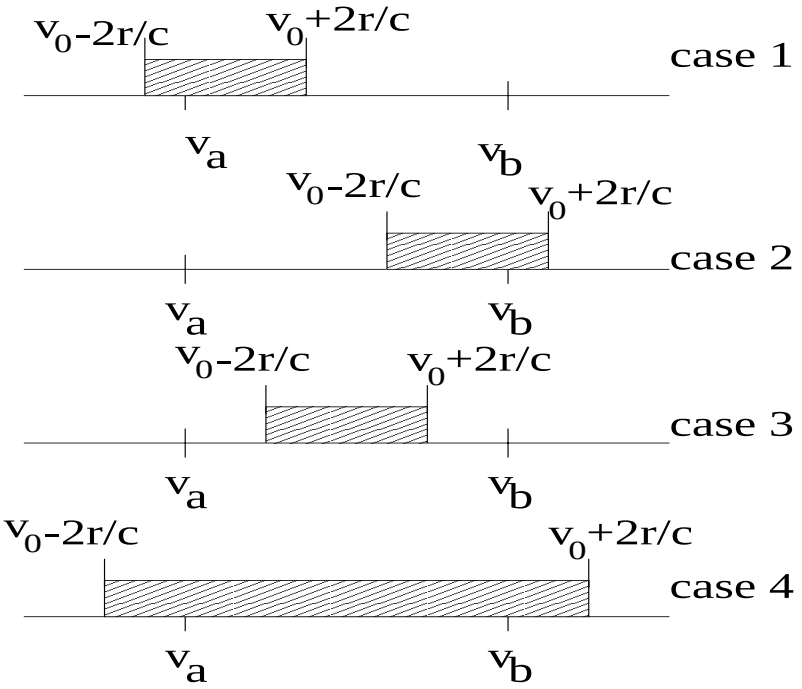



Figure 4.2: In forward traffic, connection time is truncated when the difference of encounter node speed $V$ and observer node speed $v_0$ is less than $2r/c$, i.e. $|V - v_0| \leq 2r/c$. The shaded area shows the range of encounter node speed when connection time truncation occurs.

and $E[B_1(t_0)] =$

$$\begin{cases} \frac{c(\frac{d}{t_0}-v_a)+2r\ln[(v_b-\frac{d}{t_0})(\frac{ce}{2r})]}{v_b-v_a} & t_0 \geq \max(\frac{d}{v_a+\frac{2r}{c}}, \frac{d}{v_b-\frac{2r}{c}}) \\ \frac{c(v_b-\frac{d}{t_0})+2r\ln[(\frac{d}{t_0}-v_a)(\frac{ce}{2r})]}{v_b-v_a} & t_0 \leq \min(\frac{d}{v_a+\frac{2r}{c}}, \frac{d}{v_b-\frac{2r}{c}}) \\ \frac{2r\ln[(\frac{ce}{2r})^2(v_b-\frac{d}{t_0})(\frac{d}{t_0}-v_a)]}{v_b-v_a} & \frac{d}{v_b-\frac{2r}{c}} \leq t_0 \leq \frac{d}{v_a+\frac{2r}{c}} \\ c & \frac{d}{v_a+\frac{2r}{c}} \leq t_0 \leq \frac{d}{v_b-\frac{2r}{c}} \end{cases} . \quad (4.46)$$

The derivation of the expected connection time in one node encounter $E[B_1(t_0)]$ is included in Appendix 4.6. For forward traffic, given the speed of the observer node $v_0$ and encounter node $V$, the connection time is truncated if

$$\frac{2r}{|V - v_0|} \geq c, \quad (4.47)$$

or $|V - v_0| \leq 2r/c$. That is, the connection time of forward traffic scenarios is truncated when the relative speed of the encounter node and observer node speed is less than $2r/c$. When the encounter node speed $V$ falls into the shaded area as illustrated in Figure 4.2, the connection time is truncated. The four cases on the figure correspond to the four

cases in (4.46). Cases 1 and 2 correspond to *boundary truncation*. When the observer node has a speed $v_0 \leq v_a + 2r/c$ and $v_0 \geq v_b - 2r/c$ respectively, connection time is truncated when the encounter node speed is at the boundary. Case 3 corresponds to *partial truncation*. Connection time truncation occurs if the difference of encounter node and observer node speed is less than $2r/c$. For large $r/c$, the shaded area is wide and spans over the interval $[v_a, v_b]$. A connection time truncation occurs irrespective of the encounter node speed. This corresponds to case 4 of *full truncation*. The occurrence of each case is dependent on the ratio $r/c$ and the span of the speed distribution $v_b - v_a$.

When $v_b - v_a$ is much larger than $r/c$ such that $v_b - v_a \geq 4r/c$, a observer node may experience left boundary, right boundary and partial connection time truncations depending on its mobility $t_0$. This is usually the case in highway traffic scenarios, where vehicle speed at the fast lane is much larger than that at the slow lane. When $r/c$ is larger, connections are more prone to truncations. In the case $4r/c \geq v_b - v_a \geq 2r/c$, a observer node may experience left boundary, right boundary and full truncation depending on its mobility $t_0$. That is, there exists some observer node mobility $t_0$ such that connection time is always truncated for all encounter node speed. In a typical mobile infostation network, the transmit range is small such that the ratio $r/c$ is much smaller compared with $v_b - v_a$. This case may be applicable when highway traffic is slow due to congestion. When $2r/c \geq v_b - v_a$, the transmit range is so large such that a truncation always occurs regardless of the speeds of the encounter and observer node. The three regimes are summarized in Table 4.1 which shows the range of observer node mobility such that a particular case applies. The connection times for the limiting cases at maximum and minimum observer node speed are also included.

Recall that $E[N_1(t_0)]$ is minimized when $t_0$ is the median of $F$, i.e. $F(t_0) = 1/2$. For uniform distribution, the median is equal to the arithmetic mean. It can be easily verified that $E[N_1(t_0)]$ is convex with a minimum at $t_0 = 2d/(v_a + v_b)$, i.e., when the observer node is at mean speed $v_0 = (v_a + v_b)/2$. Similarly the node encounter rate $E[N_1(t_0)]/t_0$ is also convex with a minimum at $t_0 = d/\sqrt{v_a v_b} \geq 2d/(v_a + v_b)$, where the inequality follows from the fact that arithmetic mean is greater than or equal to the geometric mean. On the other hand, $E[B_1(t_0)]$ is concave with a maximum at

| Regime | $E[B_1(t_0)]$ | Target Node Mobility $t_0$ | $E[B_1(d/v_a)]$ | $E[B_1(d/v_b)]$ |
|---|---|---|---|---|
| $v_b - v_a \geq 4r/c$ | case 1<br>case 2<br>case 3 | $d/v_a \geq t_0 \geq d/(v_a + 2r/c)$<br>$d/v_b \leq t_0 \leq d/(v_b - 2r/c)$<br>$\frac{d}{(v_b-2r/c)} \leq t_0 \leq \frac{d}{(v_a+2r/c)}$ | $\frac{2r}{v_b-v_a} \ln \frac{(v_b-v_a)ce}{2r}$ | $\frac{2r}{v_b-v_a} \ln \frac{(v_b-v_a)ce}{2r}$ |
| $4r/c \geq v_b - v_a \geq 2r/c$ | case 1<br>case 2<br>case 4 | $d/v_a \geq t_0 \geq d/(v_b - 2r/c)$<br>$d/v_b \leq t_0 \leq d/(v_a + 2r/c)$<br>$\frac{d}{(v_a+2r/c)} \leq t_0 \leq \frac{d}{(v_b-2r/c)}$ | $\frac{2r}{v_b-v_a} \ln \frac{(v_b-v_a)ce}{2r}$ | $\frac{2r}{v_b-v_a} \ln \frac{(v_b-v_a)ce}{2r}$ |
| $2r/c \geq v_b - v_a$ | case 4 | $d/v_b \leq t_0 \leq d/v_a$ | c | c |

Table 4.1: Existence of three regimes for forward traffic scenario.

$t_0 = 2d/(v_a + v_b)$. Moreover, the expected connection time as a function of speed is symmetric about the mean speed. That is, the expected connection time is the same when the observer node has a speed of $v_0$ or $v_b + v_a - v_0$. This also explains why the connection times at maximum and minimum observer node speed are equal in Table 4.1.

In the reverse traffic scenario, we substitute (4.43) into (4.24), (4.31), (4.36) to obtain

$$E[N_2(t_0)] = \lambda\Big(t_0 + \frac{d}{v_b - v_a} \ln \frac{v_b}{v_a}\Big) \tag{4.48}$$

$$E[N_2(T)] = \frac{2\lambda d}{v_b - v_a} \ln \frac{v_b}{v_a} \tag{4.49}$$

and $E[B_2(t_0)] =$

$$\begin{cases} \frac{2r \ln\left(\frac{v_b + d/t_0}{v_a + d/t_0}\right)}{v_b - v_a} & t_0 \leq \frac{d}{\max(v_a, 2r/c - v_a)} \\ \frac{2r \ln\left(\frac{d/t_0 + v_b}{2r} ce\right) - c(d/t_0 + v_a)}{v_b - v_a} & \frac{d}{\min(v_b, 2r/c - v_a)} \leq t_0 \leq \frac{d}{\max(v_a, 2r/c - v_b)} \\ c & t_0 \geq \frac{d}{\max(v_b, 2r/c - v_b)} \end{cases} . \tag{4.50}$$

The derivation of the expected time in one node encounter $E[B_2(t_0)]$ is included in Appendix 4.6. Let $V$ and $v_0$ be the speed of the encounter node and observer node respectively. A connection is truncated if

$$\frac{2r}{V + v_0} \geq c, \quad \text{or} \quad V \leq \frac{2r}{c} - v_0. \tag{4.51}$$

That is, given the observer node speed $v_0$, a connection time truncation occurs if the encounter node speed $V$ is too low. As illustrated in Figure 4.3, if the encounter

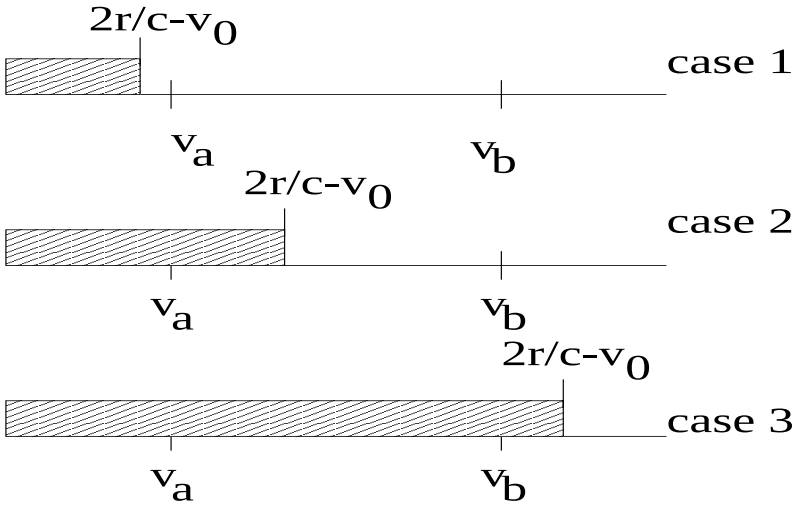



Figure 4.3: In reverse traffic, connection time is truncated when the encounter node speed is smaller than $2r/c - v_0$. The shaded area shows the range of encounter node speed when connection time truncation occurs.

| Regime | $E[B_2(t_0)]$ | Target Node Mobility $t_0$ | $E[B_2(d/v_a)]$ | $E[B_1(d/v_b)]$ |
|---|---|---|---|---|
| $r/c \leq v_a$ | case 1 | $d/v_b \leq t_0 \leq d/v_a$ | $\frac{2r}{v_b - v_a} \ln \frac{v_a + v_b}{2v_a}$ | $\frac{2r}{v_b - v_a} \ln \frac{2v_b}{v_a + v_b}$ |
| $v_a \leq r/c$<br>$v_a + v_b \geq 2r/c$ | case 1<br>case 2 | $d/v_b \leq t_0 \leq d/(2r/c - v_a)$<br>$d/(2r/c - v_a) \leq t_0 \leq d/v_a$ | $\frac{2r \ln \frac{(v_a+v_b)ce}{2r} - 2cv_a}{v_b - v_a}$ | $\frac{2r}{v_b - v_a} \ln \frac{2v_b}{v_a + v_b}$ |
| $v_a \leq r/c$<br>$v_a + v_b \leq 2r/c$ | case 2<br>case 3 | $d/v_b \leq t_0 \leq d/(2r/c - v_b)$<br>$d/(2r/c - v_b) \leq t_0 \leq d/v_a$ | c | $\frac{2r \ln \frac{v_b ce}{r} - c(v_a + v_b)}{v_b - v_a}$ |
| $v_b \leq r/c$ | case 3 | $d/v_b \leq t_0 \leq d/v_a$ | c | c |

Table 4.2: Existence of four regimes for reverse traffic scenario.

node speed falls into the shaded area, the connection time is truncated. The three depicted cases correspond to a connection with no truncation, partial truncation and full truncation. The expected connection time in one node encounter of the three cases is shown in (4.50). In case 1, the shaded area is below $v_a$. Thus there is no connection truncation at all encounter node speed. In case 2, there is partial truncation. Connection time is truncated if the encounter node speed is smaller than $2r/c - v_0$ and vice versa. In case 3, a connection time truncation occurs irrespective of the encounter node speed. This is denoted as full truncation. The occurrence of the three cases depends on speed $v_a$ and $v_b$. Four regimes can be identified and are summarized in Table 4.2.

In the first regime, the minimum speed $v_a$ is large such that $v_a \geq r/c$. Suppose the encounter node moves at speed $V$ and the observer node moves at speed $v_0$, the corresponding connection time is

$$\frac{2r}{V+v_0} \leq \frac{2r}{2v_a} \leq c. \tag{4.52}$$

Thus, there is no connection time truncation at all observer and encounter node speeds. In a highway environment, the minimum node speed $v_a$ is typically much larger than $r/c$. Thus we expect there is no connection time truncation in reverse traffic scenarios. In the second regime, $v_a \leq r/c$ and $v_a + v_b \geq 2r/c$. When there is traffic congestion on the highway, it is possible that the minimum speed is small and satisfies $v_a \leq r/c$. On the other hand, congestion may be local and occurs only in one or two lanes. The fast lanes may experience no congestion such that $v_a + v_b \geq 2r/c$ is satisfied. In this scenario, a observer node undergoes no connection time truncation if it has high mobility such that $d/v_b \leq t_0 \leq d/(2r/c - v_a)$. On the other hand, if the observer node has low mobility such that $d/(2r/c - v_a) \leq t_0 \leq d/v_a$, a connection time truncation occurs when the encounter node speed is smaller than $2r/c - v_0$. In the third regime, $v_a \leq r/c$ and $v_a + v_b \leq 2r/c$. When the maximum speed $v_b$ is also small, a observer node will undergo partial connection time truncation at high mobility when $d/v_b \leq t_0 \leq d/(2r/c - v_b)$. Again, connection time truncation occurs when the encounter node speed is smaller than $2r/c - v_0$. When the observer node has low mobility such that $d/(2r/c - v_b) \leq t_0 \leq d/v_a$, full truncation always occurs irrespective of the encounter node speed. In the fourth regime, the maximum speed $v_b$ is small such that $v_b \leq r/c$. Even if both the observer node and the encounter node move at maximum speed, the corresponding connection time is $2r/2v_b \geq c$. In practice, a mobile infostation network has a small transmit range $r$ and a moderate large connection time limit $c$. It is unlikely that the last two regimes are of importance in reverse traffic scenarios. In the usual highway traffic scenarios, it is reasonable to assume that the first regime holds most of the time. We will therefore perform our numerical experiments for the first regime only.

Although both node encounter rates $E[N_1(t_0)]/t_0$, $E[N_2(t_0)]/t_0$ and connection times $E[B_1(t_0)]$, $E[B_2(t_0)]$ are known analytically, the critical points for $\eta_1(t_0) =$

$E[Z_1(t_0)]/t_0$ and $\eta_2(t_0) = E[Z_2(t_0)]/t_0$ cannot be determined analytically as both involves the products of logarithmic functions in $t_0$. Thus it is impossible to examine the variations of $\eta_1(t_0)$ and $\eta_2(t_0)$ as as function of observer node mobility without employing numerical studies, as we do in the next section. Nevertheless, it is instructive to compare the values of $\eta_1(t_0)$ and $\eta_2(t_0)$ at limiting cases of maximum and minimum observer node speed. Specifically, we compute the ratios $\eta_1(d/v_b)/\eta_1(d/v_a)$ and $\eta_2(d/v_b)/\eta_2(d/v_a)$.

Consider the forward traffic scenario. Recall in Table 4.2 that the connection time at minimum and maximum speed is the same at all regimes by symmetry. The ratio $\eta_1(d/v_b)/\eta_1(d/v_a)$ therefore depends on node encounter rate only. Thus for all the three regimes in the forward traffic scenario, this ratio is given by

$$\frac{\eta_1(d/v_b)}{\eta_1(d/v_a)} = \frac{v_b E[N_1(d/v_b)]}{v_a E[N_2(d/v_a)]} = \frac{\left(\frac{v_b}{v_a}\right)\ln\left(\frac{v_b}{v_a}\right) - \left(\frac{v_b}{v_a}\right) + 1}{\left(\frac{v_b}{v_a}\right) - 1 - \ln\left(\frac{v_b}{v_a}\right)}. \tag{4.53}$$

It is noteworthy that (4.53) is completely determined by the ratio $v_b/v_a$ and is independent of the transmit range $r$ and connection time limit $c$. With reference to Figure 4.4(a), we observe that the ratio $\eta_1(d/v_b)/\eta_1(d/v_a)$ is always larger than 1 for all choices of $v_a$ and $v_b$. In particular, when the difference $v_b - v_a$ is large, say $v_a = 2$ and $v_b = 30$, the ratio is as large as 2.25. That is, the fraction of connection time, or the average number of connections of the observer node is more than double when observer node mobility is high.

In reverse traffic scenarios, we consider the first two regimes, namely $v_a \geq r/c$ and $\{v_a \leq r/c, v_a + v_b \geq 2r/c\}$, since these regimes are most likely to happen in realistic scenarios. Here, the node encounter rate is increasing with node speed and connection time is decreasing with node speed. For $v_a \geq r/c$, we have

$$\frac{\eta_2(d/v_b)}{\eta_2(d/v_a)} = \frac{\left[\left(\frac{v_b}{v_a}\right) + \left(\frac{v_b}{v_a}\right)\ln\left(\frac{v_b}{v_a}\right) - 1\right]\ln\left(\frac{2}{1+\frac{v_b}{v_a}}\right)}{\left[\left(\frac{v_b}{v_a}\right) - 1 + \ln\left(\frac{v_b}{v_a}\right)\right]\ln\left(\frac{1+\frac{v_b}{v_a}}{2}\right)}. \tag{4.54}$$

Although the connection time at $t_0 = d/v_a$ and $t_0 = d/v_b$ is different, it turns out that the ratio $\eta_2(d/v_b)/\eta_1(d/v_a)$ is also independent of transmit range $r$ and dependent on the ratio $v_b/v_a$ only. In this regime, no connections are truncated. The expected

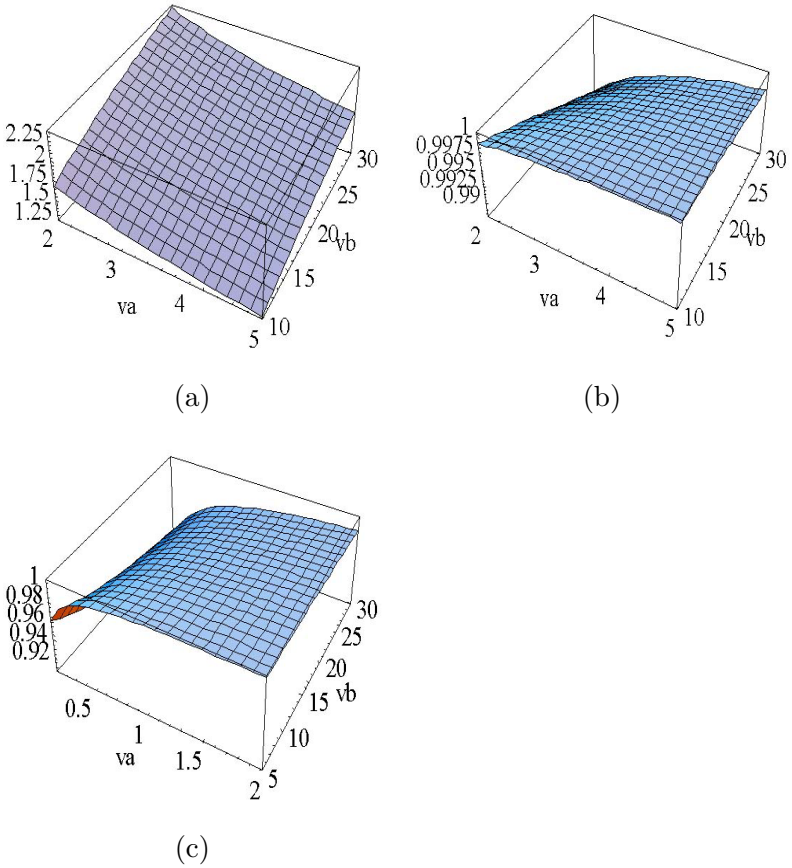
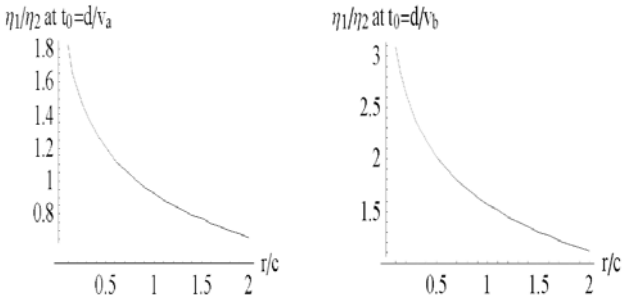



Figure 4.4: (a) Ratio of the average number of connections at maximum speed and minimum speed for forward traffic. (b) Ratio of the average number of connections at maximum speed and minimum speed for reverse traffic ($v_a \geq r/c$). (c) Ratio of the average number of connections at maximum speed and minimum speed for reverse traffic ($v_a \leq r/c$ and $v_a + v_b \geq 2r/c$).

connection times $E[Z_1(t_0)]$ and $E[Z_2(t_0)]$ are linear to the transmit range $r$. Thus $r$ is cancelled out in (4.54). As illustrated in Figure 4.4(b), the ratio $\eta_2(d/v_b)/\eta_2(d/v_a)$ is plotted. It is noteworthy that $\eta_2(d/v_b)/\eta_2(d/v_a) \approx 1$ for a large range of $v_a$ and $v_b$. It naturally leads to a hypothesis that $\eta_2(t_0)$ is independent of node mobility $t_0$, which we have confirmed in our numerical study by plotting out $\eta_2(t_0)$ vs. $t_0$ in the next section.

In the second regime, we have $0 \leq v_a \leq r/c$ and $v_a + v_b \geq 2r/c$. Since a connection undergoes partial truncation when the observer node speed is low, the expected connection time at low observer node mobility is no longer linear to the transmit range $r$.




(a) (b)

Figure 4.5: (a) Ratio of average number of connections of forward traffic to reverse traffic vs. $r/c$ when observer node speed is $v_a$. $(v_a = 2, v_b = 10, d = 1000)$. (b) Ratio of average number of connections of forward traffic to reverse traffic vs. $r/c$ when observer node speed is $v_b$. $(v_a = 2, v_b = 10, d = 1000)$.

In this case, we have

$$\frac{\eta_1(d/v_b)}{\eta_1(d/v_a)} = \frac{\left(1 + \frac{v_b}{v_b - v_a} \ln \frac{v_b}{v_a}\right) \ln \frac{2v_b}{v_a + v_b}}{\left(1 + \frac{v_a}{v_b - v_a} \ln \frac{v_b}{v_a}\right)\left(\ln \frac{ce(v_a + v_b)}{2r} - \frac{cv_a}{r}\right)}, \tag{4.55}$$

which depends on the ratio $r/c$. As illustrated in Figure 4.4(c), the expected fraction of connection time or the number of connections is almost equal for both observer node speed, though it is larger when observer node speed is minimum.

It is also instructive to examine the effect of transmit range to connection time limit ratio $r/c$ on $\eta_1(t_0)/\eta_2(t_0)$. We consider the cases where $t_0 = d/v_a$ and $t_0 = d/v_b$. As reference to Figure 4.5(a),(b), we observe that $\eta_1(t_0)/\eta_2(t_0)$ decreases with $r/c$ in both cases. At high speed, $\eta_1/\eta_2 > 1$ for all values of $r/c$, indicating that forward traffic connections are superior in terms of the fraction of connection time. At low speed, however, forward traffic connections are inferior to reverse traffic connections for large $r/c$. In general, forward traffic connections are more prone to connection time truncation than reverse traffic connections. A large transmit range is not helpful since connection time is truncated in many cases.

## 4.5 Numerical Study

Here, we plot our results numerically to compare the performance of forward and reverse traffic connections at different observer node speed. The parameters $v_a = 2$, $v_b = 10$,

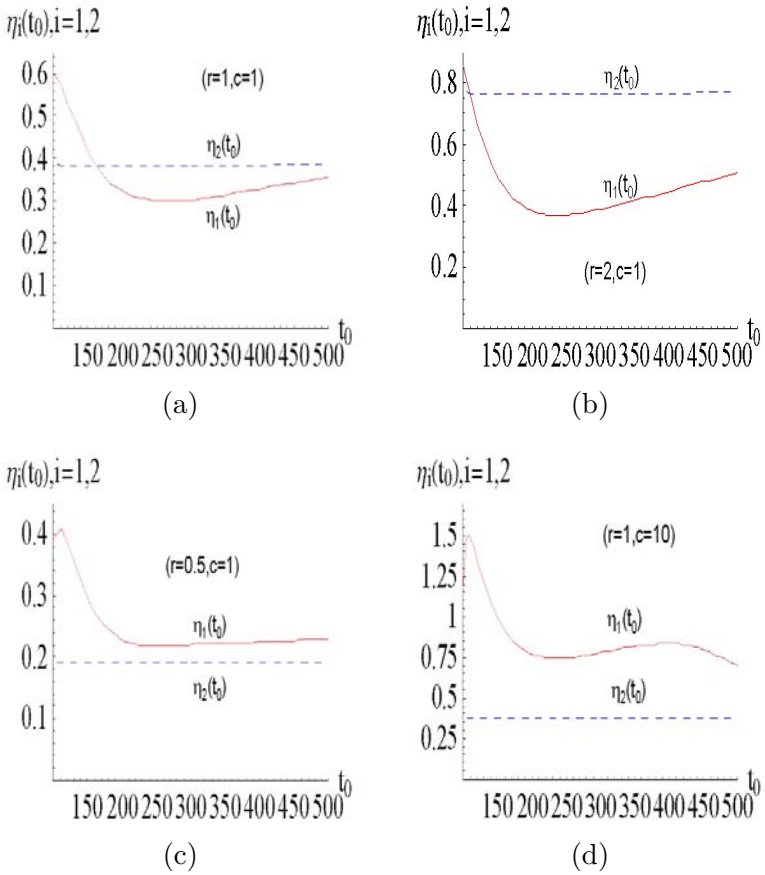



Figure 4.6: Expected number of connections $\eta(t_0)$ versus node mobility $t_0 = d/v_0$ for different transmit range $r$ and connection time limit $c$. (a) $r = 1, c = 1$ (b) $r = 2, c = 1$ (c) $r = 0.5, c = 1$ (d) $r = 1, c = 10$.

$d = 1000$ are adopted. We do not perform simulations, however. Our derivations are exact except for the boundary effect of an overshoot connection time, which is negligible since $c << \min(t_0) = d/\max(v_0) = 100$ by two orders of magnitude. With reference to Figure 4.6, the expected fraction of connection time, or expected number of connections $\eta_1(t_0)$ and $\eta_2(t_0)$ are plotted together versus $t_0$ in the range $d/v_b = 100$ to $d/v_a = 500$. At mean speed $v_0 = 6$, the corresponding $t_0$ is 166.67 unit. Consider scenario 1 for $r = 1, c = 1$. For forward traffic, $\eta_1(t_0)$ attains a global maximum of 0.6 when $t_0$ is minimum. $\eta_1(t_0)$ decreases steadily as $t_0$ increases and hits the minimum of 0.3 at $t_0 = 267.73$. Beyond that, there is a slight increase of $\eta_1(t_0)$ when $t_0$ is increased further. Similar trends are observed for other scenarios in Figure 4.6(b),(c),(d). Nevertheless, a

slight dip of $\eta_1(t_0)$ occurs at low mobility ($t_0 \approx 500$) for Figure 4.6(d). Although there are slightly more encounters at low mobility, there is a steeper decrease in connection time. Thus $\eta_1(t_0)$ is not convex in general. In the particular case of $v_0 = v_a = 0$, the observer node is stationary. The expected fraction of connection time for forward and reverse traffic should be arbitrarily close. That is, the two curves should coincide when $t_0$ is arbitrarily large. In our example, the observer node moves slowly when $v_0 = v_a = 2$. The dip in Figure 4.6(d) is consistent to our intuition that the fraction of connection time for forward and reverse traffic are close when the observer node has low mobility.

In contrast to forward traffic, the expected fraction of connection time, or expected number of connections $\eta_2(t_0)$ is almost constant at all observer node speed in reverse traffic scenarios. The relative value of $\eta_1(t_0)$ and $\eta_2(t_0)$ depends on the ratio of transmit range to connection time limit $r/c$. When $r/c$ is large (Figure 4.6(b)), it is likely that the connection time for forward traffic is truncated. Thus $\eta_1(t_0)$ is consistently smaller than $\eta_2(t_0)$ except for very high observer node speed. When $r/c$ is small (Figure 4.6(c),(d)), the connection time of each node encounter is large. In fact, if there is no connection time limit, the expected connection time for forward traffic is unbounded. The large connection time at large $c$ stipulates that $\eta_1(t_0) > \eta_2(t_0)$ at all node speed. Incidentally, when $r/c = 1$ (Figure 4.6(a)), $\eta_1(t_0)$ and $\eta_2(t_0)$ intersects at $t_0 = 162.7$, which is close to the cycle duration at mean speed $d/E[V] = 166.67$. Thus, if a observer node moves at a constant speed $v_0$ less than the mean speed $E[V]$, reverse traffic connections are more preferable. Similarly, forward traffic connections are more preferable if a node moves at a constant speed $v_0 \geq E[V]$ in this particular example.

Our results show that the data rate of forward traffic connections and reverse traffic connections is dependent on $c$. The value of $c$, in turn, is closely related to the correlation of the contents between two nodes. If nodes have highly correlated contents, any two arbitrary nodes may want to exchange only a few files with each other, effectively modeled by a small $c$. It is more efficient to maintain reverse traffic connections and exchange files with more nodes, as in the case $(r = 2, c = 1)$ shown in Figure 4.6(b). In a content distribution application, this is an appropriate strategy when most nodes get

most of the files already. Similarly, when new content is disseminated, nodes have few files in common and can be modeled by a large $c$. In this case, a node should maintain forward traffic connections to exploit the long expected connection time as warranted by the uniform speed distribution, as in the case $(r = 1, c = 10)$ shown in Figure 4.6(d).

## 4.6 Discussions

In [20], it was shown that mobility increases the capacity of a mobile infostation network. Capacity gain arises from the realization of the maximal spatial transmission concurrency in each network snapshot. Mobility comes into the picture by shuffling node locations, creating numerous instances when excellent channels between different nodes can be exploited (multiuser diversity). As a result of mobility, the sum capacity of each network snapshot translates to the long run end-to-end network throughput. It is noteworthy that in this networking paradigm, end-to-end capacity does not depend on node mobility *per se*. Node mobility, however, do impact the delay performance. The delay of a transiting packet is directly related to the time scale of the mobility process.

In this chapter we have focused on the physical implications of mobility. The fraction of connection time, or number of connections of a observer node over an interval, is determined by the rate of node encounters and the connection time of each encounter, both of which are obviously related to node mobility. It turns out that in reverse traffic scenarios, the expected number of connections is really independent of node mobility. In forward traffic scenarios, however, the expected number of connections (and thus the data rate) increases as mobility increases. Numerical results show that the expected fraction of connection time, or expected number of connections at high node mobility can be much greater than that at low mobility. In particular, in the case when node speed is uniformly distributed between 2 to 30 units, the fraction of connection time is improved by more than a factor of 2 when the observer node increases its speed from minimum to the maximum. Thus, mobility not only provides a mechanism for the exploitation of multiuser diversity. The increase of the fraction of connection time and data rate is a physical consequence of node mobility. Incidentally, this also provides an

incentive for network nodes to be mobile. If a particular mobile user wants to enjoy a higher throughput, or minimizes the downloading time of the files he is interested in, he is motivated to become more mobile and roam around the network. This mobile user in turn helps the network to disseminate data more efficiently, such that the end-to-end delay performance of other users are improved.

It is well known that mobility degrades network performance in many wireless paradigms such as cellular networks and multihop networks. In multihop networks, for instance, extraneous overhead is needed for route maintenance to cope with link failures in node mobility. On the other hand, the fraction of connection time in a fixed infostation model [16] is constant regardless of node mobility. We have shown in this chapter that the fraction of connection time, and data rate increases with node mobility in a mobile infostation network. Thus the mobile infostation network paradigm is superior to multihop networks and fixed infostation networks in its robustness to node mobility.

## Appendix I: Derivation of Expected Connection Time in Forward Traffic

$E[B_1(t_0)] =$

$$\begin{cases} \frac{c(\frac{d}{t_0}-v_a)+2r\ln[(v_b-\frac{d}{t_0})(\frac{ce}{2r})]}{v_b-v_a} & \frac{d}{v_a} \geq t_0 \geq \max(\frac{d}{v_a+\frac{2r}{c}}, \frac{d}{v_b-\frac{2r}{c}}) \\ \frac{c(v_b-\frac{d}{t_0})+2r\ln[(\frac{d}{t_0}-v_a)(\frac{ce}{2r})]}{v_b-v_a} & \frac{d}{v_b} \leq t_0 \leq \min(\frac{d}{v_a+\frac{2r}{c}}, \frac{d}{v_b-\frac{2r}{c}}) \\ \frac{2r\ln[(\frac{ce}{2r})^2(v_b-\frac{d}{t_0})(\frac{d}{t_0}-v_a)]}{v_b-v_a} & \frac{d}{v_b-\frac{2r}{c}} \leq t_0 \leq \frac{d}{v_a+\frac{2r}{c}} \\ c & \frac{d}{v_a+\frac{2r}{c}} \leq t_0 \leq \frac{d}{v_b-\frac{2r}{c}} \end{cases} \tag{4.56}$$

In forward traffic, the connection time is truncated if two nodes move at similar speed. Suppose the observer node speed is $v_0$. Connection time truncation occurs if the encounter node has a speed fall into the shaded area as in Figure 4.2. The four depicted cases corresponds to the connection time given as (4.56).

Proof:

Instead of computing over the time variable $t$, we compute the connection time w.r.t.

the speed random variable. With reference to Figure 4.2, for case 1 we have

$$\begin{cases} v_0 - 2r/c \le v_a & v_0 \le v_a + 2r/c \\ v_0 + 2r/c \le v_b & v_0 \le v_b - 2r/c \end{cases}$$

Combining the two cases, we have

$$t_0 \ge \max(\frac{d}{v_a + 2r/c}, \frac{d}{v_b - 2r/c}).$$

Also, $v_a \le v_0 \le v_b - 2r/c$. Thus this case applies when $v_b - v_a \ge 2r/c$. The corresponding expected connection time in one node encounter is

$$\begin{aligned} E[B_1(t_0)] &= \int_{v_a}^{v_0+2r/c} c\, dG(v) && (4.57) \\ &+ \int_{v_0+2r/c}^{v_b} \frac{2r}{v - v_0} dG(v) && (4.58) \\ &= \frac{c(\frac{d}{t_0} - v_a) + 2r \ln[(v_b - \frac{d}{t_0})\,(\frac{ce}{2r})]}{v_b - v_a}. && (4.59) \end{aligned}$$

For case 2,

$$\begin{cases} v_0 - 2r/c \ge v_a & v_0 \ge v_a + 2r/c \\ v_0 + 2r/c \ge v_b & v_0 \ge v_b - 2r/c \end{cases}$$

Combining the two cases, we have

$$t_0 \le \min(\frac{d}{v_a + 2r/c}, \frac{d}{v_b - 2r/c}).$$

Also, $v_b \ge v_0 \ge v_a + 2r/c$. Thus this case applies when $v_b - v_a \ge 2r/c$. The corresponding expected connection time in one node encounter is

$$\begin{aligned} E[B_1(t_0)] &= \int_{v_a}^{v_0-2r/c} \frac{2r}{v_0 - v} dG(v) && (4.60) \\ &+ \int_{v_0-2r/c}^{v_b} c\, dG(v) && (4.61) \\ &= \frac{c(v_b - \frac{d}{t_0}) + 2r \ln[(\frac{d}{t_0} - v_a)\,(\frac{ce}{2r})]}{v_b - v_a}. && (4.62) \end{aligned}$$

For case 3,

$$\begin{cases} v_0 - 2r/c \ge v_a & v_0 \ge v_a + 2r/c \\ v_0 + 2r/c \le v_b & v_0 \le v_b - 2r/c \end{cases}$$

Combining the two cases, we have

$$\frac{d}{v_b - 2r/c} \le t_0 \le \frac{d}{v_a + 2r/c}.$$

Also, since

$$v_a + 2r/c \le v_0 \le v_b - 2r/c,$$

this case applies when $v_b - v_a \ge 4r/c$. The corresponding expected connection time in one node encounter is

$$\begin{aligned} E[B_1(t_0)] &= \int_{v_a}^{v_0-2r/c} \frac{2r}{v_0 - v} dG(v) && (4.63)\\ &+ \int_{v_0-2r/c}^{v_0+2r/c} c\, dG(v) && (4.64)\\ &+ \int_{v_0+2r/c}^{v_b} \frac{2r}{v - v_0}\, dG(v) && (4.65)\\ &= \frac{2r \ln[\left(\frac{ce}{2r}\right)^2 (v_b - \frac{d}{t_0})(\frac{d}{t_0} - v_a)]}{v_b - v_a}. && (4.66) \end{aligned}$$

For case 4,

$$\begin{cases} v_0 - 2r/c \le v_a & v_0 \le v_a + 2r/c \\ v_0 + 2r/c \ge v_b & v_0 \ge v_b - 2r/c \end{cases}$$

Combining the two cases, we have

$$\frac{d}{v_a + 2r/c} \le t_0 \le \frac{d}{v_b - 2r/c}.$$

Also, since

$$v_b - 2r/c \le v_0 \le v_a + 2r/c,$$

this case applies when $v_b - v_a \le 4r/c$. The corresponding expected connection time in one node encounter is

$$E[B_1(t_0)] = \int_{v_a}^{v_b} c\, dG(v) = c \qquad \text{Q.E.D.} \tag{4.67}$$

## Appendix II: Derivation of Expected Connection Time in Reverse Traffic

$E[B_2(t_0)] =$

$$\begin{cases} \frac{2r \ln\left(\frac{v_b + d/t_0}{v_a + d/t_0}\right)}{v_b - v_a} & \frac{d}{v_b} \le t_0 \le \frac{d}{\max(v_a, 2r/c - v_a)} \\ \frac{2r \ln\left(\frac{d/t_0 + v_b}{2r} ce\right) - c(d/t_0 + v_a)}{v_b - v_a} & \frac{d}{\max(2r/c - v_a, v_a)} \le t_0 \\ & \le \frac{d}{\max(v_a, 2r/c - v_b)} \\ c & \frac{d}{\max(v_a, 2r/c - v_b)} \le t_0 \le \frac{d}{v_a} \end{cases} \tag{4.68}$$

Proof:

Instead of computing over the time variable $t$, we compute the connection time w.r.t. the speed random variable. For no truncation,

$$\frac{2r}{V+v_0} \leq c,$$

or

$$V \geq \frac{2r}{c} - v_0.$$

With reference to Figure 4.3. In case 1, there is no truncation at all observer node speed $v_0$. That is,

$$\frac{2r}{c} - v_0 \leq v_a,$$

or

$$v_0 \geq 2r/c - v_a.$$

As $r \to 0$, $v_0$ is bounded below by $v_a$. Therefore,

$$v_0 \geq \max(2r/c - v_a, v_a),$$

or

$$t_0 \leq \frac{d}{\max(2r/c - v_a, v_a)}.$$

The corresponding expected connection time in one node encounter is

$$E[B_2(t_0)] = \int_{v_a}^{v_b} \frac{2r}{v+v_0} dG(v) \tag{4.69}$$

$$= \frac{2r \ln\left(\frac{v_b + d/t_0}{v_a + d/t_0}\right)}{v_b - v_a}. \tag{4.70}$$

In case 2, there is partial truncation when the observer node speed satisfies

$$v_a \leq 2r/c - v_0 \leq v_b.$$

That is,

$$\begin{cases} v_0 \leq 2r/c - v_a \\ v_0 \geq 2r/c - v_b \end{cases}$$

Since $v_0$ is bounded below by $v_a$ as $r \to 0$, we have

$$\begin{cases} v_0 \leq \max(2r/c - v_a, v_a) \\ v_0 \geq \max(2r/c - v_b, v_a) \end{cases}$$

Combining the two cases, we have

$$\max(2r/c - v_b, v_a) \leq v_0 \leq \max(2r/c - v_a, v_a),$$

or

$$\frac{d}{\max(2r/c - v_b, v_a)} \geq t_0 \geq \frac{d}{\min(2r/c - v_a, v_a)}.$$

The corresponding expected connection time in one node encounter is

$$\begin{aligned} E[B_2(t_0)] &= \int_{v_a}^{2r/c - v_0} c \, dG(v) && (4.71) \\ &+ \int_{2r/c - v_0}^{v_b} \frac{2r}{v + v_0} dG(v) && (4.72) \\ &= \frac{2r \ln\left(\frac{d/t_0 + v_b}{2r} ce\right) - c(d/t_0 + v_a)}{v_b - v_a}. && (4.73) \end{aligned}$$

For case 3, there is truncation at all observer node speed.

$$v_b \leq 2r/c - v_0,$$

or

$$v_0 \leq 2r/c - v_b.$$

When $r \to 0$, $v_0$ is bounded below by $v_a$. Thus,

$$v_0 \leq \max(2r/c - v_b, v_a),$$

or

$$t_0 \geq \frac{d}{\max(2r/c - v_b, v_a)}.$$

$$E[B_2(t_0)] = c \qquad \text{Q.E.D.}$$

# Chapter 5

# On Network Connectivity and Energy Efficiency of Multihop Networks

## 5.1 Introduction

A mobile ad hoc network consists of mobile nodes that communicate with each other through multihop routing. The achievable capacity in these networks, however, is low as demonstrated by simulation studies [8, 12] and supported by analytical work [24]. Recently, power control [72, 78] and rate adaptation [26, 99] techniques have been proposed and shown demonstrative improvements on network capacity.

There are two basic paradigms in power control for mobile ad hoc networks. In the first paradigm, all nodes have a common transmit range that is predetermined. In the second paradigm, each node adaptively adjusts its transmit power [14, 49, 72] based on some heuristics and local channel measurements. Since nodes can adjust and transmit just enough power to the intended destination as time evolves, these algorithms are potentially superior due to the reduced interference, increased frequency reuse and improved energy efficiency. However, the bi-directionality of a link is no longer guaranteed. Since 802.11 is routinely used in ad hoc networks, all wireless links must be bi-directional such that the MAC layer functions properly. In this aspect, the first class of power control algorithms is more preferable. No extraneous signaling is needed to ensure links are bi-directional.

In this chapter, we restrict our scope to the first power control paradigm. The literature for this paradigm can be traced back to the seminal work of Kleinrock [44]. The transmit range is expressed as the mean number of neighbors of a node. Since then, there are a steady flow of followup work [28, 56, 87, 90, 91, 107] that address the same problem under different network models. The transmit range is optimized such that

the *expected forward progress*, or a related metric is maximized, where forward progress denotes the distance advancement of a packet in the direction of the destination node per transmission. The above literature, however, completely ignores the issue of network connectivity. [23] revisited the transmit range problem from a network connectivity perspective and determine the *critical transmit range* of a random network such that it is asymptotically connected with probability 1 when the number of nodes tends to infinity. In [24], it is further shown that under ideal network assumptions, a near optimal network capacity is attainable if all nodes operate at the critical transmit range. Nevertheless, analytical approaches are pursued in all these work, and the results are applicable to stationary networks only. Recently, [78] examined the effect of mobility on ad hoc networks by simulations and showed that optimal transmit range exists for mobile networks such that network throughput is maximized. The optimal transmit range is found to be increasing with node mobility.

Our literature review reveals that the optimal transmit range problem was approached with the objective of maximizing throughput and forward progress (which is also related to throughput), or ensuring network connectivity. The effect of transmit range on *energy efficiency* of packet transmissions, however, has never been studied before. We note that in practice, it is often more important to optimize for energy efficiency than throughput in a mobile ad hoc network. Since all mobile nodes are operated on stand-alone batteries, it is imperative to ensure all packet transmissions are energy efficient. Due to the relevance of energy efficiency in ad hoc networking, in this chapter we investigate the effect of transmit range control on the energy efficiency of packet transmissions. We quantify energy efficiency by defining the *energy per packet* $E_p$ metric. This represents the ratio of the total dissipated energy of all nodes to the total number of successfully received packets at the destinations. Our objective is to determine a common transmit range for all nodes such that energy per packet $E_p$ is minimized.

We performed *ns*-2 simulations to study the effect of transmit range control on energy efficiency. In the first part of the chapter, we consider stationary network only. We focus on three system parameters that affect energy efficiency of packet transmissions,

namely: energy dissipation model, offered load and path loss exponent of the channel. In our energy dissipation models, a node may expend no power or the same power as transmitting a packet during packet reception. If packet reception consumes significant energy, it is more energy efficient for nodes transmitting at a larger range and using routes that have fewer number of hops. Energy efficiency is also very sensitive to the transmit range at heavy offered load. When the network offered load is very heavy, it is highly energy inefficient for nodes to transmit at the critical range. When the path loss exponent of the channel is small, there exists an optimum transmit range that maximizes energy efficiency. Moreover, the optimum range is much larger than the critical transmit range. It turns out that energy efficiency is closely related to congestion and network connectivity. Three network connectivity regimes are identified to explain our observations, namely: *partitioned*, *weakly connected* and *strongly connected* regimes.

Our results that the optimal transmit range in a stationary network is much larger than the critical transmit range is contrary to the conclusion of some literature. In [24], Gupta and Kumar contended that a critical range just enough for network to be connected is near optimal in network capacity. The discrepancy is due to some simplifying assumptions made in [24], where network traffic is assumed to be homogeneous and is distributed evenly to all nodes. However, we show that when nodes operates at the critical transmit range (weakly connected regime), network traffic is highly non-homogeneous. Local congestion is dominant at some critical links. Since nodes have finite buffers, many packets are dropped due to buffer overflow along the critical links, leading to poor energy efficiency.

In the second part, we examine the effect of mobility on energy efficiency. The effect of mobility on the optimal transmit range was examined in [78]. It was found that the optimal transmit range is increasing with node mobility. In high mobility scenarios, a larger transmit range leads to less frequent link failure. This suppresses packet loss when a link failure occurs and the associated control overhead at route maintenance. The decrease in frequency reuse is more than compensated by more robust wireless links. In our simulations, we have used normal offered load instead of the network saturation offered load in [78]. We show that at normal offered load, there does not

exist a transmit range that optimizes network throughput. Nevertheless, an optimum transmit range exists such that energy efficiency is maximized. The optimal range turns out to be insensitive to node mobility, and is much larger than the critical range as advocated in [23, 24].

The rest of the chapter is organized as follows. We describe our simulation setup in section 5.2, followed by the discussion of simulation results in sections 5.3, 5.4 and 5.5. In 5.3, we identify three distinct *network connectivity regimes* for a stationary network as the transmit range of nodes increases. We then consider stationary networks in 5.4. The effect of system parameters on the optimum transmit range is discussed. In 5.5, the effect of transmit range control on energy efficiency is studied when node mobility is introduced. Finally, conclusions are drawn in section 5.6.

## 5.2 Simulation Setup

The transmit range of a mobile node depends on its transmitted power, $P_t$, and the propagation loss. In our study, we focus on the effect of distance attenuation, which is characterized by the path loss exponent $\beta$. Second-order effects such as shadowing and multipath fading are ignored. To model the propagation loss, we adopt the following free space model:

$$P_r(d) = \frac{P_t(4\pi)^2}{\lambda^2} d^\beta, \tag{5.1}$$

where $Pr(d)$ is the received power at distance $d$, and $\lambda$ is the wavelength.

In mobile environments, $\beta$ typically ranges from 2 to 4. In our simulation, we consider the cases $\beta = 2, 3$ and 4. When $\beta = 2$ and 3, we apply the above model. When $\beta = 4$, however, the attenuation according to the above equation becomes unrealistically large. Therefore, for this case, we use the more realistic two-ray ground model:

$$P_r(d) = \begin{cases} P_t\lambda^2/[(4\pi)^2 d^2] & d < 4\pi h_t h_r/\lambda \\ P_t h_t^2 h_r^2/d^4 & d \geq 4\pi h_t h_r/\lambda. \end{cases} \tag{5.2}$$

In this model, when the communication distance increases beyond the crossover distance (far field), the path loss exponent changes from $\beta = 2$ to $\beta = 4$. We note that in the

far field, the free-space and the two-ray models have the same signal rolloff and differ only in the constant factor preceding the function $d$.

To study energy efficiency, it is important to understand how energy is consumed. Clearly, a mobile node consumes energy either it is transmitting, receiving, or staying idle. For state of the art hardware technology, the idle power of a node is comparatively small and is ignored in our study. On the other hand, the reception of a packet involves meticulous signal processing techniques from synchronization to decoding to equalization. The power consumption can be significant compared to the power of transmitting a packet. Therefore, it is necessary to examine the energy expenditure on packet reception as well as packet transmission. In our study, we consider two energy dissipation models. In model 1, a node consumes no energy when it receives a packet. In model 2, a node consumes the same amount of energy whether it transmits or receives a packet. In either case, the energy consumption of a packet transmission is given by the transmit power times the packet duration. These models represent two extreme cases and allow us to examine how energy consumption in packet reception affects energy efficiency of the network.

The simulations are performed on *ns*-2 [1], with its wireless extensions developed by the Monarch project [2]. We assume that there are 100 mobile nodes distributed uniformly in a $1000m$ by $1000m$ area. For a given channel model with known $\beta$, we simulate 16 power levels such that the transmit range corresponds to $75m$ to $450m$ at a step of $25m$.

The mobile nodes emulate 914 MHz Lucent WaveLAN DSSS radio interfaces. The nominal bit rate is 2 Mbps. Omni-directional antennas with 0 dB gain are used, and antennas are placed $1.5m$ above the ground. The receive threshold is 3.652e-10 W or -64.37 dBm, which determines the minimum SIR required for successful decoding of a received packet. The carrier sense threshold is 1.559e-11 W or -78.07 dBm. Any packet with a SIR more than the threshold may interfere with reception of another packet.

Nodes move in the network under the random waypoint mobility model. Node movements consist of alternate mobility and pause epochs. During a mobility epoch, a node moves with constant speed in a particular direction. We assume that the speed

is uniformly distributed between 0 and *max_speed*. In the pause epoch, nodes stop at the current position for a fixed duration of *pause_time*. We characterize node mobility using the parameter *max_speed* and keep the *pause_time* equal to 1 second in all mobility scenarios. Four different values of *max_speed* are investigated in our study, namely $v = 0, 5, 10, 20m/s$. These values correspond to the stationary, fast pedestrian, slow and fast vehicular scenarios.

The traffic is generated through a CBR application over UDP [8,12], which simulates the performance of the best effort delivery paradigm. The offered load can be varied by any of the three parameters, namely packet transmission rate, packet size and the number of traffic *flows* in the network, where each flow denotes a predetermined source destination pair that is randomly chosen. We simulate three offered loads regimes, as shown in Table 5.1. The traffic types 1 to 3 correspond to a network operating in the *light*, *normal* and *saturation* load regimes respectively. Packet sizes are chosen such that fragmentation occurs on neither the network nor the MAC layer.

In this chapter we use the *dynamic source routing* (DSR) algorithm [33] in our simulations, since DSR shares many of the salient characteristics typical to reactive routing algorithms. The DSR runs on top of the 802.11b standard with a channel reservation mechanism enabled by the use of *request to send* (RTS) and *clear to send* packets. In general, packet loss can result from contention in wireless transmissions, unavailability of route due to mobility, or buffer overflow due to congestion. Nevertheless, the RTS/CTS mechanism in the 802.11b standard is efficient in combating the hidden terminal problem. We note that in the first part of our simulations, we consider stationary networks only. By factoring out mobility into our consideration, most packet loss that occurs are due to congestion. This allows us to examine in detail the inter-relationship between network connectivity, congestion and energy efficiency. Node mobility is incorporated in the second part of our simulations so that the relationship between node mobility and energy efficiency can be studied.

To summarize, we have four system parameters in our simulation model, namely: offered load (light, normal and saturation), path loss exponent ($\beta = 2, 3, 4$), energy dissipation for receiving packets (Model 1 and 2), and node mobility (stationary, pedestrian,

| Traffic type | packet rate | packet size | number of flows | total load |
|---|---|---|---|---|
| 1 | 5/s | 64Byte | 20 | 51.2Kbps |
| 2 | 10/s | 64Byte | 20 | 102.4Kbps |
| 3 | 20/s | 768Byte | 20 | 2.458Mbps |

Table 5.1: Traffic parameters adopted in the numerical studies

slow and fast vehicular). These four parameters together define a *network scenario*. For each network scenario, ten topology realizations are simulated. Each simulation lasts for 300 sec. Each flow starts at a staggered time that is uniformly distributed between 0 and 100 sec. Simulation data is logged during the interval between 100 sec. and 300 sec. to ensure the network has reached a steady state. In each topology realization, all flows are monitored.

We focus on two performance metrics in this chapter. The goodput, $G$, denotes the fraction of packets that is correctly received. It is essentially proportional to throughput and is between 0 and 1. We evaluate the energy efficiency in terms of energy per packet $E_p$. This represents the ratio of the total dissipated energy of all nodes to the total number of successfully received packets at the destinations. The performance metrics of a particular network scenario are computed over all monitored flows and averaged over all topology realizations of the network scenario.

## 5.3 Network Connectivity Regimes and Goodput

In this section and the next, we consider only the stationary scenario. As the nodes are not moving, there are only two reasons for packet losses. If the traffic load is sufficiently light, packets can be lost only if the destination is not reachable by the source. If the traffic load is heavy, packets may be lost due to buffer overflow, which is a manifestation of network congestion. These observations enable us to define the critical range and the optimal range for a given stationary network scenario, as explained below.

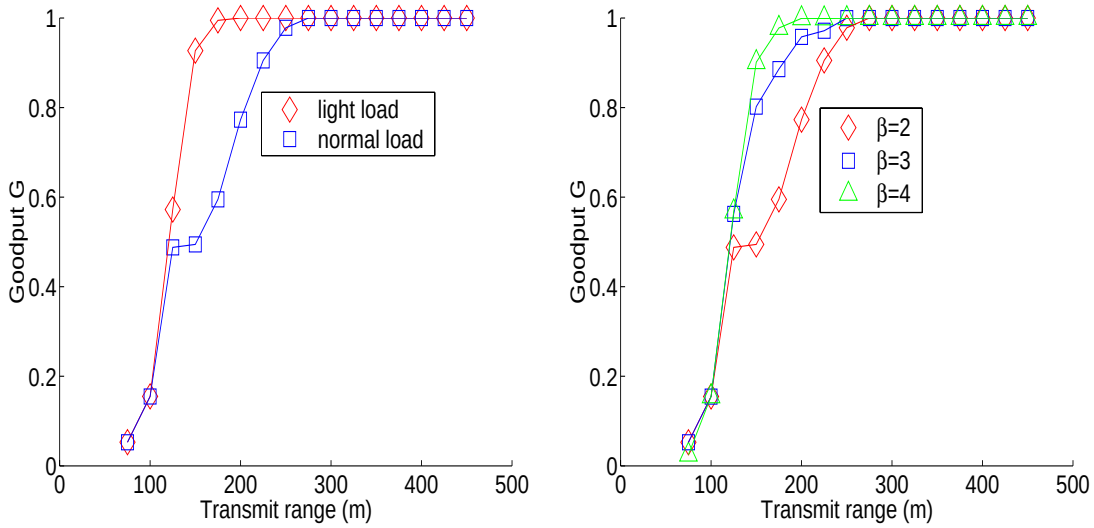



Figure 5.1: Goodput vs. Transmit Range. (a) Comparison of light and normal offered load regimes when path loss exponent $\beta = 2$, (b) Comparison of different path loss exponents $\beta$ at normal offered load scenario.

### 5.3.1 Network Connectivity

In Figure 5.1(a), the goodput at light and normal offered load scenarios is plotted for path loss exponent $\beta = 2$. In the case of *light* offered load, almost all packet losses are due to network partitioning. At a transmit range of $r = 75m$, the network is heavily partitioned. Packets generated at the source nodes are queued in the node buffers waiting for transmission. If a source has no neighbor nodes in its communication range, the DSR algorithm gives up finding a route upon several failed attempts at route discovery. No data transmission is ever attempted. As new packets are continually generated, buffer overflow ultimately occurs and the packets are dropped at the source nodes.

As the transmit range increases, network partitioning is less extensive. Routes can be established for some flows. The goodput increases with the transmit range since an increasing fraction of all flows establishes routes successfully. At a range of $r = 175m$, the network is almost connected under all topology realizations, with a goodput $G \simeq 1$. We define it as the *critical range* $r_c$ of the network scenarios. When

nodes operate beyond the critical range, all nodes are reachable to each other, and the network is *connected*. Similarly, when the transmit range is smaller than the critical range ($r < r_c$), the network is partitioned. We denote the network as operating in the *partitioned regime* when $r < r_c$.

Refer to Figure 5.1(a) again, at *normal* offered load the goodput increases more gently. The goodput hits the maximum of 1 when the transmit range is close to $r = 275m$. In our light and normal offered load simulations, the same topology realizations are used. Thus packet losses in the range $175m < r < 275m$ at normal offered load is not due to network partitioning. In particular, when $r = 175m$ the goodput at normal offered load is barely above 0.6. Operating at the vicinity of this transmit range is referred as the *weakly connected regime* ($r \simeq r_c$). We infer that packet losses are mainly due to congestion in this regime.

When the transmit range is increased further, the network evolves from weakly connected to *strongly connected* ($r \gg r_c$). At the *strongly connected regime*, multiple routes exist for each flow. The network traffic is more evenly distributed to all nodes and links. This reduces the occurrences of traffic hot spots, leading to an improved goodput. At very large transmit power, all nodes have direct links to each other. The network reduces to a broadcast network. Surprisingly, the network goodput does not suffer at large transmit range at normal offered load. The decrease of frequency reuse with large transmit range does not adversely affect the network goodput. In most reported ad hoc network simulations [8, 12], the network is operated at normal offered load $100Kbps$. This is much smaller than the nominal bit rate of $2Mbps$. The CSMA multiple access mechanism in the 802.11 MAC efficiently schedules all packet transmissions even if there is no frequency reuse. Thus the goodput is increasing with the transmit range.

### 5.3.2 Optimal Transmit Range

In [24], it was shown that under some ideal network assumptions, network capacity is near optimal when nodes are operated at the *critical range*. Although an increase in the transmit range reduces the number of hops to the destination node, it also decreases the number of simultaneous transmissions in the network. Suppose the transmit range is

doubled, the number of hops to the destination is halved on the average. However, the coverage area of a node is also quadrupled. In order to avoid a collision, all other nodes in the range of a receive node should not transmit, reducing the number of concurrent transmissions in the network by four times. It turns out that the tradeoff between the number of hops and the number of simultaneous transmissions always favors nodes with the smallest range such that the network is connected.

Our simulation results in Figure 5.1, however, indicate that a network is prone to congestion when it is operated in the weakly connected regime ($r \simeq r_c$). This discrepancy is due to the simplistic assumptions made at [24]. The derivation in [24] rests on the assumption that the total offered load (including multihop forwarding) equals the total one hop capacity, i.e. the maximum number of concurrent transmissions times the bandwidth. Routes are hypothetical and defined by drawing a straight line between the source and destination nodes. The number of hops is then proportional to the distance between the end nodes.

We note that in the above model, there is an implicit assumption of traffic homogeneity, in which all packet transmissions are distributed evenly to all nodes. As long as the total number of one hop transmissions at one snapshot can be accommodated to all simultaneously transmitting nodes, it is assumed all transmissions can be received successfully at the destination nodes. There is no consideration of congestion due to the non-homogeneity of traffic over space. In reality, the spatial distribution of traffic is not homogeneous. Congestion occurs at local hot spots and packets are dropped. For example, in our simulation, there is no network partitioning at $r = 175m$. However, the network remains weakly connected in the sense that some links are critical to network connectivity. The network will be partitioned otherwise if the *critical links* are removed. As a result, many flows route through the critical links as the intermediate paths. Local congestion occurs along the critical links and significantly degrades the throughput performance.

Our argument is graphically illustrated in Figure 5.2. Observe that the edge between node $A$ and node $B$ is a critical link. All nodes on the left have to route through link $A - B$ to reach any node on the right. In this example, three flows are merging into

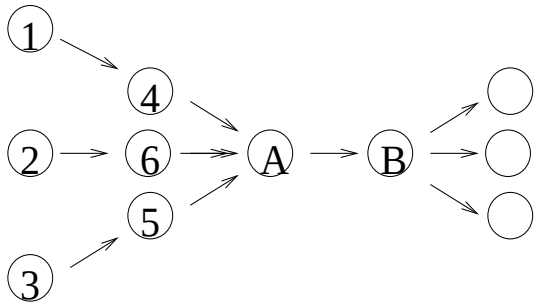



Figure 5.2: Constraint on frequency reuse along a multihop route due to the heavy merging traffic at node $A$ preceding the critical link.

the critical link. Due to the incoming traffic, there are many transmissions in the proximity of node $A$. The sum of interference for nodes several hops away from node $A$ can be very large. This suppresses node A from seizing the channel and proceed to data transmission. Consequently, a backlog of incoming packets is built up at node $A$, creating a traffic bottleneck. When the buffer is full finally, packets are dropped due to congestion. Although the network is not partitioned, the proliferation of network traffic preceding the critical links leads to congestion and packet dropping.

### 5.3.3 Effect of Path Loss Exponent

In Figure 5.1(b), the goodput at normal offered load is plotted when the value of path loss exponent is varied. We observe that as $\beta$ increases, the goodput also increases correspondingly. In particular, when the network is weakly connected ($r \simeq r_c$), the goodput is approximately 0.6,0.9,0.95 for $\beta = 2, 3, 4$ respectively. This indicates there is few congestion for $\beta = 3$ and almost no congestion for $\beta = 4$.

In general, when the path loss rolloff $\beta$ is steeper, frequency reuse is more efficient. Consider a network of $n$ nodes. Suppose the intended receiver is located at a distance $r$ from the transmitter. Due to the constraint of frequency reuse, a transmission is not successful if there is another transmission within a distance $(1 + \Delta)r$ from the intended receiver, where $\Delta > 0$ is a measure of frequency reuse distance. Let $\gamma_{min}$ be the minimum signal interference ratio (SIR) at the receiver for successful reception. It is straightforward to show that $\gamma_{min} = (1 + \Delta)^{\beta}$, or $\Delta = \gamma_{min}^{\frac{1}{\beta}} - 1$. Thus, when

path loss exponent $\beta$ increases, $\Delta$ decreases. The decrease of frequency reuse distance with increasing $\beta$ allows more simultaneous transmissions in the network, leading to an increase in goodput. Our results are consistent to [41], where network throughput is derived under simplified assumptions. It was shown that the average throughput $\nu$ per node can be supported if

$$\nu \leq \frac{c}{rn\Delta^2}$$

for some constant $c$. It is obvious that when $\beta$ increases, $\Delta$ decreases and $\nu$ increases.

Refer to Figure 5.2, the increase in goodput with large $\beta$ can be understood as follows. When $\beta = 2$, the frequency reuse distance is $1 + \Delta(\beta)$ is large. If any of the nodes 1 to 6 transmits, the interference at node $A$ is large such that $A$ refrains from transmission. When $\beta$ increases, the frequency reuse distance decreases correspondingly. This effectively reduces the number of interferers that affect node $A$, and increases the packet transmission probability. As a consequence, it is more unlikely for node $A$ to build up a backlog and drops the packets when buffer overflows.

## 5.4 Optimal Energy Per Packet $E_p$

In this section, we study the effect of transmit range on energy efficiency of an ad hoc network. As a prelude, consider the simple scenario that there are a number of nodes locating on a straight line. Suppose a node wants to transmit a packet to another node that is not the closest neighbor of the source node. It is easy to show that relaying the packet by intermediate nodes requires less energy than transmitting it direct to the destination node. This phenomenon is due to the nonlinear path loss attenuation. This simple observation suggests that it is energy efficient to choose a small transmit range.

However, things are not that simple. First of all, energy is spent not only on transmission, but also on packet reception and decoding. We will study this effect by means of the two energy dissipation models. Second, network congestion may occur if the network operates at the weakly connected regime. This complicates the problem; the simple observation may no longer be valid. We will discuss this issue in the next subsection first. Afterwards, we present our results using two different dissipation

models. Finally, we discuss the effect of path loss exponent.

### 5.4.1 General Trend at $\beta = 2$

In Figure 5.3(a),(b), energy per packet $E_p$ is plotted versus transmit range for two energy dissipation models. The $E_p$ of light and normal offered load scenarios are plotted on the same graph for comparison. Similarly, in Figure 5.3(c),(d) energy per packet $E_p$ is plotted versus transmit range for two energy dissipation models at network saturation. These results are obtained with path loss exponent keeping constant at $\beta = 2$. With the exception of light offered load scenario, the variation of $E_p$ for other offered load scenarios and different energy dissipation models follows a general trend as the transmit range increases. $E_p$ hits a local maximum as the transmit range increases. Further off the local maximum, $E_p$ dips into a local minimum before increasing again as the transmit range is increased further.

To explain the general trend, we consider in the following the normal offered load scenario under energy dissipation model 1, where there is no energy consumption for packet reception. In the *partitioned regime* ($r < r_c$), nodes have small transmit power. Due to network partitioning, routes exist only between source-destination pairs that are in close proximity. Since energy consumption per hop and the number of hops in a route is small, all successful packet transmissions are energy efficient.

The majority of source nodes, however, fail to discover a valid route to the destination nodes at DSR route discovery. The source nodes make a few attempts to broadcast *Route Request* packets locally. If no *Route Reply* packets are received after some timeout, route discovery is aborted. The source nodes perform route discovery again after some timeout for a few more times before finally giving up discovering a route. In a heavily partitioned network, few nodes receive the Route Request packets and reply with the Route Reply packets in the first place. Thus energy dissipation due to control messages can be ignored. Moreover, the size of a Route Request or Route Reply packet is only a fraction of a data packet. Thus negligible energy is expended on control packets during route discovery.

For the source nodes that give up route discovery after several attempts, incoming

packets are queued at the source nodes due to the lack of routes. These packets are dropped subsequently when node buffers are full. Since the data packets are never sent to the air, there is no energy penalty in dropping those data packets. Thus no energy is expended on the data packets that are eventually dropped. This explains the small value of energy per packet $E_p$ in the partitioned regime.

As the network evolves to the *weakly connected regime* ($r \simeq 175m$) when transmit range increases, $E_p$ increases to a local maximum at $r = 150m$ at normal offered load scenario. As discussed in the previous section, when the network is weakly connected, congestion occurs along the critical links and many packets are dropped. This is detrimental to energy efficiency, as well as network goodput. Recalled from Figure 5.1(a) that at the critical range $r = 175m$, almost half of the packets are dropped. Some of the dropped packets may have already been forwarded for a few hops. Since significant energy is expended on the dropped packets, the overall energy per packet $E_p$ is artificially higher.

As the transmit range is increased further, there are more routes between any source destination pair in the *strongly connected regime*. The network traffic is more evenly distributed across the whole network. This reduces packet loss and drives down $E_p$ to a local minimum at $r = 200m$. After congestion is resolved, a further increase in the transmit range is no longer beneficial. Since there is no gain in goodput by transmitting at a larger power, $E_p$ increases with further increase of transmit range.

In general, one would like to minimize the energy per packet as much as possible without sacrificing network connectivity. This corresponds to the transmit range within the weakly and strongly connected regime such that $E_p$ is minimum. That is, the local minimum on the $E_p$ curve is the optimum point for network operation.

For comparison purpose, we also plot the behavior of $E_p$ for light offered load scenarios on the same graph. It is obvious that $E_p$ for the light and normal offered load regime matches closely in the partitioned and strongly connected regimes. The $E_p$ for light offered load scenario, however, is slightly larger. This is because the number of data packets is smaller at light offered load, whereas the amount of control overhead remains the same. On the other hand, we observe discrepancies arise for the weakly

connected regime ($r \simeq r_c$). Recall that the maximum and minimum in $E_p$ at normal offered load corresponds to the build up and resolution of congestion. Similar trends cannot be observed at light offered load, where $E_p$ is strictly increasing with the transmit range. The absence of local extrema at light offered load is consistent to the fact the congestion is absent at light offered load scenario.

Nevertheless, in practical ad hoc networks, offered load is non-negligible. Figure 5.3 shows that a minimum exists for both normal and heavy network offered loads and for both energy dissipation models under consideration. This is an interesting observation since for normal offered load, goodput (Figure 5.1(a)) is always increasing with the transmit range. The determination of an optimal transmit range based on goodput is quite arbitrary. Our results in Figure 5.3 shows that under different network load and energy dissipation models, an optimum transmit range exists such that energy efficiency is maximized for the case when $\beta$ is small. Furthermore, this optimum range is larger than the critical range.

### 5.4.2 Energy Dissipation Model

Figure 5.3(b) shows the result for energy model 2, where power consumption of packet transmission and reception is the same. As the transmit range traverses the three network connectivity regimes, we observe the same qualitative trend for $E_p$. The energy per packet quickly rises to a local maximum due to congestion, falls again to a local minimum, and rises again as a further increase in transmit power incurs no goodput improvement. We observe that the energy dissipation is much larger for energy model 2. At the optimal range of $r = 275m$, the energy dissipation is approximately 12 $mW$, which is approximately 100 times the energy dissipation for model 1 at the same range. The significant energy consumption for model 2 is due to unsolicited packet reception. By default, nodes operate in promiscuous mode in DSR. A node listens to all packets even if the packets are not addressed to itself. Whenever a node receives a packet, it eavesdrops the packet header, extract any new routes and updates its route cache. In particular, when the transmit range is large, each node has many neighbors. This leads to a lot of unsolicited packet reception. A large fraction of energy is thus spent

to packet reception rather than packet transmission.

As an illustration, consider two transmit ranges $r = 75m$ and $r = 175m$. When $r = 75m$, the expected number of neighbors of a node is $\mu\pi r^2 = \frac{100}{1000^2}\pi(75)^2 = 1.761$, where $\mu$ is the node density. Similarly, when $r = 175m$, the expected number of neighbors of a node is $\mu\pi r^2 = \frac{100}{1000^2}\pi(175)^2 = 9.619$. We observe as the transmit range increases from $r = 75m$ to $175m$, $E_p$ is increased by 3.5 times for energy dissipation model 1 in Figure 5.3(a). The corresponding increase for $E_p$ for energy dissipation model 2 is 28 times in Figure 5.3(b). It is evident that the steeper increase in energy expenditure with range for model 2 is due to the increase in unsolicited packet reception at large transmit range.

We also observe the optimal transmit range for energy dissipation model 1 and 2 is $r = 200m$, and $r = 275m$ respectively. When a node expends energy to receive a packet, multihop routing is less attractive. Our results show that when it is expensive to receive packets, the optimal transmit range is larger such that the average number of hops in a route is reduced. This minimizes the energy lost along a route due to packet reception.

In Figure 5.3(c)(d), $E_p$ is plotted for both energy dissipation models at network saturation. At a range of $r = 250m$, $E_p$ in energy model 1 and 2 are respectively 0.8 $mW$ and 70 $mW$. Similar to the normal offered load scenario, the energy consumption for model 2 is two orders of magnitude more than that of model 1. Again, this observation can be explained by the unsolicited packet reception argument. We note that at a range of $r = 250m$, $E_p$ of the energy dissipation models at normal offered load is 0.15 $mW$ and 13 $mW$. Thus, the energy consumption at network saturation is approximately 5.33 times larger than the normal offered load scenarios for both energy dissipation models. At network saturation, congestion is a network-wide phenomenon. Since packet loss due to congestion is common, $E_p$ is much larger than that in the normal offered load scenarios.

We observe that when the network offered load is large, energy per packet $E_p$ is sensitive to the choice of transmit range. In particular, the ratio of maximum and minimum in $E_p$ for the network saturation scenario under energy dissipation model 1

is 1.66, compared with 1.17 in the normal offered load scenario. Similarly, the ratio of maximum and minimum in $E_p$ for the network saturation scenario under energy dissipation model 2 is 1.25, compared with 1.07 in the normal offered load scenario. At network saturation, the network is more prone to congestion at the weakly connected regime. The extra stress in offered load leads to severe packet loss along the critical links connecting the network. This implies that when the network is operated in some more stressful scenarios, it is imperative that an optimum transmit range should be used. If a critical range is used for stationary networks, such as a deployment of sensors, the improper setting of the network topology will lead to poor energy efficiency and the premature depletion of battery energy in sensors.

### 5.4.3 Path Loss Exponent

The effect of the path loss exponent $\beta$ on energy efficiency is also examined. In Figure 5.4(a)(b), the $E_p$ at normal offered load is plotted against the transmit range for different $\beta$. In both cases, we use energy dissipation model 2. We observe that an optimal $E_p$ does not exist when $\beta = 3, 4$. Recall the discussion on three network connectivity regimes, When $\beta = 2$, congestion is severe at the *weakly connected regime*. This leads to a local maximum in $E_p$, where many packets are dropped. The subsequent increase in transmit range decreases the energy per packet since packet loss due to congestion is resolved. In contrast, there is few congestion when $\beta = 3$ and almost no congestion for $\beta = 4$ at the *weakly connected regime*. Since an increase in transmit range beyond the critical range does not reduce packet loss, it is not energy efficient for nodes to transmit beyond the critical range.

For the cases $\beta = 3, 4$, we also observe that both goodput $G$ and energy per packet $E_p$ are increasing with the transmit range. In general, a network should be operated at a transmit range that is the most energy efficient without sacrificing network connectivity. With this rationale in mind, operating the network at *critical range* $r_c$ is a good choice. At this range, minimum energy is expended per packet while network connectivity is almost guaranteed by allowing a goodput of 0.95 and 0.99 respectively when $\beta = 3, 4$.

In typical wireless environments, a path loss exponent of $\beta = 3, 4$ is common. The

assertion of [24] on the optimality of the critical range is valid in many cases since congestion is mostly absent in the weakly connected regime when $\beta = 3, 4$. However, in some application environments such as airborne platforms where free space path loss dictates, or in short range communication networks where there is a direct line of sight between the communicating nodes, $\beta = 2$ holds and the optimal transmit range is very different from the critical range due to congestion.

## 5.5 Effect of Mobility

Finally, the effect of mobility on the optimum transmit range is also studied. Similar study was reported in [78], in which the transmit range that maximizes throughput is determined. It was found that at network saturation, the optimum transmit range increases with node mobility. Although a larger transmit power reduces the number of simultaneous transmissions over space, the decrease in frequency reuse is more than compensated by a more robust topology. Link failures are less frequent when the transmit range is large. This impacts the network throughput favorably due to less packet loss during link failure, and the reduction of the associated control overhead in route maintenance.

We have performed similar simulations to find the optimum transmit range such that energy per packet is minimized in different mobility scenarios. Instead of studying network saturation, we consider normal offered load scenarios in our simulations, which models the approximate traffic load we expect in a real ad hoc network. Four mobility scenarios are considered, namely: stationary, fast pedestrian, slow and fast vehicular scenarios. In each case, we use energy dissipation model 2 and assume $\beta = 2$.

As shown in Figure 5.5, the energy per packet is plotted versus transmit range for different mobility scenarios. The optimum range is $r = 275m$ in stationary and pedestrian scenarios. For slow and fast vehicular scenarios, the optimum transmit range is $r = 300m$ and $r = 325m$ respectively. Thus, the optimum transmit range also increases slightly with node mobility when energy per packet is used as the metric. Nevertheless mobility has little effect ($\pm 8.33\%$) on the optimum transmit range, with

$r = 300m \pm 25m$. Note that the optimal range is much larger than the critical distance $r_c = 175m$ in our simulations.

We argued earlier that it is energy inefficient to operate a network in the weakly connected regime owing to congestion. In the presence of node mobility, the energy inefficiency of operating a network at a small transmit range is more apparent. From Figure 5.5, the ratio of the local maximum and local minimum in $E_p$ is 1.1241 for the stationary scenario. The ratio increases slightly to 1.7202 for the fast pedestrian scenario and hits 2.7674 and 3.8903 respectively for the slow and fast vehicular scenarios. We note that in all mobility scenarios, the local maximum occurs at a transmit range close to the critical range. Thus it is highly inefficient for the network to operate at the critical range, especially in high mobility scenarios.

It is intuitively obvious that an increase in node mobility leads to more packet loss, and thus, decreases energy efficiency. Thus the energy per packet $E_p$ for each value of the transmit range is larger than the corresponding values of a more sedentary mobility scenario. The increase in the ratio of the local maximum to local minimum, on the other hand, indicates that node mobility is more detrimental to energy efficiency of the network in the weakly connected regime (where the local maximum occurs). When the network is weakly connected, the network is prone to link failure due to node mobility. The increased occurrence of link failures induces more packet loss along a route and demotes energy efficiency. A large transmit range effectively solves the problem of link failure, at the expense of reduced spatial concurrency. This, however, does not impact energy efficiency nor the throughput adversely. At normal load, the system bandwidth is large enough such that no congestion occurs even if nodes have infinite power. The decrease of spatial concurrency does not trigger any congestion that leads to packet loss at normal offered load.

Figure 5.6 depicts the corresponding goodput versus transmit range for the same mobility scenarios. We observe that at normal load, goodput is increasing with transmit range under all mobility scenarios. It is ambiguous to define an optimum transmit range based on goodput. On the other hand, when energy efficiency is used as the optimization metric, we show that an optimum transmit range always exists for all mobility

scenarios. Moreover, in contrast to [78], our optimum transmit range is not sensitive to mobility. The optimum transmit range is about $r = 300m$ under all mobility scenarios we considered. It is, however, very different from the critical range as advocated in [24].

The insensitivity of the optimal transmit range to mobility can be attributed to the sharp increase of goodput for all mobility in Figure 5.6. The goodput approaches 1 around $r = 300m$ under all mobility scenarios. Although a decrease in transmit power reduces the expended power in each packet transmission, significant energy is wasted on the large fraction of packets that are dropped in the network, which contributes to a higher overall energy per packet $E_p$.

## 5.6 Conclusion

In this chapter, we address the effect of transmit range control on the energy efficiency of both stationary and mobile ad hoc networks. We show that with an optimal choice of transmit range, significant energy savings can be obtained in some scenarios. In the first part of the chapter, we consider stationary network. The dependence of energy per packet on several system parameters in a stationary network is examined. This includes path loss exponent of the channel, energy dissipation model, and the offered load. In particular, we show that when the path loss exponent is small, a critical transmit range is suboptimal in throughput and energy efficiency. An optimal transmit range exists that maximizes energy efficiency, which is much larger than the critical transmit range. Congestion plays an important role that underlines our observations and is intimately connected to network connectivity. Three network connectivity regimes are identified as the transmit range of all nodes increases. In the second part, we also examine the effect of mobility on energy efficiency. Our results show that at normal offered load, there does not exist a transmit range that optimizes the network throughput. Nevertheless, an optimum transmit range exists such that energy efficiency is maximized. The optimal range turns out to be invariant to node mobility, and is much larger than the critical range as advocated in some literature.

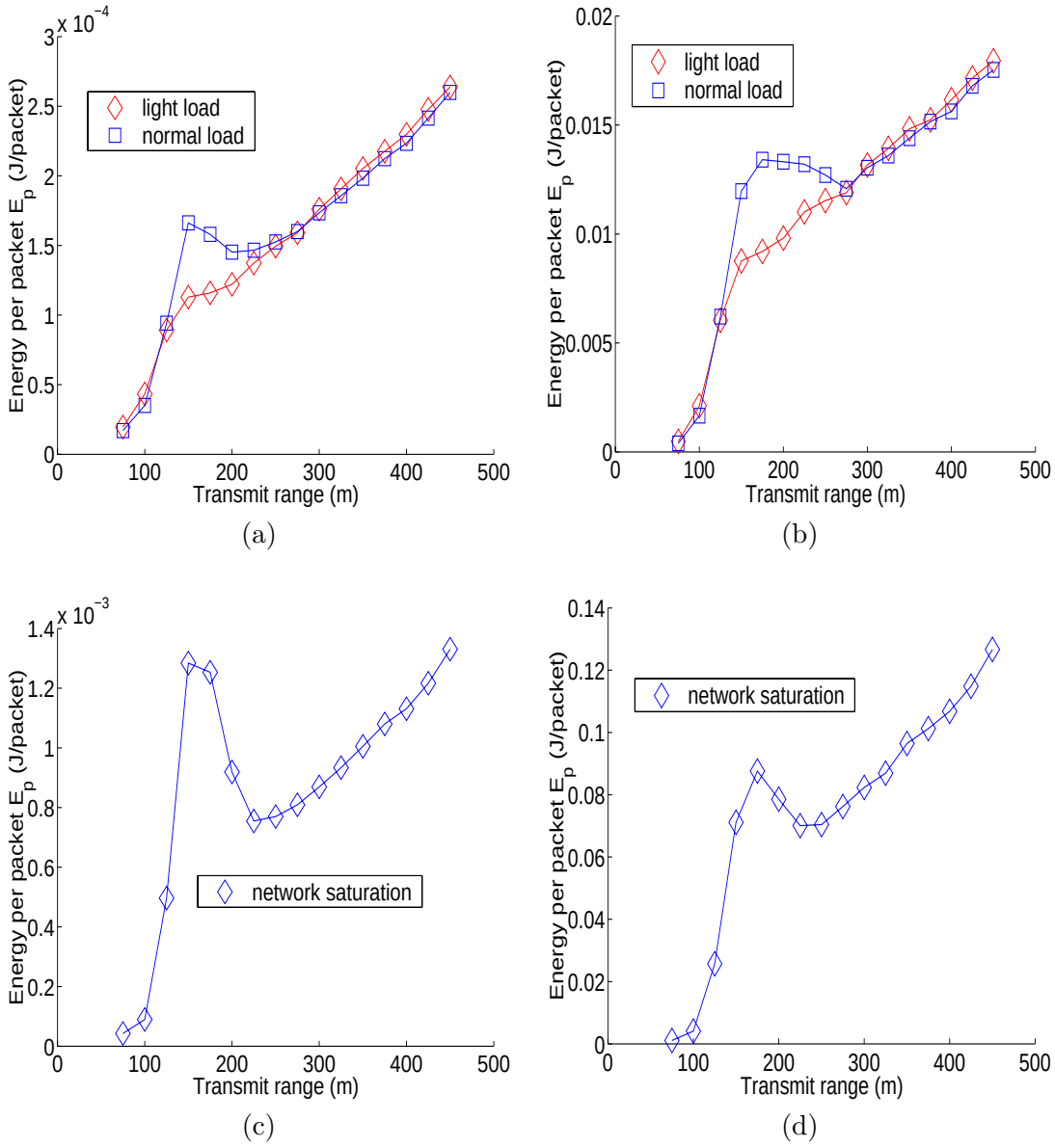



Figure 5.3: Energy per Packet vs. Transmit Range. (a) Normal offered load, energy model 1, (b) Normal offered load, energy model 2, (c) Network saturation, energy model 1, (d) Network saturation, energy model 2.

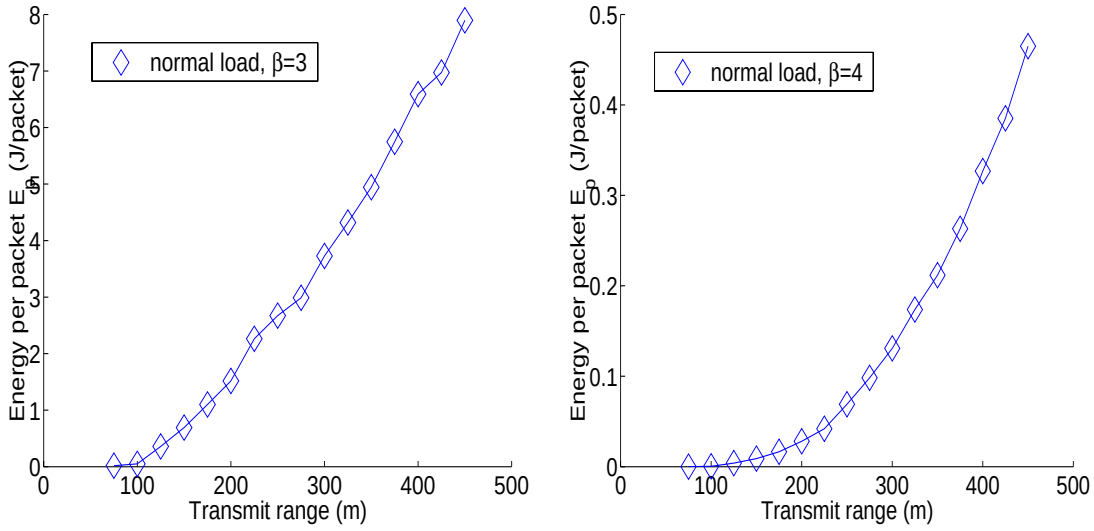



Figure 5.4: Energy per Packet vs. Transmit Range. (a) Normal offered load, energy model 2, path loss exponent $\beta = 3$. (b) Normal offered load, energy model 2, path loss exponent $\beta = 4$.

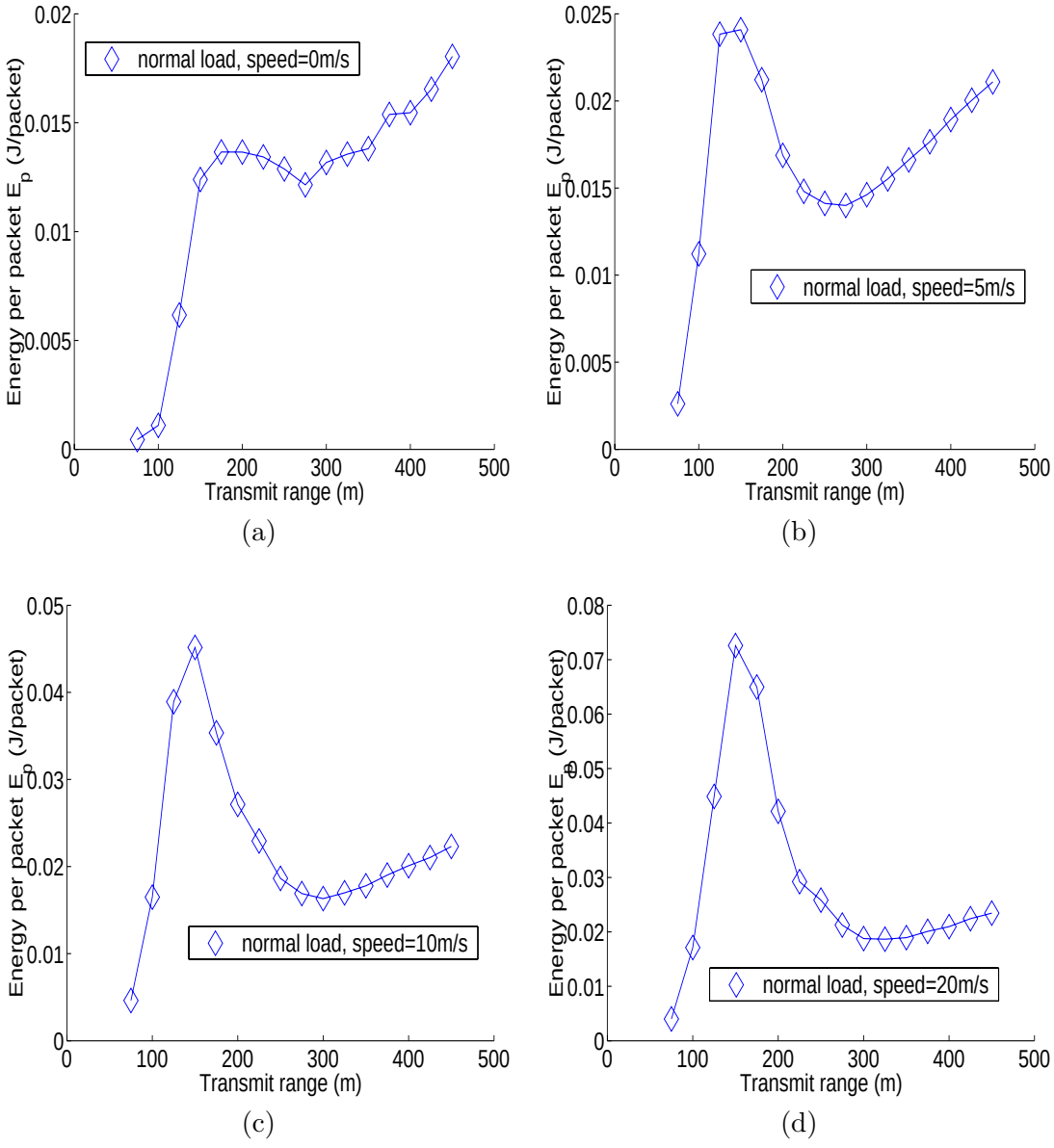



Figure 5.5: Energy per Packet vs. Transmit Range. Normal offered load, energy model 2, $\beta = 2$. (a) stationary scenario (speed=0m/s), (b) fast pedestrian scenario (speed=5m/s), (c) slow vehicular scenario (speed=10m/s), (d) fast vehicular scenario (speed=20m/s).

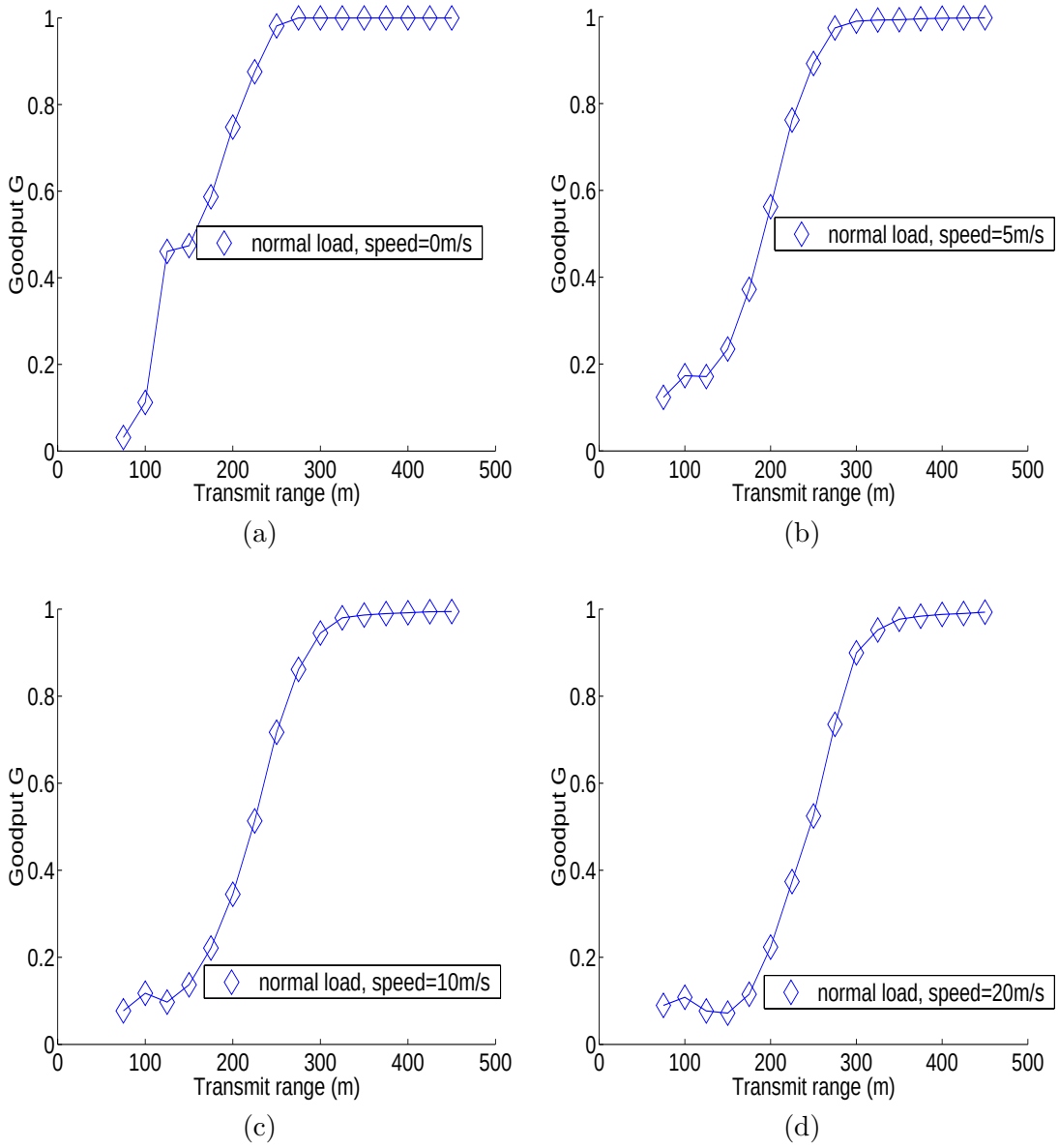



Figure 5.6: Goodput vs. Transmit Range. Normal offered load, energy model 2, $\beta = 2$. (a) stationary scenario (speed=0m/s), (b) fast pedestrian scenario (speed=5m/s), (c) slow vehicular scenario (speed=10m/s), (d) fast vehicular scenario (speed=20m/s).

# Chapter 6

# Inter-relationships of Performance Metrics and System Parameters in Mobile Ad Hoc Networks

In the previous chapter, we have studied the effect of transmit range on energy efficiency of multihop ad hoc networks. It is somewhat surprising that the critical range of Gupta and Kumar does not yields optimal throughput in the stationary scenario. In this chapter, we will look into the network behavior of multihop ad hoc networks. By using an alternate graphical representation of simulation results, we observe interesting inter-relationships between the performance metrics and the system parameters of the network. This sheds light to the mechanisms that underline the network behavior. Our result confirms that the conjecture that throughput improvement in node mobility is due to load balancing is errorneous, another interesting result.

## 6.1 Introduction

A mobile ad hoc network consists of mobile nodes which communicate with each other through multihop routing. Due to the dynamically changing topology, network routing is an important issue. Numerous routing algorithms [13, 22, 25, 30, 33, 47, 50, 54, 60, 62, 63, 73, 79, 84, 93, 94] have been proposed to facilitate efficient packet delivery in mobile environments. The focus is on the relative performance of routing algorithms, which are often characterized by a few performance metrics such as packet goodput, delay and path length. On the other hand, the network behavior for mobile ad hoc networks is not well understood. There are no systematic studies on the correlations between various performance metrics and system parameters such as node mobility and offered load. We show in this paper by interpreting the simulation results in an alternate graphical representation, interesting relationships are revealed.

Our graphical interpretation of simulation results is similar to that used by [26]. The performance metrics of individual *flows* (source destination pair) are plotted. Whereas [26] use this graphical representation to compare metrics on individual flows, we extend the use of these graphs to investigate the correlations between various metrics and system parameters. This has led to some new observations not reported before, yielding insight to the inter-relationships between the goodput, path length and node mobility and offered load. For example, by studying the relationships of path length and goodput to speed simultaneously, we resolve a conjecture that goodput improvement under high mobility is due to the load balancing effect. Furthermore, we have introduced the concept of *fraction of congested flows* as a new performance metric. This and some other metrics such as *fairness* could be easily visualized from our graphs and are important in characterizing network performance.

Our data interpretation described in this paper is general. Many inter-relationships between performance metrics and system parameters could be obtained. In this paper, however, we focus on the inter-relationships between goodput, path length and the system parameters only. The goodput (packet delivery ratio) $G$ denotes the fraction of packets that is correctly received. The path length $L$ denotes the number of hops a packet travels along a route. Inter-relationships on packet delay is not discussed. The rest of the paper is organized as follows. In section 2, we discuss the inadequacies of using ensemble averaging in obtaining the performance metrics. A graphical representation similar to that used by [26] is described. Section 3 describes our simulation setup and our main observations are discussed in section 4.

## 6.2 Ensemble Averaging in Performance Metrics

In the routing literature [8, 12], the performance metrics of a network are obtained using ensemble averaging. For a given node mobility, a number of mobility scenarios are created. A number of flows in each scenario is monitored for some duration. We obtain the performance metrics by averaging over all the monitored flows. The metrics are usually plotted against system parameters such as mobility [8,12] or offered load [12].

Although an average performance metric is useful in ranking the performance of different routing algorithms, it tells little about the performance metric of individual flows in each scenario. In different mobility realizations of the network, node placement and the traffic patterns are different. The aggregate network performance in each scenario varies. Even within a mobility scenario, individual flows also exhibit heterogeneous performance characteristics due to different path lengths and the non-uniform spatial distribution of the offered load in the network. Thus when the ensemble average of a performance metric is plotted, the variation of individual flows within the network is not captured.

In [8, 12], the performance metrics are plotted against mobility. Each point on a graph indicates the average performance metric for a specific value of node mobility. Adjacent points are obtained from mobility scenarios with different node mobility. Since these metrics are averaged over flows that have very different characteristics, the graphs in [8, 12] show zigzag patterns. The large variations in performance over different flows undermine the validity of the average metric. The trends of the performance metrics with system parameters are not obvious. This hinders our objective to find the inter-relationships between the performance metrics and system parameters.

An average performance metric sometimes yields a misleading conclusion too. A classical example is when the average packet delay is plotted. In a typical ensemble of monitored flows, some flows may have lousy routes and many packets are dropped. Since these flows have higher packet delay in general, the average packet delay will be artificially smaller since few packets from these flows reach the destination node. In the comparison of two routing algorithms, if one algorithm drops a lot of packets from lousy routes, the average delay is then artificially smaller, rendering the comparison between two algorithms meaningless.

In this paper, we advocate an alternate graphical representation of simulation results that is similar to the approach of [26]. In [26], the throughput for each of 50 mobility patterns for the 20m/s and 30m/s mean speeds used in the simulations are plotted. The patterns are sorted in the order of throughputs at 20m/s. It is demonstrated that for some mobility patterns the goodput improves with mobility. We use similar graphs

in our data interpretation with some modifications. We assign a *pattern number* to each monitored flow. The pattern numbers are assigned such that the flows are ordered in the order of increasing goodput, path length, or delay. We plot the performance metric of each monitored flow versus the pattern number. Under the above assumptions, the throughput plot for each of the 50 mobility patterns for the 30m/s mean speed is different. The same pattern number no longer refers to the same flow in each of the two mean speeds. Since the throughput plot for each mean speed is an increasing function, it is obvious to observe the inter-relationship of throughput versus mobility.

## 6.3 Simulation Setup

The simulations are performed on *ns-2* [1], with its wireless extensions developed by the Monarch project [2]. The simulations consist of 50 mobile nodes that are distributed uniformly in a 1500m by 300m area. The propagation model consists of a simple path loss model with attenuation due to distance only. The path loss exponent is chosen to be $\beta = 4$. The default parameters of the wireless radios are used, such that each node has a transmit range of $250m$.

The mobile nodes emulate 914MHz Lucent WaveLAN DSSS radio interfaces. The transmission bandwidth is 11 MHz, and the nominal bit rate is 2 Mbps. Omni-directional antennas with 0dB gain are used, and antennas are placed 1.5m above the ground. The receive threshold is -64.37 3Bm, which determines the minimum SIR required for successful decoding of a received packet. The carrier sense threshold is -78.07 dBm. Any packet with a SIR more than the threshold may interfere with reception of another packet.

Nodes move in the network under the random waypoint mobility model. We characterize the mobility using the parameter *max_speed* and keep the *pause time* equal to 1 second in all mobility scenarios. Each node has a velocity that is uniformly distributed between 0 and *max_speed*. Four different values of *max_speed* are investigated in this numerical study, namely $v = 0, 2, 10, 20$ m/s. These values correspond to the stationary, pedestrian, slow and fast vehicular scenarios.

The traffic is generated through a CBR application over UDP [8,12]. This simulates the routing performance of the best effort delivery paradigm. The offered load could be varied by any of the three parameters, namely packet transmission rate, packet size and the number of traffic flows in the network. We simulate four offered load regimes, as shown in Table 6.1. The traffic types 1 to 4 correspond to the network operating in the light, normal, heavy, and saturation load regimes respectively. Packet sizes are chosen such that fragmentation occurs on neither the network nor the MAC layer. The number of flows are kept constant for each traffic type. A fixed fraction of flows from each scenario are taken for measurement. This allows uniform sampling among all mobility scenarios without bias for a particular scenario realization.

In this paper we use the *dynamic source routing* (DSR) routing algorithm [33] in our simulations, since DSR shares many of the salient characteristics typical to reactive routing algorithms. The DSR runs on top of the 802.11b standard with a channel reservation mechanism enabled by the use of *request-to-send* (RTS) and *clear-to-send* packets. In general, packet loss can result from contention in wireless transmissions, unavailability of route due to mobility, or buffer overflow due to congestion. Nevertheless, the RTS/CTS mechanism in the 802.11b standard is efficient in combating the hidden terminal problem. Thus, most packet loss are due to mobility or congestion. In the light traffic regime, we factor out the congestion effect of all routing protocols. The performance sensitivity of routing protocols to mobility is investigated. In normal traffic regime, realistic traffic scenarios of practical interest are simulated. The simulation results give us performance estimates to realistic network scenarios. In heavy traffic load and network saturation regime, delay is unbounded. Although it is undesirable to operate a network in these regimes, saturation throughput determines the capacity of the network. Thus it is instructive to perform experiments under all four traffic scenarios.

We have selected the network size such that network partitioning does not occur in any mobility scenario. Nevertheless, the DSR routing algorithm may fail to discover a route in heavy traffic regimes or high mobility scenarios. By convention, if there is no connection for a flow for the whole simulation, we define the goodput to be 0, and the

| Traffic type | packet rate | packet size | number of flows | total load |
|---|---|---|---|---|
| 1 | 5 | 64Byte | 20 | 51.2Kbps |
| 2 | 10 | 64Byte | 20 | 102.4Kbps |
| 3 | 10 | 512Byte | 20 | 819.2Kbps |
| 4 | 20 | 768Byte | 20 | 2.458Mbps |

Table 6.1: Traffic parameters adopted in the numerical studies

delay and path length to be infinity.

Altogether we have four mobility scenarios and four offered load regimes. For each of the sixteen *network scenarios*, ten topology realizations are simulated. Each simulation lasts for $300s$. Each flow starts at a staggered time that is uniformly distributed between 0 and $100s$. Simulation data is logged during the interval between $100s$ and $300s$ to ensure the network has reached to a steady state. In each topology realization, 5 out of the 20 flows are monitored. Thus, for each network scenario, we have logged the data of 50 monitored flows.

## 6.4 Simulation Results

### 6.4.1 Dependence of Path Length $L$ on Speed

The time averaged path length $L$ of each flow of each offered load regime is plotted in Figure 6.1. In each subgraph, the path length in each mobility scenario is plotted. The average path length of a flow is obtained by averaging the number of hops each packet traverses in the flow during the experiment. In the heavy traffic regimes of Figure 6.1(c)(d), we consistently observe that the path length decreases as mobility increases for each pattern number. The path length difference is as large as 4 hops for some pattern numbers. This explains the prevalence of short routes in high mobility scenarios of the heavy traffic regimes. A similar trend is observed in the non heavy traffic regimes of Figure 6.1(a)(b). The relative difference in path length in different mobility scenarios is smaller. In particular, for short path lengths (1-2 hops) the trend is reversed and mobility increases the path length slightly.

In the DSR protocol, each node continuously snoops into every packet it receives. The length of an existing route may be *shortened* or *lengthened* as time evolves due to mobility. Whenever a node along a route detects there is a shorter path to the destination node, a route change is triggered. Thus under higher mobility, route optimization is triggered more often, leading to shorter routes. However, for flows with a short path length of 1 or 2 hops, mobility usually *lengthens* a route. This explains that pattern 1 to 23 in Figure 6.1(a)(b) have longer routes in high mobility. In general, the path length of long routes (path length $L \geq 3$) decreases in mobility, whereas the path length of short routes increases in mobility. The variations of path length in high mobility scenarios is thus smaller.

In the heavy traffic regimes, there is a larger disparity of path length in different mobility scenarios. In particular we observe from Figure 6.1(c)(d) that the path length of 80% of the monitored flows is less than 3 hops at speed 20m/s. When the network is under the stress of *heavy traffic* and *high mobility*, only short routes (1 to 2 hops) are discovered during route discovery. In these regimes, the packet delay incurred at each hop is in the order of 10 seconds. In route discovery, if a *route request* (RREQ) packet traverses along a long route, the round trip delay is sufficiently long such that route discovery is aborted. Thus, in high mobility scenarios, the source and destination nodes are intermittently connected. When the nodes are in proximity, route discovery is successful; otherwise, route discovery fails. This explains the prevalence of short routes in high mobility scenarios of the heavy traffic regimes.

Incidentally, in the stationary scenario of the light offered regime of Figure 6.1(a), we observe that the time averaged path length of each flow is an integer. In general, route changes are due to mobility or congestion. Since there is no mobility and congestion in the stationary scenario of Figure 6.1(a), there are no route changes over the duration of simulation. Thus the time averaged path length of individual flow must be an integer. A staircase pattern is also observed in the stationary scenario of the normal offered load regime. In Figure 6.1(b), we observe that for pattern 1 to 35, the path length follows a staircase pattern. This shows that there are few route changes when the path length are short. For patterns 36 onwards, the staircase pattern disappears. This indicates

that route changes due to congestion are common for these flows. We observe that these flows have path lengths of more than 4 hops. Thus we could also infer that flows with longer path lengths are more susceptible to route changes due to congestion.

### 6.4.2 Dependence of Path Length $L$ on Node Distribution

Referring to Figure 6.1(a) the stationary scenario of the light offered load regime, there are 14 flows with a path length of 1 hop. As the path length $L$ increases, the number of flows with path length $L$ decreases. In this scenario, there is neither congestion nor route changes due to mobility. The proliferation of routes with short path length is therefore not related to congestion nor node mobility. Consider a scenario in which nodes are uniformly distributed on a line of length 1. Denote $X$ and $Y$ as the location of two arbitrary nodes, such that $X \in [0,1], Y \in [0,1]$. We define the distance between the two nodes as $Z = |X - Y|$, where again we have $Z \in [0,1]$. To compute the probability distribution of $Z$, we have

$$\begin{aligned} & Pr[Z \leq z] && (6.1)\\ = \;& Pr[|X - Y| \leq z] && (6.2)\\ = \;& Pr[Y - z \leq X \leq Y + z] && (6.3) \end{aligned}$$

Since $X$ and $Y$ are independent, the event $Pr[Z \leq z]$ corresponds to the shaded area in Figure 6.2, implying

$$\begin{aligned} & Pr[Z \leq z] && (6.4)\\ = \;& 1 - 2(1/2)(1 - z^2) && (6.5)\\ = \;& 2z - z^2 \qquad 0 \leq z \leq 1 && (6.6) \end{aligned}$$

and the corresponding PDF is

$$f_Z(z) = \begin{cases} 2(1-z) & 0 \leq z \leq 1 \\ 0 & \text{o.w.} \end{cases} \tag{6.7}$$

(6.7) indicates that the prevalence of short routes is a direct consequence of uniform node distributions in the network. Since path length is roughly proportional to route

distance, a route with short path length is more probable. The same trend of path length is also observed in other mobility scenarios and offered load regimes since all the path length curves $L$ in Figure 6.1 are concave upwards.

The mean path length $E[Z]$ could be derived from Equation (6.7) to be 1/3, which is one third of the network dimension. In our simulation, we have used a scenario size of 1500m $\times$ 300m. Since each node has a nominal range of 250m, the scenario resembles a one dimensional network. The mean route length is thus 500m. Consider the stationary scenario in the light offered load regime. The mean path length of all 50 flows is found to be 2.95 hops. This agrees with our computations for a one dimension network. Most routes require a minimum path length of 3 hops to traverse a distance of 500m.

### 6.4.3 Improved Goodput $G$ due to Load Balancing

In Figure 6.3, the goodput $G$ of each offered load regime is plotted. In each subgraph, the goodput in all mobility scenarios is plotted versus pattern number. In the light offered load regime of Figure 6.3(a), we observe that mobility leads to a slight deterioration of goodput. Due to the light offered load, few packets in transit are lost due to buffer overflow in the node preceding a broken link. Most packets are queued in the node buffers during route maintenance. Although goodput degrades with mobility, the discrepancy is small because packet loss is uncommon.

Similarly, the goodput also deteriorates in the vehicular scenarios of the normal offered load regime of Figure 6.3(b). At normal load, packet loss is due to both mobility and congestion. During route failure, most packets in transit are lost due to buffer overflow. Thus, goodput is very sensitive to mobility. However, the goodput in the pedestrian scenario is higher than that in the stationary scenario. When nodes are stationary, packet loss is due to local congestion along some flows. For convenience, we define a flow as congested if the goodput of the flow is smaller than 0.8. Thus there are 9 congested flows in the stationary scenario, compared with 5 for the pedestrian scenario. This is consistent with observations in literature [8, 12], in which goodput is shown to improve with speed. It is conjectured that the improved goodput in mobility is due to the load balancing effect. Some flows that pass through the congestion hot

spots achieve improved goodput through rerouting brought about by node mobility.

### 6.4.4 Improved Goodput $G$ due to Reduced Effective Load

The load balancing effect could correctly explain the goodput improvement with speed when localized congestion occurs. When we consider the heavy traffic regimes, the load balancing argument is inapplicable since network-wide congestion is experienced at all nodes. However, the goodput improvement with speed is still observed, albeit for a different network mechanism.

In the heavy traffic regimes of Figure 6.3(c)(d), we observe that many flows have better goodput in higher mobility. This is remarkably different from other traffic regimes of Figure 6.3(a)(b), where mobility degrades the goodput performance.

In heavy traffic, congestion is a network-wide phenomenon. All nodes are backlogged with packets. Rerouting due to mobility should not bring about any goodput improvement at all. We claim that the improved goodput in mobility is due to the decreased effective network load. As discussed earlier, most flows have shorter routes in high mobility. The total number of transmissions to forward a packet to the destination node is dramatically reduced. This effectively decreases the total network traffic, enabling a higher goodput for all flows. To see this, we consider the network saturation regime. We compare the total number of transmissions to deliver one packet for each monitored flow in each mobility scenario. Refer to Figure 6.1(d), by computing the area under the path length paths in Figure 6.1(d), we find the total number of transmissions to deliver 1 packet for each of the 50 flows is 171.9184 hops in the stationary scenario. The average path length for each flow is then 3.4384 hops. In the fast vehicular scenario, the total number transmission for 50 flows is 93.3077 hops. Thus the average path length for each flow is 1.8662 hops. Compared the normalized path length of each flow, the effective network load in the fast vehicular scenario is only 54.27% of that in the stationary scenario.

The corresponding number of delivered packets in a fixed duration is proportional to the sum of the areas under the goodput graphs in Figure 6.3(d) weighed by the packet rate. Suppose $T_{intarr}$ is the mean packet interarrival time. In the stationary

scenario, the expected number of delivered packets for 50 flows is 6.0541 packets. The normalized number of delivered packets for a flow in a time $T_{intarr}$ is thus 0.1211. In the network saturation scenario, the expected number of delivered packets for 50 flows is 8.6093 packets. The normalized number of delivered packets for a flow in a time $T_{intarr}$ is thus 0.1722. This amounts to a 42.21% increase in goodput.

Similarly, we also compare the total number of transmissions to deliver one packet for each monitored flow in the heavy offered load regime. Refer to Figure 6.1(c), the average number of transmissions is 3.6 hops per flow in the stationary scenario, and 2.0986 hops per flow in the fast vehicular scenario. The effective load in the fast vehicular scenario is only 58.29% of that in the stationary scenario. From the goodput graph in Figure 6.3(c), the expected number of delivered packets for a flow in a time $T_{intarr}$ is 0.2880 in the stationary scenario. compared with 0.3277 in the fast vehicular scenario. This amounts to a modest 13.79% increase in goodput. In general, in high mobility scenarios, the reduction of effective network load has a more prominent effect to packet loss due to mobility. Thus we expect to obtain even better network goodput in higher mobility scenarios.

### 6.4.5 Determination of the Fraction of Congested Flows

In Figure 6.4, the goodput $G$ of all mobility scenarios is plotted. In each subgraph, the goodput in all offered load regimes are plotted against the pattern number. We note that Figure 6.4 and Figure 6.3 are plotted from the same results. Whereas the goodput in different mobility scenarios are compared in Figure 6.3, goodput in different offered load regimes are compared here in Figure 6.4.

Consider the stationary scenario in Figure 6.4(a). Recall that we defined a flow as congested if the goodput was less than 0.8. Thus, in the normal offered load regime, roughly 20% of all flows experience congestion. In the heavy offered load and the network saturation regimes, the fraction of congested flows is respectively 80% and 100%. Nevertheless, the above definition is arbitrary. We show below that it is possible to classify the flows into the congested or uncongested regimes independent of a specific threshold.

Consider the light offered load regime of Figure 6.4. In this regime, packet loss is due to mobility in every flow. It is clear that the goodput curves for all patterns could be fitted by a straight line as a function of pattern number. Thus the variations of goodput could be modeled by an uniform distribution with some mean and variance. As mobility increases from stationary to pedestrian to vehicular scenarios, the mean goodput drops slightly and the variance increases.

More generally, packet loss is due to a combination of mobility and congestion. Consider the normal offered load regime in the following. In the slow vehicular scenario of Figure 6.4(c), the goodput curve could be fitted into a piecewise linear function. From patterns 1 to 12, the goodput variations with pattern number have a steeper slope. For other patterns, the goodput variations follow a more gentle slope. The observation of two regimes could be explained by the cause of packet loss. For patterns 11 to 50, packet loss is dominated by mobility. Similar to the light offered load regime, the goodput variations of each flow could be modeled by an uniform distribution. For patterns 1 to 10, congestion dominates over mobility. Congestion is more detrimental to goodput than mobility. Also, there are more variations in goodput depending on the extent of congestion. The goodput for the congested flows is thus modeled by an uniform distribution with a smaller mean and larger variance. Similarly, the goodput variations in the fast vehicular scenario of Figure 6.4(d) could also be fitted by a piecewise linear function. By observing the intersection of the fitted lines, we infer that patterns 1 to 11 is in the regime where congestion is the main reason for packet loss.

In the heavy offered load and the network saturation regimes, the goodput could not be fitted nicely by a piecewise linear function. In these regimes, congestion is a network-wide phenomenon. Thus there are flows in which goodput deterioration due to mobility and congestion is both prominent.

In our simulations, the delay in uncongested flows typically does not exceed 1 second. The corresponding delay in congested flows is in the order of 10 seconds. Most applications in ad hoc network today have tight delay constraints. It is highly undesirable to deliver packets over some flows that experience congestion. The fraction of congested flows that is therefore an important performance metric to consider.

### 6.4.6 Dependence of Fairness on Offered Load, Speed and Path Length

In homogeneous ad hoc networks, all nodes are peers and they cooperate to forward packets for each other. Applications in military and rescue operations fall into this context. In these networks, it is desirable for each flow to attain the same goodput independent of the offered load, mobility and path length.

Referring to Figure 6.4, we observe that fairness deteriorates quickly with the offered load in each mobility scenario. When the offered load is light, all flows have a goodput close to 1 irrespective of mobility. In the normal offered load regime, local congestion occurs for some flows. We discussed earlier that the variable goodput could be modeled by a uniform distribution with some mean and variance depending on the cause of packet loss. If packet loss is dominated by congestion, the goodput exhibits randomness with a larger variance. Thus, in general, fairness is very sensitive to the presence of congestion. In particular, when congestion dominates in the heavy traffic regimes, there is a high asymmetry in the goodput of the monitored flows.

We also observe the sensitivity of fairness to offered load decreases in higher mobility. This is easily visualized by comparing the stationary and the fast vehicular scenarios in Figure 6.4(a)(d). At high mobility the goodput curves of different load regimes are more closely spaced. Heavy load regimes have improved goodput due to the decreased effective network load while light load regimes suffer from packet loss due to mobility.

Whereas fairness is sensitive to the offered load, it is insensitive to mobility. Refer to Figure 6.3(a), all flows have a goodput close to 1 in all mobility scenarios. The goodput is insensitive to mobility in the light offered load regime. Similarly, in the normal offered load regime of Figure 6.3(b), packet loss due to mobility leads to a slight deterioration of goodput. In the heavy traffic regimes, the reduction of effective network load in high mobility scenarios leads to a slight improvement in goodput. Thus, for each offered load regime, the disparity of goodput among all flows in different mobility scenarios is small. This is intuitively plausible since fairness depends on the resource allocation of each flow in the network. Although mobility impacts the efficiency of route discovery, the resource allocation of each flow does not depend on mobility once a route

is found. Thus, in all node mobility of interest in this study, mobility does not impact the fairness. In extremely high mobility, however, route discovery is inefficient. The flow will be intermittently connected, leading to poor goodput performance.

We have argued that fairness is intimately connected to the resource allocation in each flow. Since the path length of a route determines the amount of forwarding and thus the resource requirement of a flow, it is instructive to investigate the dependence of fairness to path length $L$. Refer to Figure 6.5, the goodput $G$ for each mobility scenario is plotted. In each subgraph, the goodput of all patterns in all offered load regimes are plotted ordered in increasing path length. This allows us to inspect the effect of path length $L$ on fairness in each offered load regime and mobility scenario.

As shown in the figure for all mobility scenarios, we observe some correlation between the path length $L$ and goodput $G$. When the path length is long, the goodput is likely to be smaller. Consider the light offered load regime. Although routes with long path lengths have lower goodput, fairness is not a problem in this regime since the goodput of the worst flow in each mobility scenario is close to 1. Consider the normal offered load regime. In the stationary and pedestrian scenarios of Figure 6.5(a)(b), only the flows with long path lengths have small goodput. In these scenarios, packet loss is dominated by congestion. We could infer that long routes are more susceptible to congestion. This is intuitively reasonable since it is more likely to route through some local congestion hot spots if the path length is long. In the slow and fast vehicular scenarios, we also observe that goodput is likely to be smaller for longer path lengths. However, not all flows with long path lengths have small goodput. This could be explained by load balancing due to path rerouting for these flows.

Consider the heavy offered load and network saturation regimes. In these regimes, long routes are shut down completely. With reference to Figure 6.5(a)(b), 40% of all flows with the longest path length have a goodput close to 0. In the slow and fast vehicular scenarios the fraction of flows that are shut down decreases to 20% and 12% respectively due to the decreased effective network offered load. In general, flows with long path length are shut down completely in heavy traffic regimes. Only local communication is possible. We conclude that fairness is sensitive to the path length

when the offered load is large.

### 6.4.7 Dependence of Path Length $L$ on Offered Load

In Figure 6.6, the path length $L$ of all mobility scenarios is plotted. In each subgraph, the path length of all patterns in all offered load regimes are plotted. These are the same results of Figure 6.1. Path length in different offered load regimes are compared here, whereas path length in different mobility scenarios are compared in Figure 6.1.

Consider the stationary and pedestrian scenarios of Figure 6.6(a)(b). We observe that $L$ generally increases when the offered load increases. This phenomenon implies that at high traffic intensity, either routes with long range could not be used, or successful transmissions are limited to small ranges only. The latter argument is flawed since the RTS/CTS channel reservation mechanism in 802.11 is effective in resolving contention, even if the network is under the stress of excessive traffic. The paradox could be explained as follows. The RTS/CTS mechanism in 802.11 effectively prevents collisions of data packets. Route request (RREQ) packets however, can't use the RTS and CTS mechanisms since they are broadcast packets. If heavy congestion occurs, collisions between RREQ packets are likely. Since collisions are more likely to occur for longer hops, RREQ packets may never reach the destination node if a route consists of hops with large distance progress. Only routes with large number of hops and small per hop distance progress are discovered in route discovery. In general, as the offered load increases, congestion is prominent and may impede the transmissions of the broadcast RTS and CTS packets. Thus path lengths are longer due to congestion.

Consider the slow and fast vehicular scenarios of Figure 6.6(c)(d). At high mobility, the trend of our observations is reversed. We observe as offered load increases, the path length decreases. In high mobility scenarios, mobility has a more significant impact compared to congestion. Recall in our discussion for path length $L$ vs. speed. At high mobility, the path length will decrease more drastically for heavy load regimes. In general, congestion and mobility have opposite effects on path length. At high mobility, the effect of mobility dominates. Therefore, path length is smaller for higher offered load.

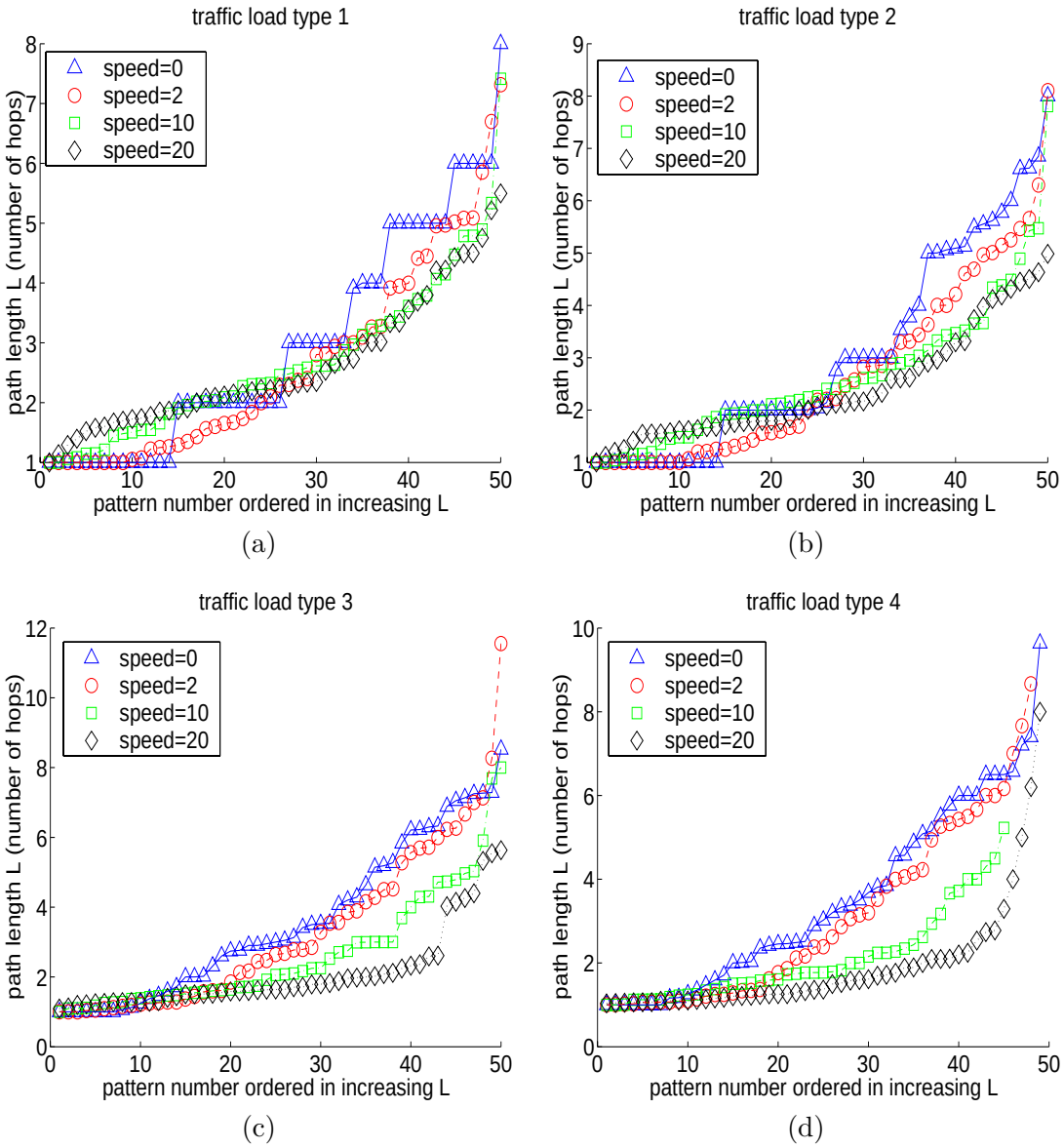



Figure 6.1: Path length vs. pattern number in all mobility scenarios. (a)light offered load regime, (b)normal offered load regime, (c)heavy offered load regime, (d)network saturation regime.

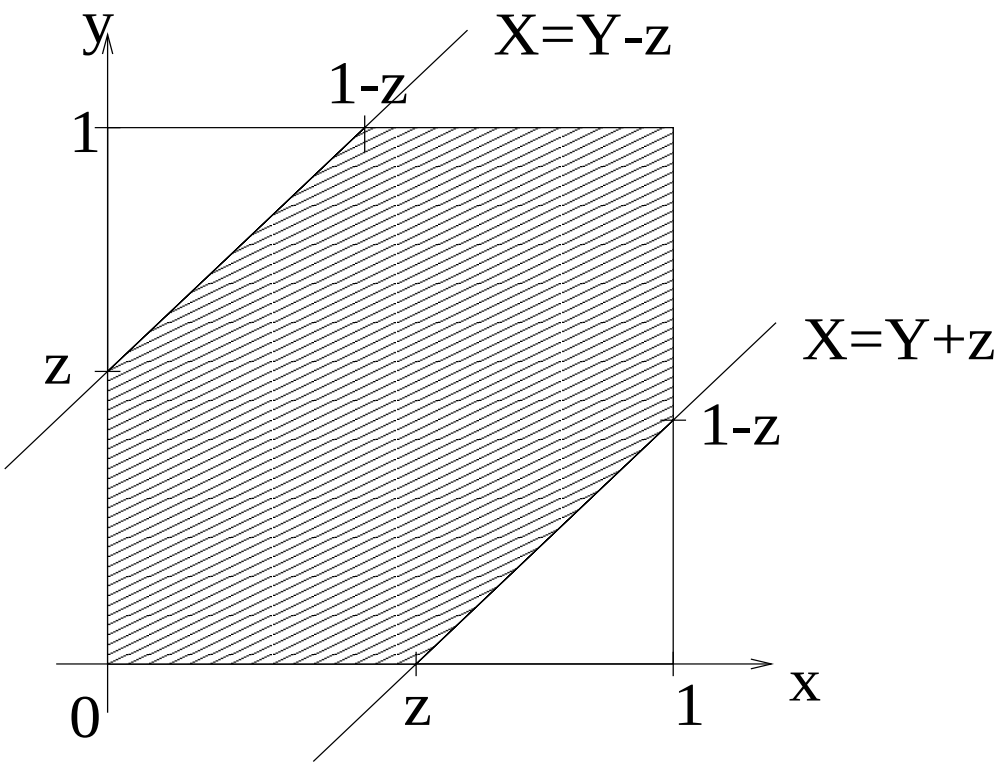



Figure 6.2: Illustration for computing $Pr[|X - Y| \leq z]$

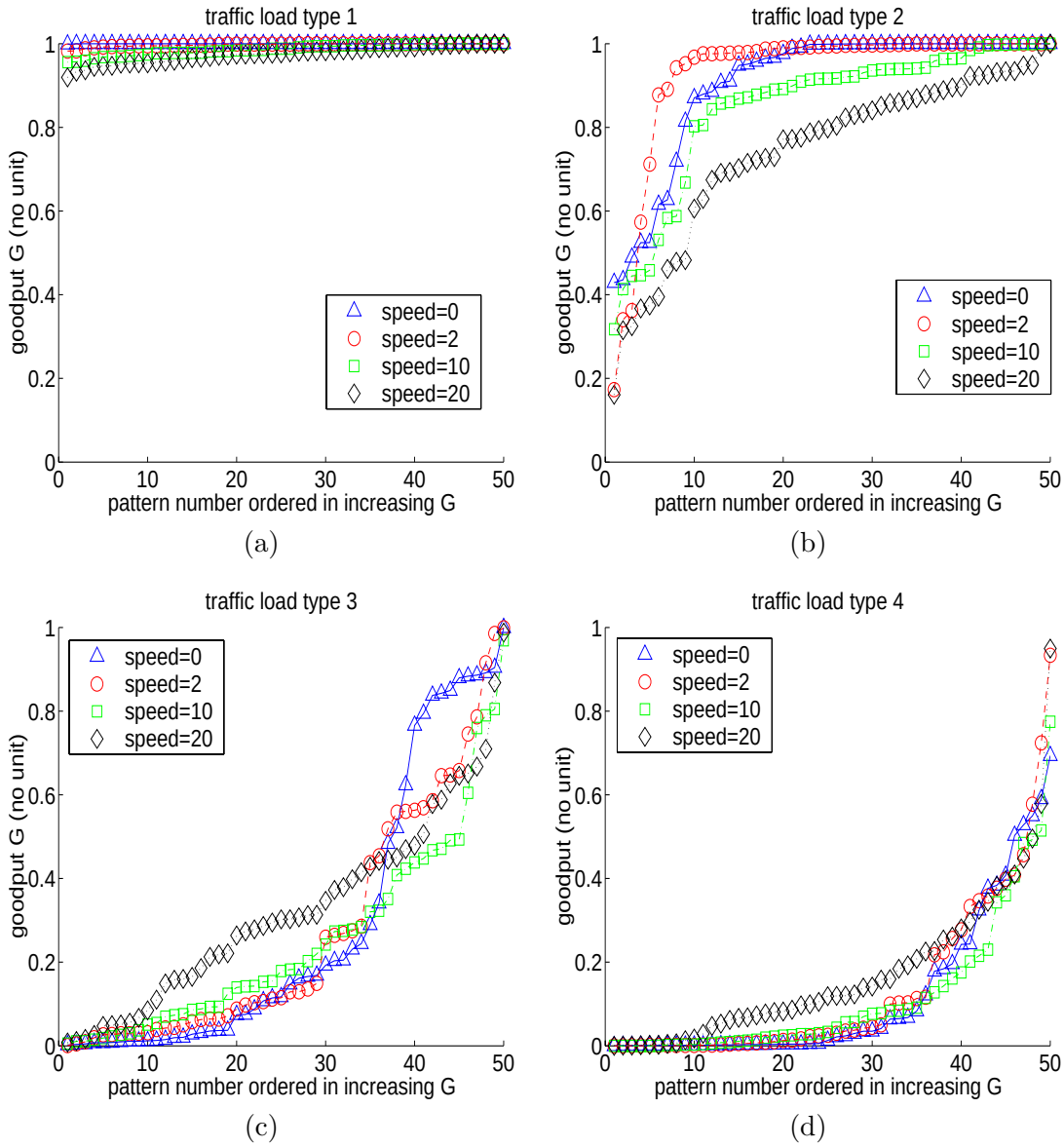



Figure 6.3: Goodput vs. pattern number in all mobility scenarios. (a)light offered load regime, (b)normal offered load regime, (c) heavy offered load regime, (d)network saturation regime.

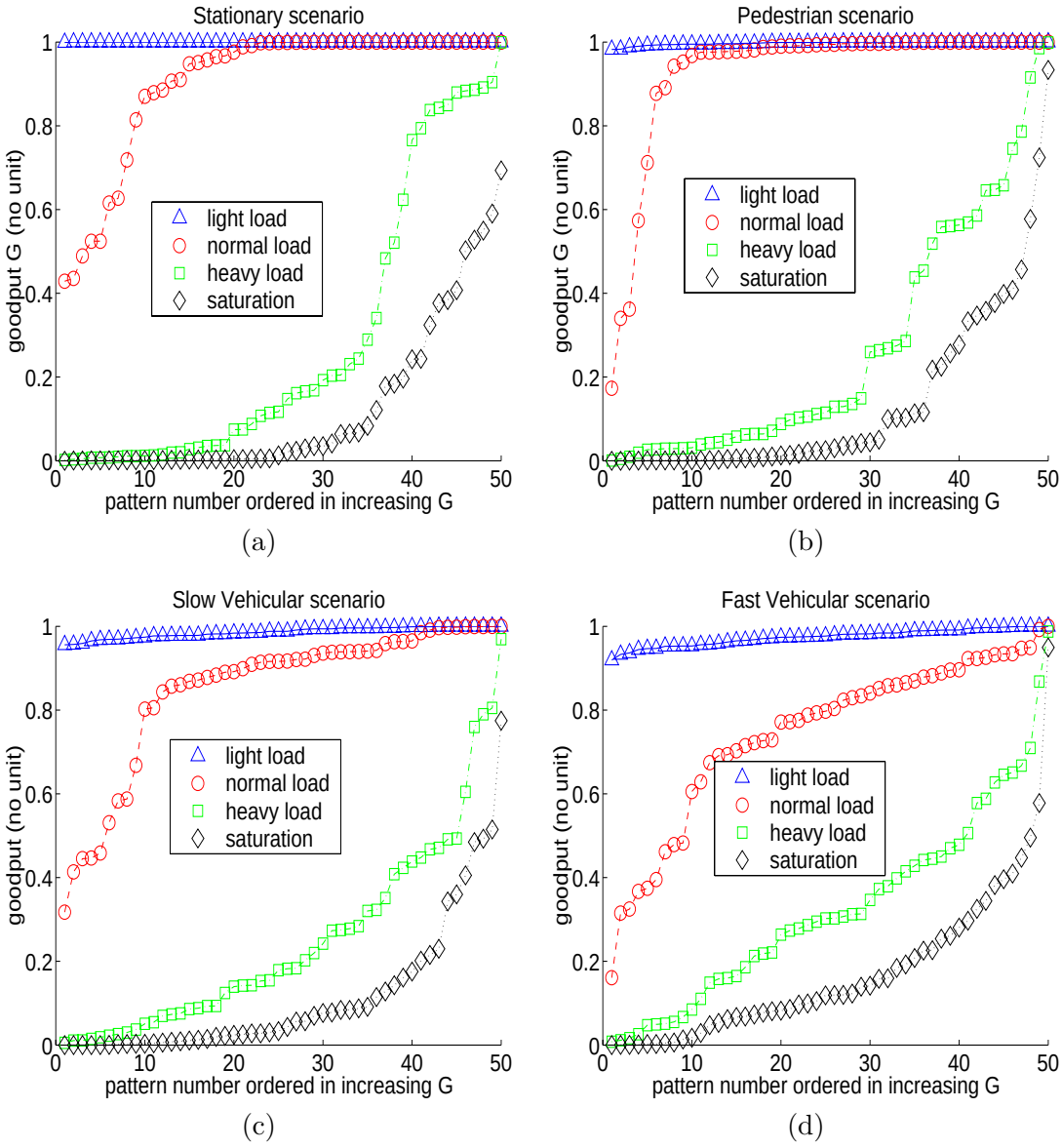



Figure 6.4: Goodput vs. pattern number in all offered load regimes. (a)stationary scenario, (b)pedestrian scenario, (c)slow vehicular scenario, (d)fast vehicular scenario.

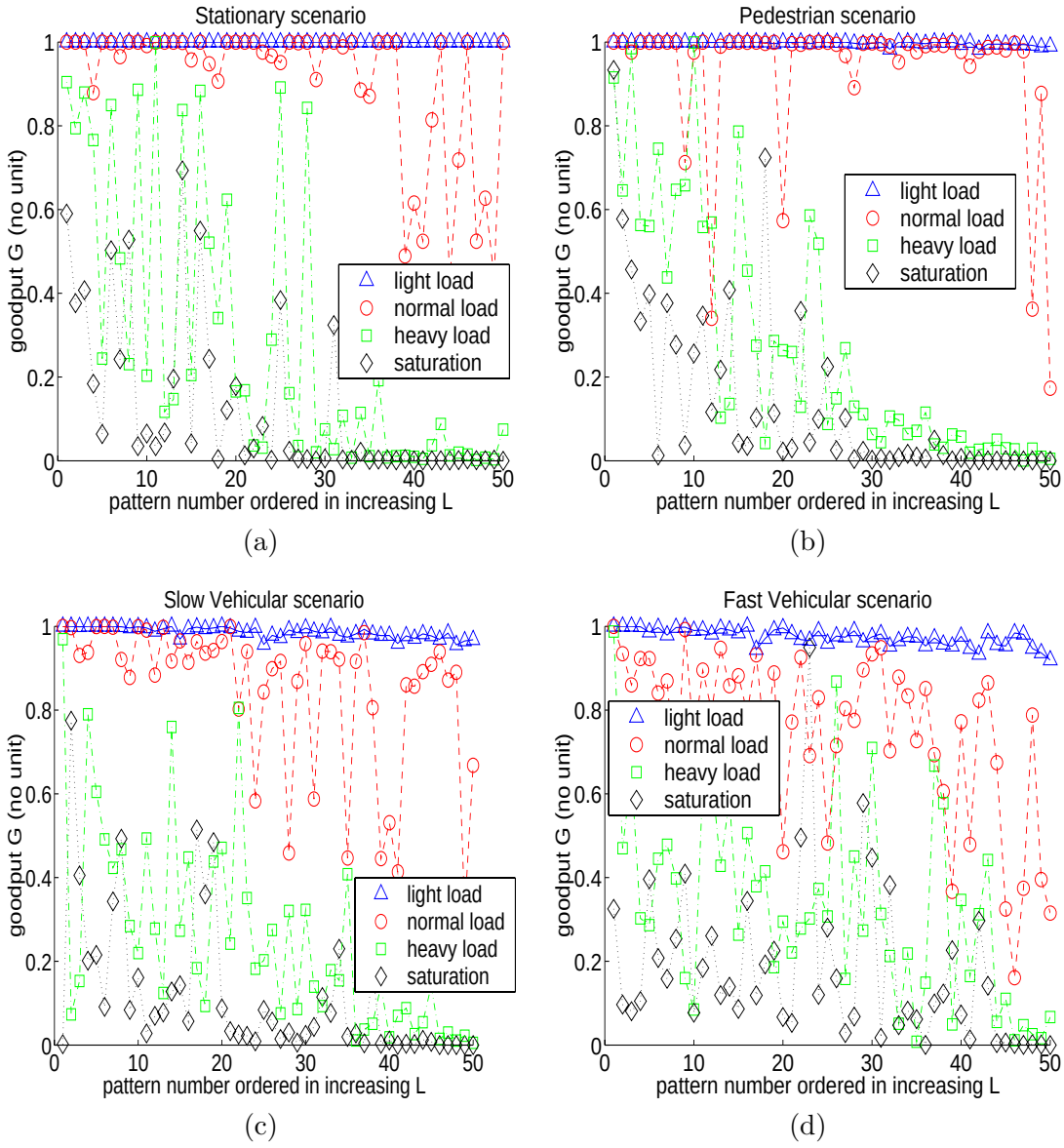



Figure 6.5: Goodput vs. pattern number in all offered load regimes. (a)stationary scenario, (b)pedestrian scenario, (c)slow vehicular scenario, (d)fast vehicular scenario.

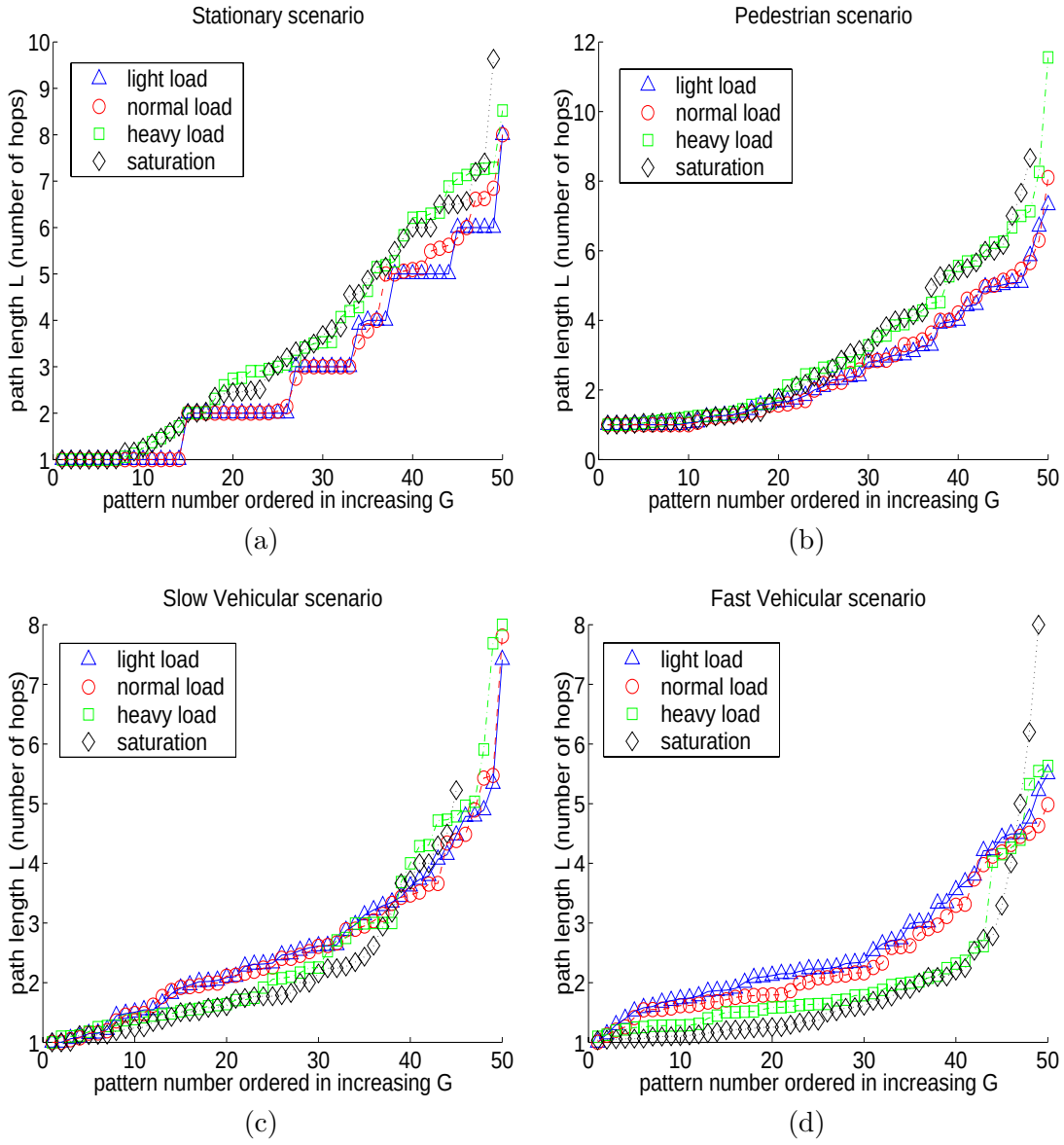



Figure 6.6: Path length vs. pattern number in all traffic regimes. (a)stationary scenario, (b)pedestrian scenario, (c)slow vehicular scenario, (d)fast vehicular scenario.

# Chapter 7

# Optimal Price Decremental Strategy for Dutch Auctions

## 7.1 Introduction

In the previous chapters, we have examined various issues that affect network behavior. A thorough understanding of these issues is important so that the mobile ad hoc network could be operated at the optimal regimes. Nevertheless, the success of a networking paradigm eventually lies in the pervasiveness and the importance of the applications running on top of the network stack. In this chapter, we look into the example of an online wireless Dutch auction application for 3G/4G mobile cellular networks. We focus on mobile cellular networks rather than mobile ad hoc networks, since the former is a mature technology with a wide subsriber base. There is a lot of commercial interest on finding a killer application that generates extra revenues for the network operator.

In a Dutch auction, the price of an item decrements from the starting price at regular intervals. A bidder may buy the item at any time and stop the auction at the current price. This chapter presents an optimal price decrement strategy in a Dutch auction, such that the expected revenue of the auction host is maximized. Properties of the optimal solution and a simple iterative solution methodology are discussed. Numerical studies show that significant gain could be obtained compared with a simple reference strategy.

## 7.2 Online Dutch Auction

With recent advances in wireless standards, such as the IMT2000 for cellular networks as well as the HIPERLAN and the IEEE802.11 standards for wireless local area networks,

there has been great expectation that wireless data applications will soon become popular just like wireless telephony. In anticipation of this development, there have been many attempts in testing pilot applications on various wireless platforms. The Information Engineering Department at the Chinese University of Hong Kong has established a site, jawap.net, based on the Wireless Application Protocol(WAP) to provide a host of wireless applications, including an implementation of an Dutch Auction System.

The Dutch auction is said to originate in the Netherlands and uses a descending-price format unlike the so-called "English Auction". In the Dutch scheme, when an object is presented to interested buyers for bidding the price will start at a high value and progressively decreases downward until a buyer bids for the object by making a declaration. If multiple declarations are made, any common resolution scheme can be invoked to break the tie. It is possible to extend this scheme for the auctioning of multiple units of an object. Successful bidders bidding at the same price will each receive their unit at the bid price. If the number of units is not sufficient to cover all the bids, a tie-breaking rule is invoked. If sufficient units of the object are still available, the auction will continue until all the units are sold or a reserved price level is reached.

The Dutch auction is one of the four major auctioning schemes ( [3] or the survey article [46],) that also include the English auction[1], first price sealed bid auction[2], and second price sealed bid auction (also known as the Vickrey auction)[3]. The seminal work of Vickrey [95] analyzed and compared these four common kinds of auction rules. Since the process of an auction is very complicated involving the auctioneer, the seller, and multiple bidders, it is common to make certain simplifying assumptions about these players in order to make the analytical model tractable. A key concept concerns the idea of the value of the object under auction. The value can be *private*, that is, a person buys an item for his/her own consumption without an objective to resell, or *common*, in

[1]In the English auction, bidders compete with each other by offering progressively higher bids until one bidder remains, who is committed to buy at last price of the last bid.

[2]In a sealed bid auction, all the buyers submit a bid at the same time. The buyer offering the highest bid is committed to buy the object the bid price.

[3]In such an auction, the buyer making the highest bid is committed to buy the object, but at a price equal to the second highest bid.

which case the buyer bids with the intention of resell and has to estimate the valuation offered by prospective buyers. In the latter case, the competitors are clearly a helpful source to obtain such a valuation estimate. A bidder is said to be *risk neutral* if he/she bids exactly accordingly to his/her evaluation of the object. A bidder who is likely to bid above his/her evaluation to increase the chance of winning is called a *risk averse* bidder.

Vickrey's paper [95] assumes that each bidder is risk-neutral and knows the value of the object to himself/herself but not the value to other bidders. Moreover, the model is assumed to consist of *symmetric* bidders, that is, individual *valuation* of the object is i.i.d. Since then, there has been a steady output of follow-up work on optimal auction design. One important piece of work was due to Myerson [55]. He extended the work of Vickrey in two directions. Firstly, the case of *asymmetric* bidders is considered, in which individual valuations are independent but not necessarily identically distributed. Secondly, different viable ways of selling the object was considered rather than just a prespecified set of auction rules. Under this framework, the optimal auction design problem to maximize the expected revenue of the auction host is solved. There have been many more additional works on the design of optimal auction rules. The well known auction rules are compared under various relaxations of the stated assumptions [52], [74], [57].

With the advent of the World Wide Web, online auctions have become increasingly popular. Moreover, the arrival of 3G and high speed wireless local area networks have made the idea of hosting auctions to serve mobile users via wireless communication devices, such as enhanced cellular phones or personal digital assistants, practical in the near future. Coupled with the concept of *micropayment*, one can envision the possibility of auctioning all sorts of items which may have only small monetary value or are time-critical, such as tickets for an upcoming concert or seats on air flights.

For these applications, the expected time for completing an auction and the amount of signaling messages needed to conduct an auction are important consideration factors. These two elements are not considered in classical auctioning models. For Internet based auctions, in particular those available on wireless accesses, the issue of communication

cost cannot be ignored. From this perspective, the Dutch auctioning scheme intuitively has an advantage over the English auctioning scheme. For the former scheme, no bidder is required to bid more than once, whereas for the latter scheme bidders may have to bid several times whether they are successful at the bidding or not. Moreover, although the possibility of multiple bids at the same value occurs in both cases, it can happen at most once in the auctioning process for an item in a Dutch scheme and multiple times for the English scheme.

Another motivation of our work comes from the observation that in the literature, a common assumption is that the bidding offers take values from a continuum. This is an idealization of an actual bidding process. In practice, bid increments in an English auction or bid decrements in a Dutch auction is a discontinuous process. While a small discontinuous price decrement would minimize inaccuracies due to discrete optimization, it would also prolong the auctioning process. For auctions conducted over the Internet, the value of a discontinuous decrement could have significant implications on the communication cost. In particular, for a Dutch auction, the optimal strategy for price decrement is an interesting issue.

In this chapter, we present an analysis of an Internet based Dutch auctioning system. In any auction, there are three player roles that one can consider, the buyer, the seller, and the auction host. Traditional analyses on auctioning tend to focus on the roles played by the buyer or the seller. The role of the auction host is defined in terms of the type of auctioning system used. For auctions conducted on Internet, the auction host is bestowed with a new set of controlling mechanisms and faces a new set of objectives. Assuming the revenue of the auction host comes from commission based on the realized bid revenue and the varying part of an auctioning cost is proportional to the duration of the bidding process, one can formulate an objective function based on these two factors. In a Dutch auction, an important controlling mechanism available to an auction host is through price decrement strategy. The structure of the optimal price decrement strategy in an Internet based Dutch auctioning system is analyzed here via the Karush-Kuhn-Tucker condition. Moreover, we also established a numerically efficient algorithm to determine the optimal strategy. Numerical studies were carried out and we showed

that under certain conditions, the simple uniform decrement strategy can be close to the optimal strategy. These results form a small, first step to generalize the earlier works in the literature on auction in the new context of the Internet based environment.

The rest of the chapter is organized as follows. In section 3, we describe the system and optimization model. We maximize the expected revenue to the auction host by dynamically varying price decrement at each iteration. Knowledge on the number of bidders and the probability distributions of their valuations are exploited. In section 4, properties of the optimal solution are presented. We also show how the original problem can be reduced to a one dimensional numerical search problem. In section 5, we present numerical examples to illustrate the properties of the optimal solution. Performance comparison between the optimal strategy and a simple uniform decrement strategy is also given. Section 6 offers some concluding remarks.

## 7.3 System and Optimization Model

### 7.3.1 System Model

In an auction, there are three distinct player roles, namely the buyer, seller and auction host. In an Internet based auction, the auction host usually acts as the application server and provides the necessary information to implement the auction. It disseminates current price information to all bidders (*logon users*) regularly, and ends the auction when it receives a buying request from the users or the auction timeout is reached.

We assume only one item is sold in the auction. Initially, the auction starts at a given price $c_0$. A price is kept constant for a fixed interval until the next iteration. At iteration $k$, the price falls to $c_k$, under the constraint $c_{min} \leq c_k \leq c_{k-1}$. The auction will last for $M+1$ iterations, where $M$ is predetermined.

Denote $X_i$ as the valuation of bidder $i$. Let $n$ be the number of bidders. Since the auction host is also the application server it knows the value of $n$. In this work, we assume $n$ is constant for the duration of the auction. We also assume the valuations of the $i^{th}$ bidder, $X_i$, for all $i \in [1, n]$ are i.i.d. random variables drawn from a known distribution $F_X(.)$. In literature this is known as the case of *symmetric bidders*. However, our

solution methodology also works for *asymmetric bidders*. This is the case where bidder valuations are drawn from independent but not necessarily identical distributions.

Also denote $Y = \max(X_1, ..., X_n)$ as the maximum valuations of the bidders. It is straightforward to compute the cumulative distribution function *cdf* and probability density function *pdf* of $Y$ for the cases of symmetric or asymmetric bidders. We denote them by the notation $F(Y)$ and $f(Y)$ respectively. For simplicity, we assume that $f(Y)$ is a continuous positive function in the range $c_{min}, c_0$. Subsequently we will work with $Y$ directly since the sold price depends on the random variable $Y$. If the current selling price $c_k$ is lower than $Y$, at least one user will immediately make a bid and end the auction. If $c_M > Y$, the item will not be sold at the auction.

### 7.3.2 Optimization model

Suppose the item is sold at iteration $k$. $c_k$ is the selling price at iteration $k$, and $T$ is a non-negative time discounting increment at each iteration. Thus, the revenue is $R_k = c_k - kT$ if the item is sold at iteration $k, k \in \{0, ..., M\}$. If the item is not sold at the end of auction (i.e. $c_M > Y$), we define the corresponding revenue as $R_{M+1} = 0$.

The expected revenue upon selling the item is

$$\begin{aligned} p(\mathbf{c}) &= \mathbf{E}_{k\in\{0,...,M+1\}}[R_k] && (7.1)\\ &= \mathbf{E}_{k\in\{0,...,M\}}[c_k - kT] && (7.2)\\ &= c_0(1 - F(c_0)) + \sum_{k=1}^{M}(c_k - kT)(F(c_{k-1}) - F(c_k)). && (7.3) \end{aligned}$$

In our optimization model, we incorporated a time discounting factor $T$. The meaning of $T$ can be interpreted in two different scenarios. First of all, $T$ can represent the cost of using server resources in a wireless Dutch auction. Typically, the auction period spans only for minutes or hours. The auction server may update the price on a per minute or second basis. At each iteration, the application server has to broadcast a message to update the current price to all the clients (bidders). This generates a lot of data traffic and uses up bandwidth resources. Moreover, the amount of processing and bandwidth overhead can be regarded as being constant at each iteration. Thus, the resource usage is adequately modeled by a constant parameter $T$.

Alternatively, this model can also apply to online auction web sites where auction period spans over longer periods of time, such as days or weeks. It is common in these cases that the price of an item is dropped gradually on a daily basis. In this scenario, the amount of network traffic generated is insignificant. Rather, there is a time discounting factor on the value of the good to account for storage and maintenance cost incurred on the auction host.

The present problem belongs to a class of general nonlinear optimization problems with inequality constraints. The problem is to

$$\max_{(c_1,c_2,...,c_M)} p(\mathbf{c}) = \max_{(c_1,c_2,...,c_M)} c_0(1-F(c_0)) + \sum_{k=1}^{M}(c_k - kT)\Big(F(c_{k-1}) - F(c_k)\Big)$$

subject to the constraints

$$\begin{aligned} g_1(\mathbf{c}) &= c_1 - c_0 \leq 0, \\ g_2(\mathbf{c}) &= c_2 - c_1 \leq 0, \\ &\vdots \\ g_M(\mathbf{c}) &= c_M - c_{M-1} \leq 0, \\ g_{M+1}(\mathbf{c}) &= c_{min} - c_M \leq 0. \end{aligned}$$

We note that many alternative formulations such as dynamic programming [7] are possible. Our nonlinear programming formulation is desirable for practical implementation because, as we will show in the next section, the multivariable optimization problem of determining the selling price $M$-tuple $c_k, k = 1, 2, ..., M$ can be reduced to a one dimensional numerical search problem. Since an auction host typically has hundreds or thousands of items for sale, the reduction of computation complexity in price setting is desirable for running a large auction hosting site.

## 7.4 Properties of the optimal solution

In the literature, optimization of nonlinear functions subject to inequality constraints is well studied. The Karush-Kuhn-Tucker (KKT) Theorem is one of the powerful tools commonly employed.

Let $\mathbf{c}$ be any point in the feasible set. Denote $J(\mathbf{c}) = \{j : g_j(\mathbf{c}) = 0\}$. If $\nabla g_j(\mathbf{c})$ are mutually linearly independent for all $j \in J(\mathbf{c})$, then $\mathbf{c}$ is a *regular point*. The well known theorem due to Karush,Kuhn and Tucker [10] provides a necessary condition for a point to be a local maximizer, commonly known as the Karush-Kuhn-Tucker (KKT) condition, presented as follows.

*Let* $\mathbf{c}^*$ *be a regular point and local maximizer for the problem of maximizing* $p$ *subject to* $\mathbf{g}(\mathbf{c}) \leq 0$. *Then there exists a vector* $\mathbf{u}^* \in \Re^{M+1}$ *such that*

$$\mathbf{u}^* \geq 0 \tag{7.4}$$

$$\nabla p(\mathbf{c}^*) = \nabla \mathbf{g}(\mathbf{c}^*)\mathbf{u}^* \tag{7.5}$$

$$\mathbf{u}^{*T}\mathbf{g}(\mathbf{c}^*) = 0 \tag{7.6}$$

*where*

$$\mathbf{u}^* = \begin{pmatrix} u_1^* \\ u_2^* \\ \vdots \\ u_{M+1}^* \end{pmatrix} \qquad \mathbf{g}(\mathbf{c}^*) = \begin{pmatrix} g_1(\mathbf{c}^*) \\ g_2(\mathbf{c}^*) \\ \vdots \\ g_{M+1}(\mathbf{c}^*) \end{pmatrix}, \tag{7.7}$$

and $\nabla \mathbf{g}(\mathbf{c}^*)$ is the $M \times M+1$ matrix whose $i$-th column is $\nabla g_i(\mathbf{c}^*)$.

We refer to the vector $\mathbf{u}^*$ as the Karush-Kuhn-Tucker (KKT) multiplier vector. In the literature, a point satisfying the KKT condition (equation 7.4-7.6) is called a *critical point*. It follows from the KKT Theorem that a local maximizer is a critical point but not necessarily vice versa. We also define the global maximum in the feasible set as $\mathbf{c}^{**}$. Thus, if $\mathbf{c}^{**}$ is regular, it is also within the set of all critical points.

For the stated optimization problem, the constraint stated by equation 7.4 can be rewritten as:

$$\mathbf{g}(\mathbf{c}) = \begin{pmatrix} c_1 - c_0 \\ c_2 - c_1 \\ \vdots \\ c_{min} - c_M \end{pmatrix}. \tag{7.8}$$

Therefore,

$$\nabla \mathbf{g}(\mathbf{c}) = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 1 & -1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 \end{pmatrix}. \tag{7.9}$$

It is obvious that any selection of $M$ column vectors from $\nabla \mathbf{g}$ are mutually linearly independent. Hence any point, $\mathbf{c}$, in the feasible set with the cardinality of $J(\mathbf{c})$ less than $M+1$ is a regular point. The case where $J(\mathbf{c}) = \{1, 2, ..., M+1\}$ corresponds to $c_0 = c_1 = ... = c_M = c_{min}$, does not belong to the feasible set since $c_0 \neq c_{min}$. Thus, every point in the feasible set is regular. It follows from KKT Theorem that all local maxima are regular and should satisfy the KKT equations. As a result, $\mathbf{c}^{**}$ can be found by searching over the set of all critical points. Hereafter, we denote a critical point by $\mathbf{c}^*$.

On substitution of $\mathbf{g}(\mathbf{c}^*)$ and $\nabla \mathbf{g}(\mathbf{c}^*)$ into equation 7.5, one obtains

$$\begin{aligned} \nabla p(\mathbf{c}^*) &= \left( \frac{\partial p}{\partial c_1}(\mathbf{c}^*), \quad \frac{\partial p}{\partial c_2}(\mathbf{c}^*), \quad \ldots, \quad \frac{\partial p}{\partial c_M}(\mathbf{c}^*) \right) \\ &= \left( u_1^* - u_2^*, \quad u_2^* - u_3^*, \quad \ldots, \quad u_M^* - u_{M+1}^* \right). \end{aligned} \tag{7.10}$$

The last KKT condition in equation 7.6 leads to the conclusion

$$u_k^* g_k(\mathbf{c}^*) = u_k^*(c_k^* - c_{k-1}^*) = 0, \qquad \text{for} \quad k = 1, 2, ..., M+1. \tag{7.11}$$

since $c_k^* - c_{k-1}^*$'s are non-positive for all $k$.

There are standard algorithmic approaches to solve the class of convex programming problems in which the objective function is concave and the feasible set is convex. However, we show in the appendix that our objective function is not concave in general. As a result, local search technique is applied to identify local maxima. The search procedure is repeated with different initial points to discover as many distinct local maxima as possible. The best of these local maxima is chosen as the solution. Numerical

computations for this heuristic approach over the feasible set for local maxima can be quite extensive. This is an important consideration when the number of iterations is large. In this case an optimization problem in $M$ variables needs to be considered. However, it turns out that by exploiting our knowledge of the structure of the critical points, we could determine the global optimum $\mathbf{c}^{**}$ by reducing the problem to a one dimensional search problem. This is the main result provided by theorem 3. In the following, we present some basic properties of the optimal solution and describe an iterative solution methodology for finding the global maximum $\mathbf{c}^{**}$.

Suppose one implements an auction following the optimal price vector $\mathbf{c}^{**}$. At iteration $j$, the current price is $c_j^{**}$. Define the *subproblem starting at iteration j* as one in which there are $M-j$ remaining iterations, starting from the price $c_j^{**}$. The problem of finding the optimal price vector for this problem is equivalent to solving the problem,

$$\max_{(c_{j+1},\ldots,c_M)} \mathbf{E}_{k\in\{j+1,\ldots,M+1\}}[R_k|Y < c_j^{**}] \tag{7.12}$$

$$= \max_{(c_{j+1},\ldots,c_M)} \mathbf{E}_{k\in\{j+1,\ldots,M\}}[c_k - kT|Y < c_j^{**}]. \tag{7.13}$$

**Proposition 1** *For any $j$, $j \in \{1,...,M-1\}$, $(c_{j+1}^{**}, c_{j+2}^{**}, \ldots, c_M^{**})$ is the optimal price vector to subproblem starting at iteration $j$.*

**Proof:**

$$\max_{(c_{j+1},c_{j+2},\ldots,c_M)} \mathbf{E}_{k\in\{j+1,\ldots,M\}}[c_k - kT|Y < c_j^{**}] \tag{7.14}$$

$$= \max_{(c_{j+1},c_{j+2},\ldots,c_M)} \sum_{k=j+1}^{M} (c_k - kT)\frac{(F(c_{k-1}) - F(c_k))}{F(c_j^{**})} \tag{7.15}$$

$$= \frac{1}{F(c_j^{**})} \max_{(c_{j+1},c_{j+2},\ldots,c_M)} \sum_{k=j+1}^{M} (c_k - kT)(F(c_{k-1}) - F(c_k)). \tag{7.16}$$

It is obvious that the above expression is optimized when $c_k = c_k^{**}$ for $k = j+1,...,M$.

$\diamond$

The previous result states that if the number of bidders $n$ is constant throughout the auction period, then we need to compute $\mathbf{c}^{**}$ only once at the start of the auction. In practice, $n$ may change from time to time as bidders may join or leave during an

auction. In that case, one needs to get an update on the value of $n_j$ at iteration $j$ and computes the new optimal price vector.

When the number of iterations $M$ is large, numerical optimization becomes more complicated due to the number of variables involved. In theorem 1 and 3, we show that the multivariable optimization problem can be reduced to a one dimensional search problem.

**Theorem 1** *Suppose* $\mathbf{c}^*$ *is a critical point and there exists an integer* $i$ *such that* $c_i^* = c_{i+1}^*$, *where* $i = \{0, 1, ..., M-1\}$. *Then for all* $j > i$, $j \in \{i+1, ..., M\}$, $c_i^* = c_j^*$.

**Proof:**

Any feasible point must satisfy the condition:

$$c_0 \geq c_1^* \geq ... \geq c_k^* \geq c_{k+1}^* \geq ... \geq c_M^* \geq c_{min}. \tag{7.17}$$

We claim that there does not exist an integer $i$, $0 \leq i \leq M-2$, and a $j$, $i < j \leq M-1$, such that:

$$c_0 > c_1^* > ... > c_i^* = c_{i+1}^* = ... = c_j^* > c_{j+1}^* \geq ... \geq c_M \geq c_{min}. \tag{7.18}$$

We prove this statement by contradiction. By the Karush-Kuhn-Tucker Theorem we have to find $\mathbf{u}^* \geq \mathbf{0}$ so that equations 7.4-7.6 are satisfied. For convenience, we define the functions $h$ and $h_M$ as

$$h(c_{k-1}, c_k, c_{k+1}) = \frac{\partial p}{\partial c_k} \tag{7.19}$$

$$= F(c_{k-1}) - F(c_k) + f(c_k)(c_{k+1} - c_k - T) \qquad k \in \{1, \ldots, M-1\} \tag{7.20}$$

$$h_M(c_{M-1}, c_M) = \frac{\partial p}{\partial c_M} \tag{7.21}$$

$$= F(c_{M-1}) - F(c_M) + (c_M - MT)(-f(c_M)). \tag{7.22}$$

Since $g_k(\mathbf{c}^*) < 0$ for $k \in \{1, ..., i\}$, by equation 7.11,

$$u_k^* = 0 \qquad k \in \{1, ..., i\}. \tag{7.23}$$

At the point $\mathbf{c}^*$, by equation 7.10 the following equations hold:

$$\begin{aligned}
\tfrac{\partial p}{\partial c_k}(\mathbf{c}^*) &= h(c^*_{k-1}, c^*_k, c^*_{k+1}) = u^*_k - u^*_{k+1} = 0, \qquad k \in \{1, ..., i-1\}, \\
\tfrac{\partial p}{\partial c_i}(\mathbf{c}^*) &= h(c^*_{i-1}, c^*_i, c^*_{i+1}) = -u^*_{i+1}, \\
\tfrac{\partial p}{\partial c_{i+1}}(\mathbf{c}^*) &= h(c^*_i, c^*_{i+1}, c^*_{i+2}) = u^*_{i+1} - u^*_{i+2}, \\
&\vdots \\
\tfrac{\partial p}{\partial c_{j-1}}(\mathbf{c}^*) &= h(c^*_{j-2}, c^*_{j-1}, c^*_j), = u^*_{j-1} - u^*_j, \\
\tfrac{\partial p}{\partial c_j}(\mathbf{c}^*) &= h(c^*_{j-1}, c^*_j, c^*_{j+1}) = u^*_j - u^*_{j+1}.
\end{aligned} \tag{7.24}$$

Since

$$g_{j+1}(\mathbf{c}^*) = c^*_{j+1} - c^*_j < 0, \tag{7.25}$$

equation 7.11 implies that $u^*_{j+1} = 0$.

$$\begin{aligned}
\frac{\partial p}{\partial c_j}(\mathbf{c}^*) &= u^*_j - u^*_{j+1} = u^*_j && (7.26)\\
&= h(c^*_{j-1}, c^*_j, c^*_{j+1}) && (7.27)\\
&= F(c^*_{j-1}) - F(c^*_j) + f(c^*_j)(c^*_{j+1} - c^*_j - T) && (7.28)\\
&= f(c^*_j)(c^*_{j+1} - c^*_j - T) && (7.29)\\
&< 0. && (7.30)
\end{aligned}$$

That is $u^*_j < 0$, hence the non-negativity condition on $\mathbf{u}^*$ is not satisfied. By contradiction we show that the scenario stated in equation 7.18 cannot hold. $\diamond$

**Corollary 1** *Suppose* $\mathbf{c}^*$ *is a critical point and*

$$c_0 > c^*_1 > ... > c^*_i = c^*_{i+1} = ... = c^*_M \geq c_{min}. \tag{7.31}$$

*for some* $0 \leq i < M$. *Then*

$$\begin{aligned}
\tfrac{\partial p}{\partial c_j}(\mathbf{c}^*) &= 0 \qquad j \in \{1, ..., i-1\}, \\
\tfrac{\partial p}{\partial c_j}(\mathbf{c}^*) &\leq 0 \qquad j \in \{i, ..., M-1\}.
\end{aligned} \tag{7.32}$$

**Proof:**

The proof of theorem 1 shows that

$$\frac{\partial p}{\partial c_i}(\mathbf{c}^*) = -u^*_{i+1} \leq 0. \tag{7.33}$$

Since $c^*_i = c^*_{i+1} = ... = c^*_M$, we observe that

$$\begin{aligned}
\frac{\partial p}{\partial c_{i+1}}(\mathbf{c}^*) &= \frac{\partial p}{\partial c_{i+2}}(\mathbf{c}^*) = ... = \frac{\partial p}{\partial c_{M-1}}(\mathbf{c}^*) && (7.34)\\
&= h(c^*_i, c^*_{i+1}, c^*_{i+2}) && (7.35)\\
&= F(c^*_i) - F(c^*_{i+1}) + f(c^*_{i+1})(c^*_{i+2} - c^*_{i+1} - T) && (7.36)\\
&= -Tf(c^*_{i+1}) && (7.37)\\
&\leq 0. \quad \diamond && (7.38)
\end{aligned}$$

We now introduce the notation of a sequence-valued function

$$\hat{\mathbf{c}}(s) = (\hat{c}_0, \hat{c}_1, \hat{c}_2, ..., \hat{c}_M) = (c_0, s, \hat{c}_2, ..., \hat{c}_M). \tag{7.39}$$

The domain for $s$ is defined in the range $c_{min} \leq s \leq c_0$. The elements of $\hat{\mathbf{c}}$ are defined recursively in the following way:

Assume that elements up to $\hat{c}_k$ have been defined. Let $t$ be the solution to:

$$h(\hat{c}_{k-1}, \hat{c}_k, t) = 0. \tag{7.40}$$

(Note that by our assumption of the pdf and the definition of $h$, $t$ always exists and is unique.) If $c_{min} \leq t \leq \hat{c}_k$ and $t - (k+1)T > 0$, then $\hat{c}_{k+1} = t$. Otherwise, define $\hat{c}_{k+1} = \hat{c}_k$.

Note that $\hat{\mathbf{c}}(s)$ defines a 1-parameter family of critical points satisfying the KKT conditions. However, not all critical points can be represented by $\hat{\mathbf{c}}(s)$ for some $s$. Suppose $\mathbf{c}^*$ is a critical point and

$$c_0 > c^*_1 > ... > c^*_i = c^*_{i+1} = ... = c^*_M \geq c_{min} \tag{7.41}$$

for some $0 \leq i < M$. If $c^*_1 = \hat{c}_1$, then it follows directly from equation 7.24 and the definition of $\hat{\mathbf{c}}$ that

$$c^*_k = \hat{c}_k \qquad k \in \{1, ..., i-1\}. \tag{7.42}$$

However, $c_k^*$ and $\hat{c}_k$, for $k \geq i$, may not be equal. In general,

$$c_k^* \geq \hat{c}_k \qquad k \in \{i, ..., M-1\}. \tag{7.43}$$

To show that an optimal solution can be obtained by searching the family of critical points defined by $\hat{\mathbf{c}}(s)$, we need the following observation:

**Theorem 2** *Define $R_j = c_j^{**} - jT$, for $j \in \{1, \ldots, M\}$. If $R_j \leq 0$, then $c_k^{**} = c_{k-1}^{**}$ for $k \in [j, M]$. If $R_j > 0$, then $c_j^{**} < c_{j-1}^{**}$ if $c_{j-1}^{**} > c_{min}$.*

**Proof:**

Suppose $R_j < 0$. We have $R_M \leq R_{M-1} \leq ... \leq R_{j+1} \leq R_j < 0$.

$$\begin{aligned} p(\mathbf{c}) &= \sum_{k=1}^{M} R_k(F(c_{k-1}) - F(c_k)) & (7.44) \\ &= \sum_{k=1}^{j-1} R_k(F(c_{k-1}) - F(c_k)) + \sum_{k=j}^{M} R_k(F(c_{k-1}) - F(c_k)). & (7.45) \end{aligned}$$

If $c_k^{**} < c_{k-1}^{**}$ for any $k \in [j, M]$, $p$ can be increased by setting $c_k^{**} = c_{k-1}^{**}$ for $k \in [j, M]$. A contradiction.

Suppose $R_j = 0$ and $c_j^{**} < c_{j-1}^{**}$. $p$ can be increased by changing $c_j^{**}$ to any value in the interval $(c_j^{**}, c_{j-1}^{**})$, and setting $c_k^{**} = c_{k-1}^{**}$ for $k \in [j+1, M]$. Again a contradiction.

On the other hand, suppose $R_j > 0$. If $c_j^{**} = c_{j-1}^{**}$ then all $c_k^{**}$'s must be equal for $k \geq j$ according to theorem 1. Therefore, $p(\mathbf{c})$ can be increased by setting $c_j^{**}$ to a value in the interval $(c_{min}, c_{j-1}^{**})$ while keeping $R_j > 0$. A contradiction. Hence, $c_j^{**} < c_{j-1}^{**}$. $\diamond$

Now we are ready to prove our main result:

**Theorem 3** *If $\mathbf{c}^{**}$ is an optimal solution, then $\hat{\mathbf{c}}(s) = \mathbf{c}^{**}$ when $s = c_1^{**}$.*

**Proof:**

Suppose $c^{**} = (c_0^{**}, c_1^{**}, \ldots, c_M^{**})$ is an optimal solution. Set $s = c_1^{**}$.

Define $R_j = c_j^{**} - jT$. Suppose $R_j > 0$ for all $j \in \{2, \ldots, M\}$, then it follows from corollary 1 and theorem 2 that $c_{j-1}^{**} > c_j^{**}$ and $h(c_{j-2}, c_{j-1}, c_j) = 0$ unless $c_{j-1}^{**} = c_{min}$. It follows from the definition of $\hat{\mathbf{c}}(s)$ that

$$\hat{c}_k = c_k^{**} \tag{7.46}$$

for all $k$.

Suppose $R_j \leq 0$ for some $j \in \{2, \ldots, M\}$. If $j < M$, notice that $c_M^{**} = c_{M-1}^{**} = \ldots = c_j^{**}$ since $R_k < 0$ for $k > j$. Hence, it follows from the definition of $\hat{\mathbf{c}}(s)$ that

$$\hat{c}_k = c_k^{**} \tag{7.47}$$

for all $k$. $\diamond$

According to this theorem, by doing a one-dimensional numerical search for $\hat{\mathbf{c}}(s)$ within the feasible set, one can obtain an optimal solution to the problem. We now describe two more observations on the structure of the optimal solution. First of all, the following theorem shows that depending on whether $F(.)$ is convex or concave, the sequence of price difference of the optimal strategy, $c_k^{**} - c_{k+1}^{**}$, satisfies the following inequalities.

**Theorem 4** *Given* $\mathbf{c}^{**}$ *makes the form* $c_0 > c_1^{**} > ... > c_i^{**} = c_{i+1}^{**} = ... = c_M^{**} \geq c_{min}$. *If* $F(.)$ *is convex, then*

$$c_{k-1}^{**} - c_k^{**} < c_k^{**} - c_{k+1}^{**} + T \qquad k \in \{1, ..., i-1\}. \tag{7.48}$$

*If* $F(.)$ *is concave, then*

$$c_{k-1}^{**} - c_k^{**} > c_k^{**} - c_{k+1}^{**} + T \qquad k \in \{1, ..., i-1\}. \tag{7.49}$$

**Proof:**

By Corollary 1,

$$\frac{\partial p}{\partial c_k}(\mathbf{c}^{**}) = 0 \qquad k \in \{1, ..., i-1\}$$

. That is,

$$F(c_{k-1}^{**}) - F(c_k^{**}) + f(c_k^{**})(c_{k+1}^{**} - c_k^{**} - T) = 0. \tag{7.50}$$

Consider the case when $F(.)$ is convex.

$$\frac{F(c_{k-1}^{**}) - F(c_k^{**})}{c_{k-1}^{**} - c_k^{**}} > f(c_k^{**}) \tag{7.51}$$

or

$$F(c_{k-1}^{**}) - F(c_k^{**}) + f(c_k^{**})(c_k^{**} - c_{k-1}^{**}) > 0. \tag{7.52}$$

Subtracting equation 7.50 from equation 7.52, we have

$$f(c_k^{**}) \left[ (c_k^{**} - c_{k-1}^{**}) - (c_{k+1}^{**} - c_k^{**} - T) \right] > 0 \tag{7.53}$$

or

$$c_{k-1}^{**} - c_k^{**} < c_k^{**} - c_{k+1}^{**} + T. \tag{7.54}$$

The case when $F(.)$ is concave can be proven in the same way. $\diamond$

When $X$ is uniformly distributed as $U(a, b)$, the pdf of $Y$ is convex. In the special case $T = 0$, we note that the price difference $c_k^{**} - c_{k+1}^{**}$ is increasing with time, whereas the probability $F(c_k^{**}) - F(c_{k+1}^{**})$ is decreasing. This conforms to our intuition that price levels should be closely packed at intervals where pdf of $Y$ is large, such that the item could be sold at a price $c_k$ close to $Y$.

When $X$ is normal distributed as $N(\mu, \sigma^2)$,

$$F(y) = Q\left(\frac{\mu - y}{\sigma}\right)^n, \tag{7.55}$$

$$f(y) = nQ\left(\frac{\mu - y}{\sigma}\right)^{n-1} \frac{1}{\sqrt{2\pi}\sigma} \exp(\frac{-(\mu - y)^2}{2\sigma^2}). \tag{7.56}$$

It could be shown that $F(y)$ is convex when $y \leq \psi$ and concave otherwise, where $\psi$ is solution to the equation

$$(n-1)\exp(\frac{-x^2}{2}) + \sqrt{2\pi}xQ(x) = 0, \qquad x = \frac{(\mu - \psi)}{\sigma}. \tag{7.57}$$

Since the pdf of $Y$ is largest at $\psi$, the price difference is decreasing at first and starts increasing again as $c_k^{**}$ passes through $\psi$. That is, price levels are more closely packed around $Y = \psi$.

A uniform price decrement strategy is used as a reference in the numerical studies. The price vector starts at $c_0$ and falls to $c_{min}$ in $M$ equally spaced steps. We hereafter refer this strategy as the *uniform price decrement strategy*. It turns out that this strategy is optimal in the trivial case as shown in the following theorem.

**Theorem 5** *Uniform decrement strategy is optimal when*

*(1)* $X \sim U(a,b)$ $\qquad c_0 \leq b$ *and* $a \leq c_{min}$

*(2)* $n = 1$

*(3)* $T = 0$

*The optimal price levels are given by*

$$c_k^{**} = \left(\tfrac{M-k}{M}\right) c_0 - \left(\tfrac{k}{M}\right) c_{min}, \qquad c_{min} \geq \frac{c_0}{M+1}, \tag{7.58}$$

$$c_k^{**} = \left(\tfrac{M+1-k}{M+1}\right) c_0, \qquad c_{min} \leq \frac{c_0}{M+1}. \tag{7.59}$$

**Proof:**

We will show that $\mathbf{c}^*$ defined in equation 7.58 and equation 7.59 satisfies the KKT conditions. Then we prove that $p$ is concave in the feasible set. Since $p$ is concave, the optimality of $\mathbf{c}^*$ is proved.

$$F(y) = \begin{cases} \frac{y-a}{b-a} & a \leq y \leq b \\ 0 & \text{otherwise} \end{cases} \qquad f(y) = \frac{1}{b-a} \qquad a \leq y \leq b$$

Substitute to equation 7.3 we have

$$p = \frac{1}{b-a} \sum_{k=1}^{M} c_k (c_{k-1} - c_k) \tag{7.60}$$

Taking partial derivatives w.r.t. $c_k$

$$\frac{\partial p}{\partial c_k} = \frac{1}{b-a} \left[ (c_{k-1} - c_k) - (c_k - c_{k+1}) \right] \qquad k \in \{1, ..., M-1\} \tag{7.61}$$

$$\frac{\partial p}{\partial c_M} = \frac{1}{b-a} (c_{M-1} - 2c_M) \tag{7.62}$$

Consider the case $c_{min} \geq \frac{c_0}{M+1}$. Substitute the price vector $\mathbf{c}^{**}$ equation 7.58 to

equation 7.61, we have

$$\frac{\partial p}{\partial c_k}(\mathbf{c}^{**}) = 0 \qquad k \in \{1, ..., M-1\} \tag{7.63}$$

$$\frac{\partial p}{\partial c_M}(\mathbf{c}^{**}) = \frac{1}{b-a}\left[-c_{min} + \left(\frac{c_0 - c_{min}}{M}\right)\right] \tag{7.64}$$

$$= \frac{1}{b-a}\left[\frac{M-1}{M}(\frac{c_0}{M+1} - c_{min})\right] \tag{7.65}$$

$$\leq 0. \tag{7.66}$$

$g_k(\mathbf{c}^{**}) < 0 \quad \forall k \in \{1, ..., M\}$. By construction $u_k^* = 0, k \in \{1, ..., M\}$, so that equation 7.11 is satisfied. Moreover, $u_{M+1}^{**} \geq 0$. Thus we have

$$\frac{\partial p}{\partial c_k}(\mathbf{c}^{**}) = u_k^{**} - u_{k+1}^{**} = 0 \qquad k \in \{1, M-1\} \tag{7.67}$$

$$\frac{\partial p}{\partial c_M}(\mathbf{c}^{**}) = u_M^{**} - u_{M+1}^{**} \tag{7.68}$$

$$= -u_{M+1}^{**} \leq 0. \tag{7.69}$$

Thus the KKT condition 2 is also satisfied. Therefore there exists a non-negative $\mathbf{u}^*$ that satisfies all the KKT conditions.

Consider the case $c_{min} \leq \frac{c_0}{M+1}$. We substitute the price vector $\mathbf{c}^{**}$ equation 7.59 to equation 7.61, yielding

$$\frac{\partial p}{\partial c_k}(\mathbf{c}^{**}) = 0 \qquad k \in \{1, ..., M-1\} \tag{7.70}$$

$$\frac{\partial p}{\partial c_M}(\mathbf{c}^{**}) = \frac{1}{b-a}(c_{M-1}^{**} - 2c_M^{**}) \tag{7.71}$$

$$= 0 \tag{7.72}$$

on simplification. Since $g_k(\mathbf{c}^{**}) < 0$ for $k \in \{1, ..., M+1\}$, by equation 7.11 $u_k^{**} = 0, k \in \{1, ..., M+1\}$ . Thus we have $u_k^{**} - u_{k+1}^{**} = 0$. All the KKT conditions are satisfied again for this $\mathbf{c}^{**}$.

To show that $\mathbf{c}^{**}$ is a global maximizer, we proceed to prove the Hessian matrix $\mathbf{H}$

for the objective function $p$ is negative semi-definite. It is trivial to show that

$$\mathbf{H} = \frac{1}{b-a}\begin{pmatrix} -2 & 1 & 0 & \cdots & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ & & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 1 & -2 & 1 & 0 \\ 0 & \cdots & 0 & 1 & -2 & 1 \\ 0 & \cdots & \cdots & 0 & 1 & -2 \end{pmatrix}$$

Apply the Gerschgorin's theorem on each row of $\mathbf{H}$, we show that $\max(\lambda) \leq 0$. Thus, $\mathbf{H}$ is negative semi-definite and $p$ is concave. $\diamond$

So far Theorem 3 is the most important observation. Suppose an auction host knows the statistics of the individual valuation $F$. The auction host only needs to search for different values of $s$ for the optimal value $c_1^{**}$. The nature of the optimization problem stipulates that if $s = c_1^{**}$, then $\mathbf{c}^{**} = \hat{\mathbf{c}}(s)$ is the optimum solution to the optimization problem. Given $s$, $\hat{\mathbf{c}}(s)$ can be determined easily by recursively solving simple algebraic equations (7.40) $M-1$ times. Thus, an exhaustive search of $s$ leads to the solution for the optimal price settings.

## 7.5 Numerical Studies

In this section, we firstly present several numerical examples to illustrate the properties of $\mathbf{c}^{**}$. Then, the optimal strategy is compared to the *uniform price decrement strategy* in the following subsection.

### 7.5.1 Illustration of properties of optimal solution

In figure 7.1, $X$ is uniformly distributed as $U(700, 1000)$. There is no discounting factor, i.e. $T = 0$. The optimal price $c_k^{**}$ at iteration $k$ is plotted for $n = 1, 5, 10, 20, 50$ respectively. When $n = 1$, we observe that the uniform decrement strategy is optimal. Since $c_{min} \geq \frac{c_0}{M+1}$, $\mathbf{c}^{**}$ is given by equation 7.58. As $n$ increases, the pdf of $Y$ shifts to the right. Thus the price decrement rate is more gradual. When $n = 20$, $c_M$ is

approximately equal to 870. When $n = 50$, $c_M$ is approximately equal to 930. In both cases, we note that the probability $Pr[Y \leq c_M]$ is very small. We also observe that the cost difference $c_k^{**} - c_{k-1}^{**}$ is increasing with $k$. The observation is in agreement to equation 7.48 since $F_Y(y)$ is convex.

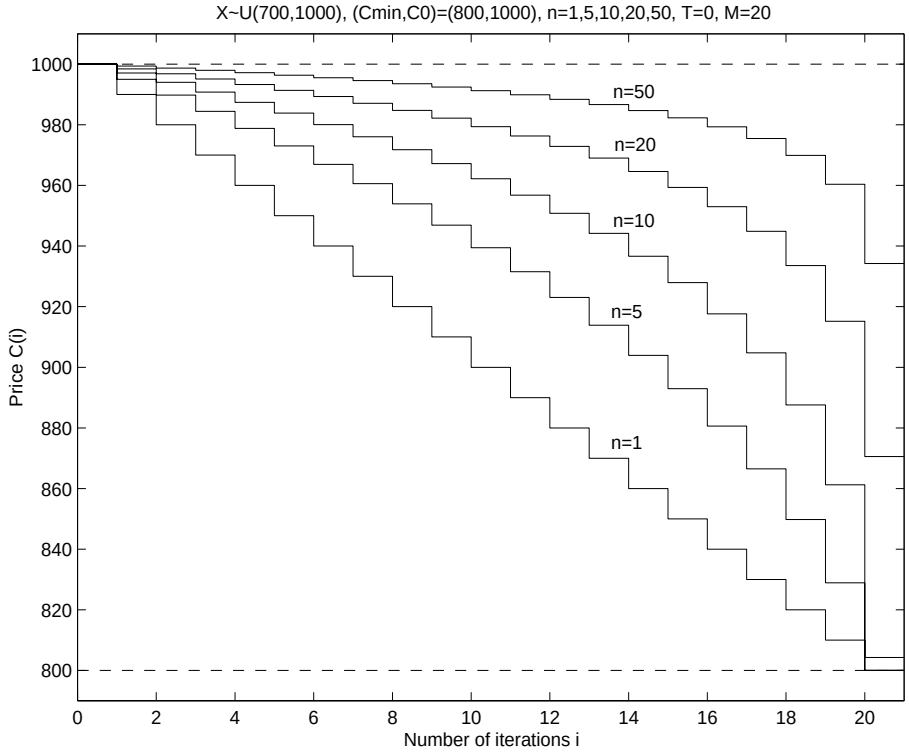



Figure 7.1: Example 1: $X$ uniformly distributed, $T = 0$

In the following examples, we assume the valuation of a bidder $X$ is normal distributed with mean 850 and variance $50^2$. The pdf of maximum valuation $Y$ for different $n$ is shown on figure 7.2. Note that the pdf becomes more peaked and shifts to the right as $n$ increases. Figure 7.3 shows the case when $T = 0$ and $X$ is normal distributed as $N(850, 50^2)$. For all values of $n$, the price difference is decreasing at first and increasing towards the end. This agrees with our results for normal distributed $X$'s, since $F(Y)$ changes from convex to concave as $Y$ increases. When $n = 1$, the pdf is maximum around $\mu = 850$. Thus, the initial drop in price is fast. After that, the drop in price is about the same in each iteration. On the other hand, when $n = 50$, the pdf is maximum around 960 and tends to zero around 900. Thus, the drop in price is slowest around 960 and becomes faster after it passes through the pdf maxima. Note that for

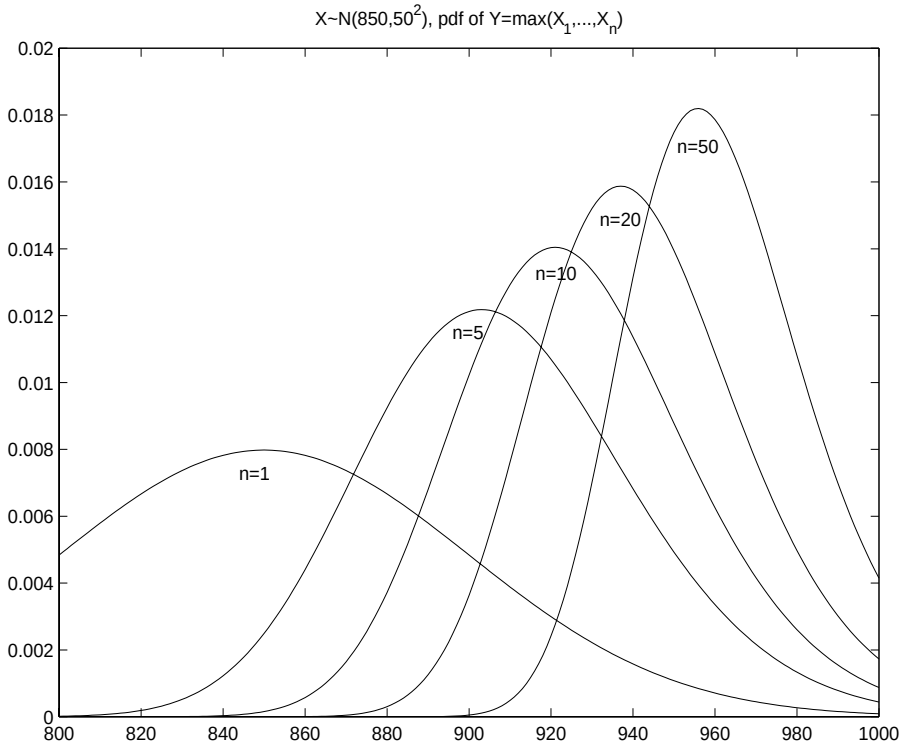



Figure 7.2: pdf of $Y$ for $n = 1, 5, 10, 20, 50$

$n = 10, 20, 50$, the iterations ends essentially at the point when the pdf is essentially zero.

In figure 7.4, a discounting factor of $T = 20$ is introduced in each iteration. The price $c_k^{**}$ at iteration $k$ is plotted for the cases $n = 1, 5, 10, 20, 50$. We observe the inclusion of a non-zero discounting factor $T$ leads to faster price decrements. As one can read from figure 7.3, the maxima of $f(Y)$ occur around $850, 900, 920, 940, 960$ respectively when $n = 1, 5, 10, 20, 50$. The optimal price decrement $c_1^{**}$ lies in the vicinity of these pdf maxima. We note that when $n = 50$, the simulation result is suboptimal. The auction ends at iteration 15 at a price higher than $c_{min}$. Thus, the expected revenue can be further increased by additional price decrements. The discrepancy is due to numerical inaccuracies that occur at iterations when the pdf of $Y$ at the selling price is extremely small. In this case, the pdf of $Y$ around $c_{15}$ is less than $10^{-9}$. Despite the numerical inaccuracies, the proposed method tends to yield solutions that are nearly optimal since the difference in the expected revenue is small.

In figure 7.5, we change the discounting factor to $T = 50$. The optimal price

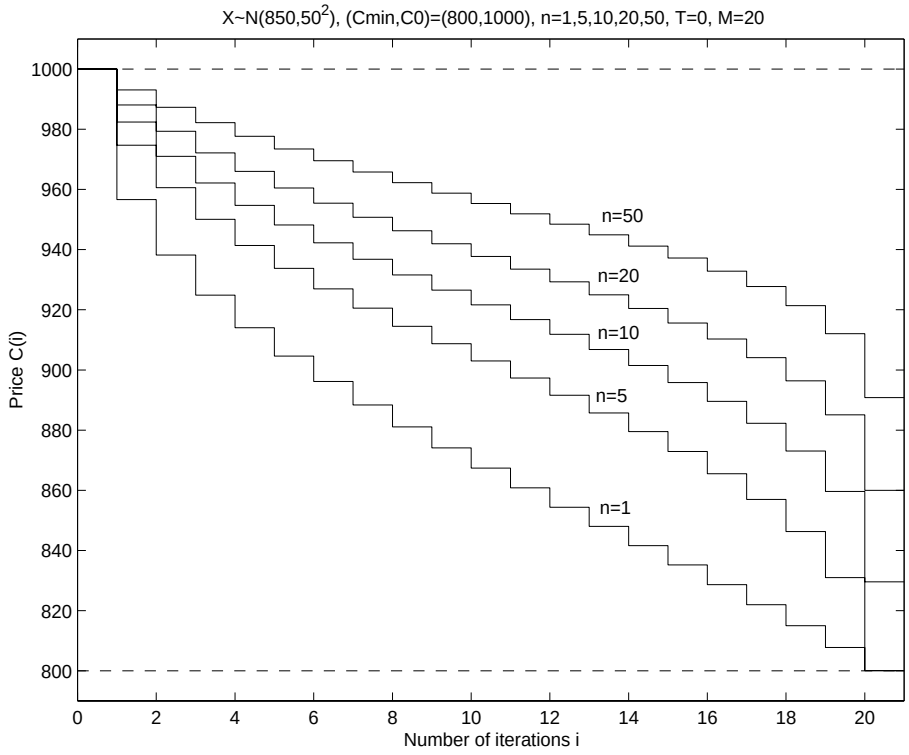



Figure 7.3: Example 2: $X$ normal distributed, $T = 0$

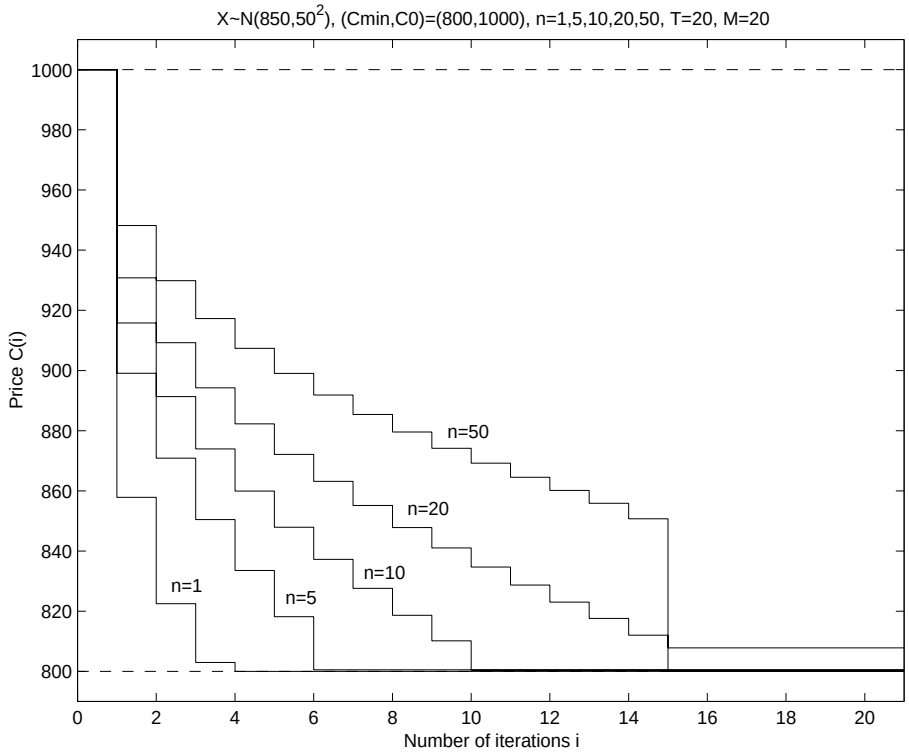



Figure 7.4: Example 3: $X$ normal distributed, $T = 20$

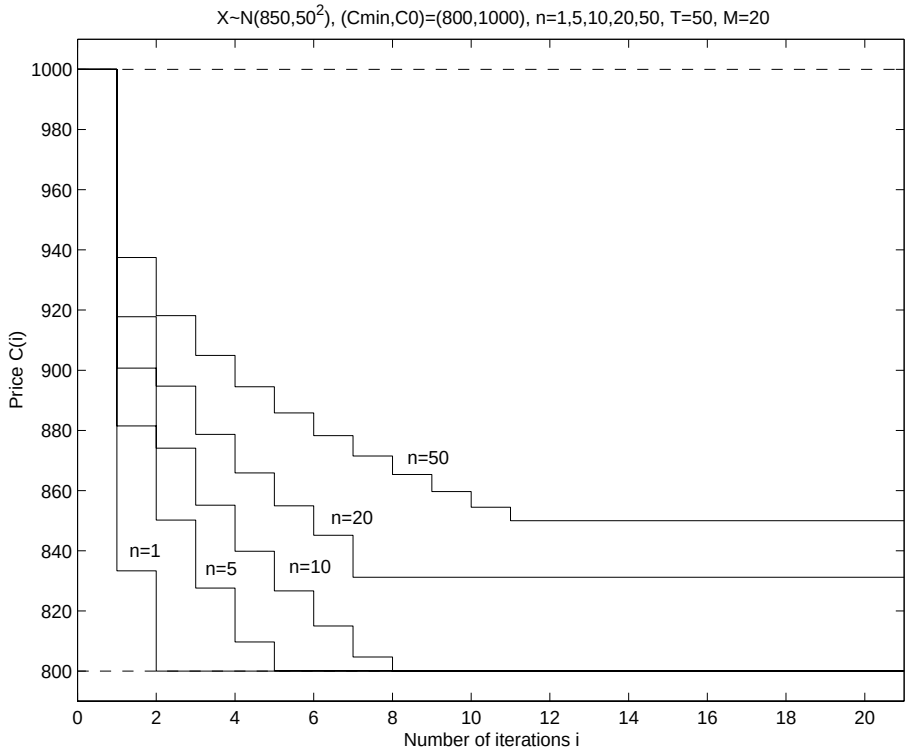



Figure 7.5: Example 4: $X$ normal distributed, $T = 50$

decrements for different $n$ are shown. As predicted, the price decrement is even steeper compared to the cases where $T = 20$ and $T = 0$. In the case $n = 20$ and $n = 50$, the iterative solution fails to touch $c_{min}$ when the auction ends due to the resolution inaccuracy in the search. Again, a near optimal solution is obtained since the pdf when the auction ends is very small (less than $10^{-10}$).

In figure 7.6, the price decrements are compared for different discounting factor $T = 0, 5, 10, 20, 50$. The number of bidders is $n = 10$. As $T$ increases, the price decrement is steeper. Thus, if the resource usage is expensive, the auction host would prefer a strategy with faster price decrements.

### 7.5.2 Comparison with the uniform decrement strategy

Comparison is done in terms of the expected revenue $p$. The ratio $p(\mathbf{c}^{**})/p(\mathbf{c_{ref}})$ is shown in Table 7.1. We also compare the expected time to sell an item, as shown in Table 7.2. Suppose a strategy $\mathbf{c}$ is used. Define the expected time to sell an item $T_s$, given that it is actually sold when the auction ends as $\mathbf{E}[T_s|Y \geq c_M]$. It is straightforward to

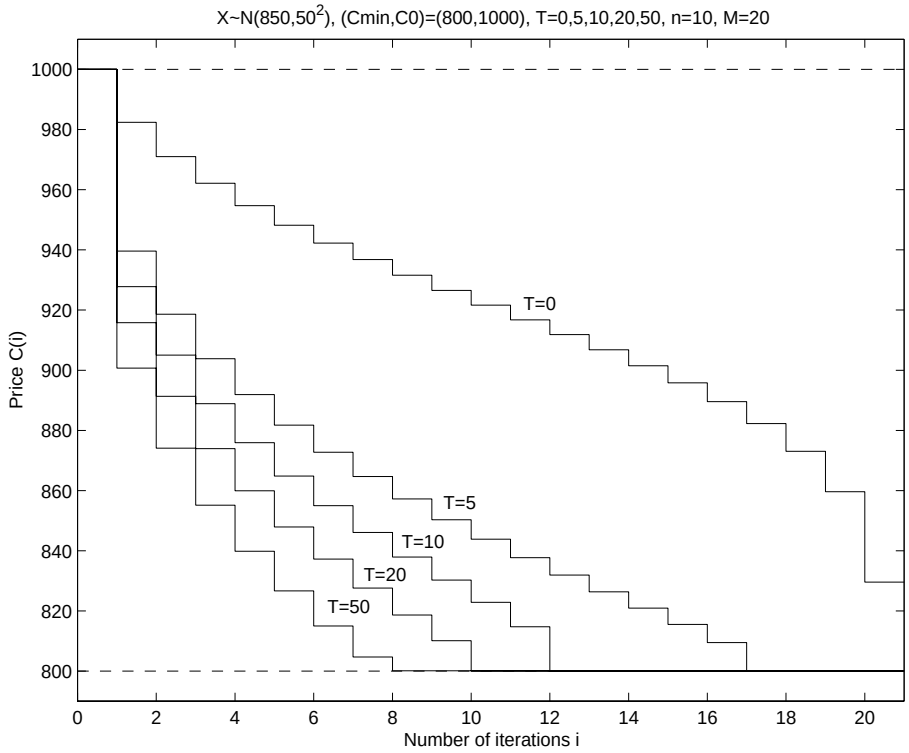



Figure 7.6: Example 5: $X$ normal distributed, $n = 10$

show that

$$\mathbf{E}[T_s|Y \geq c_M] = \frac{\sum_{k=1}^{M} k(F(c_{k-1}) - F(c_k))}{1 - F(c_M)} \tag{7.73}$$

$$= \frac{\sum_{k=0}^{M-1} F(c_k) - MF(c_M)}{1 - F(c_M)}. \tag{7.74}$$

As illustrated in Table 7.1, the ratio of expected revenue for the optimal strategy over the reference strategy is shown. When $T = 0$, the optimal strategy offers marginal improvement over a uniform price decrement strategy. This is true in general irrespective of the distribution of $X$. In general, the successful bidder will pay a little less than his valuation of the good under all price decrement strategies. The strategies are different mainly in the expected time that the product is sold. Since there is no discounting of revenue with time, the expected revenue of all strategies should appear the same. Applying this result to the context of wireless Dutch auction, we infer that a uniform decrement strategy is nearly optimal if the resource and processing overhead is low.

When the discounting factor is non-zero, the optimal strategy outperforms the alternate strategy by a large margin. In our studies, the difference is more remarkable

| $n$ | 1 | 5 | 10 | 20 | 50 |
|---|---|---|---|---|---|
| $T = 0, X \sim U(700, 1000)$ | 1.0000 | 1.0012 | 1.0027 | 1.0042 | 1.0058 |
| $T = 0, X \sim N(850, 50^2)$ | 1.0009 | 1.0012 | 1.0018 | 1.0023 | 1.0028 |
| $T = 20, X \sim N(850, 50^2)$ | 1.3920 | 1.2033 | 1.1444 | 1.1000 | 1.0566 |
| $T = 50, X \sim N(850, 50^2)$ | 4.8749 | 1.9413 | 1.5764 | 1.3655 | 1.1948 |

Table 7.1: Revenue ratio of the optimal and the reference strategy when $T$ and $n$ are varied.

| $n$ | 1 | 5 | 10 | 20 | 50 |
|---|---|---|---|---|---|
| $T = 0, X \sim U(700, 1000)$ | 10.50/3.67 | 8.10/5.42 | 7.30/3.25 | 7.06/1.98 | 6.89/1.22 |
| $T = 0, X \sim N(850, 50^2)$ | 11.41/11.23 | 9.87/9.69 | 9.76/7.82 | 9.57/6.19 | 8.98/4.31 |
| $T = 20, X \sim N(850, 50^2)$ | 1.65/11.23 | 1.57/9.69 | 1.49/7.82 | 1/41/6.19 | 1.29/4.32 |
| $T = 50, X \sim N(850, 50^2)$ | 1.25/11.23 | 1.24/9.69 | 1.20/7.82 | 1.15/6.19 | 1.08/4.32 |

Table 7.2: Expected time to sell an item for the optimal and the reference strategy.

when $n$ is small or $T$ is large. When $T$ is large, it is desirable to sell an item sooner to reap more profits. Reading from Table 7.2, the expected time to sell the good is much shorter for the optimal strategy. Thus, the optimal strategy is significantly better when $T$ is large. In a wireless Dutch auction, it is reasonable to assume the number of bidders $n = 5$ or $n = 10$. For the case $T = 50$, the optimal strategy is superior to the uniform strategy by 94% and 57% respectively as read from Table 7.1. Thus, in a wireless Dutch auction, we should refrain from using the uniform strategy if communication resources are expensive. Similarly, when the value of a good suffers from fast time discounting, as in perishable goods such as Dutch tulips, the uniform strategy should not be used.

From figure 7.2 we observe that the pdf shifts to the low price region when $n$ is small. For the uniform decrement strategy, it takes many iterations until the item is sold. Thus the revenue suffers from large time discounting. In the optimal strategy, the initial price decrements are steep so that the maximum valuation is reached upon several iterations. In practice, a Dutch auction is usually started at a very high initial price $c_0 \gg Y > c_{min}$. The maximum valuation $Y$ is substantially below $c_0$. The use of the optimal strategy leads to significant gain over the uniform decrement strategy by significantly shortening the auction time and the corresponding amount of time

discounting.

We have demonstrated that when the time discounting factor $T$ is large and when the number of bidders $n$ is small, the optimal strategy have the potential to outperform the uniform decrement strategy by a large margin. The underlying reason for the performance difference is that the auction time is significantly shortened for the optimal strategy as illustrated in Table 7.2. More generally, when parameters such as $c_{min}$, $\sigma$ and $M$ are varied, the auction time may be adversely prolonged by using the uniform decrement strategy. Thus under certain parameter settings, we also observe a large performance margin between the optimal and the uniform decrement strategy. Specifically, when the lower price limit $c_{min}$ and individual valuations $X_i$ are small compared with the initial price $c_0$, the auction time for the reference strategy is considerably longer. Similarly, when the variance of the individual valuations $\sigma$ is small, the maximum valuation over all bidders $Y$ is also smaller, thus prolonging the auction time of the uniform decrement strategy. Finally, when the number of allowed iterations $M$ in an auction increases, the resultant price decrement interval of the uniform strategy is finer. This also adversely affect the performance of the reference strategy relative to the optimal strategy.

| Parameters | revenue ratio |
|---|---|
| $T = 10, n = 10, c_{min} = 100, M = 20, \sigma = 50, X \sim N(300, 50^2)$ | 1.6168 |
| $T = 10, n = 10, c_{min} = 100, M = 20, \sigma = 100, X \sim N(300, 50^2)$ | 1.3256 |
| $T = 10, n = 10, c_{min} = 100, M = 20, \sigma = 25, X \sim N(300, 50^2)$ | 1.8736 |
| $T = 10, n = 10, c_{min} = 100, M = 10, \sigma = 25, X \sim N(300, 50^2)$ | 1.5414 |
| $T = 10, n = 10, c_{min} = 100, M = 30, \sigma = 25, X \sim N(300, 50^2)$ | 2.3685 |

Table 7.3: Revenue ratio of the optimal and the reference strategy when $c_{min}$, $\sigma$ and $M$ are varied.

As a simple illustration we consider 5 more numerical examples with results shown in Table 7.3. In all the five examples, the lower price limit is $c_{min} = 100$. Individual valuations are modeled as i.i.d. Gaussian random variables with mean $\mu = 300$ and variance $\sigma^2 = 50^2$. The revenue ratio of the optimal strategy relative to the uniform decrement strategy is found to be more than 60%. This confirms our intuition that when

$c_0 \gg Y > c_{min}$, the optimal strategy outperforms the reference strategy by a large margin. In the second and third examples, the variance of the individual valuations is varied as $\sigma^2 = 100^2$ and $\sigma^2 = 25^2$ respectively. The corresponding revenue ratios are 1.3256 and 1.8736. Our results show that as the variance $\sigma^2$ increases, the revenue ratio also increases. When the valuations of the bidders show smaller randomness, it is unlikely that the maximum valuation is much higher than $\mu$. This decreases the efficiency of the uniform decrement strategy considerably. Finally, in the fourth and fifth examples we vary the number of iterations to $M = 10$ and $M = 30$ respectively. The corresponding revenue ratios are 1.5414 and 3.1962. This shows that when the number of iterations is large, the optimal strategy may lead to an improvement that is quite significant, as much as three times the revenue of the reference strategy.

### 7.6 Conclusion

In this chapter, we present the optimal price decrement strategy for Dutch auction. It is shown in a system with inexpensive resources/low time discounting factor, the uniform decrement strategy is nearly optimal irrespective of the distribution of $X$. When resources are expensive/time discounting is high, the optimal strategy has steeper price decrements and is more favorable to the uniform strategy. Finally when the initial price is very high compared with the maximum valuation and the lower price limit, as in a practical Dutch auction, the optimal strategy is significantly better than the uniform strategy. We conclude that the optimal price decrement strategy is useful in a variety of contexts such as the wireless Dutch auction or online auction houses.

## Appendix

In this appendix an argument to show why the objective function $p$ is not concave in general is described.

In order to render $p(\mathbf{c})$ concave, the Hessian matrix $\mathbf{H} = \{h_{k,j}\}$ must be negative semi-definite, where

$$h_{k,j} = \frac{\partial^2 p}{\partial c_j \partial c_k}$$

Recall that

$$p(\mathbf{c}) = \sum_{k=1}^{M}(c_k - kT)(F(c_{k-1}) - F(c_k)) + c_0(1 - F(c_0)).$$

Taking partial derivatives with respect to $c_k$,

$$\frac{\partial p}{\partial c_k} = F(c_{k-1}) - F(c_k) + f(c_k)(c_{k+1} - c_k - T) \qquad k \in \{1, ..., M-1\} \tag{7.75}$$

$$\frac{\partial p}{\partial c_M} = F(c_{M-1}) - F(c_M) + (c_M - MT)(-f(c_M)). \tag{7.76}$$

Differentiating w.r.t. $c_j$ again. For $k \in \{1, ..., M-1\}$, we have

$$h_{k,j} = \begin{cases} f(c_{k-1}) & j = k-1 \\ -2f(c_k) + f'(c_k)(c_{k+1} - c_k - T) & j = k \\ f(c_k) & j = k+1 \\ 0 & o.w. \end{cases}$$

Whereas, for $k = M$, we have

$$h_{M,j} = \begin{cases} f(c_{M-1}) & j = M-1 \\ -2f(c_M) + (-f'(c_M))(c_M - MT) & j = k \\ 0 & o.w. \end{cases}$$

We observe that $\mathbf{H}$ is tri-diagonal with negative entries along the diagonals and positive entries adjacent to the diagonal entries. A sufficient condition to guarantee that $\mathbf{H}$ is negative semi-definite is to ensure its row sums are smaller than or equal to zero. However, this does not hold in general unless $T$ is very large.

To give an example, we consider the case where $M = 1$, $n = 1$, and $X$ is an exponential random variable with mean equal to 1. In this case,

$$p = c_0(1 - F(c_0) + (c_1 - T)(F(c_0) - F(c_1)).$$

Differentiating this function twice, we have

$$\frac{d^2p}{dc_1^2} = (c_1 - T - 2)e^{-c_1} \geq (c_{min} - T - 2)e^{-c_1}.$$

If $T < c_{min} - 2$, then $p$ is not concave.

To estimate the range of $T$ such that $p$ is concave, or $\mathbf{H}$ is negative semi-definite, one can use inclusion theorems on eigenvalues such as the Gerschgorin's theorem [88]. Since $\mathbf{H}$ is symmetric, all the eigenvalues are real. Applying the Gerschgorin's theorem on $\mathbf{H}$, each eigenvalue $\lambda_i$ must satisfy

$$\lambda_i \leq h_{i,i} + \sum_{j=1, j\neq i}^{M} |h_{i,j}|$$

in the feasible set. Thus $\mathbf{H}$ is negative semi-definite if $\max_i h_{i,i} + \sum_{j=1,j\neq i}^{M} |h_{i,j}| \leq 0$. On substitution, we have

$$\begin{aligned} -f(c_1) + f'(c_1)(c_2 - c_1 - T) &\leq 0 && (7.77) \\ f(c_{k-1}) - f(c_k) + f'(c_k)(c_{k+1} - c_k - T) &\leq 0 && (7.78) \\ f(c_{M-1}) - 2f(c_M) - f'(c_M)(c_M - MT) &\leq 0. && (7.79) \end{aligned}$$

Note that equation 7.77 holds in the feasible set. If equation 7.78 is true, then

$$T \geq \frac{f(c_{k-1}) - f(c_k)}{f'(c_k)} + (c_{k+1} - c_k).$$

One can set $c_k = c_{k-1} = c_0$ and $c_{k+1} = c_{min}$ for example. Then $T \geq c_0 - c_{min}$ must be satisfied to ensure $p$ is concave.

# Chapter 8

# Conclusions

## 8.1 Introduction

This thesis is a collection of research on the fundamental network behaviors that pertain to the subject of mobile ad hoc networks. A brief summary of our contributions is given in section 8.2. We have shown a mobile infostation network is superior to multihop ad hoc network when the network is operated under stress such as high node density or high node mobility. These desirable properties may have important implications on futuristic networking paradigms and change the way we design network protocols. We highlight this with the example of a pervasive sensor network in section 8.3. Finally, in section 8.4 we outline other promising research directions that is sprung from this research.

## 8.2 Thesis Summary

In the first part of the thesis, we focus on *mobile infostation networks*, a new kind of mobile ad hoc network that exploits node mobility in packet transmission. In a mobile infostation network, any two nodes communicate when they are in proximity. Under this transmission constraint, any pair of nodes is intermittently connected as mobility shuffles the node locations. We have addressed three important problems in this thesis, namely the effect of node noncooperation, transmit range and node mobility on the network performance, which are covered in depth in chapter 2,3 and 4.

Chapter 2 addresses the issue of node noncooperation in mobile infostation networks in the context of a content distribution application. All nodes have common interest to all files cached in the fixed infostations. In addition to downloading files from the

fixed infostations, nodes act as mobile infostations and exchange files when they are in proximity. We stipulate a social contract such that an exchange occurs only when each node can obtain something it wants from the exchange. Our social contract enables much higher system efficiency compared to downloading from fixed infostations only while not requiring true cooperation among nodes. We show by analysis and simulations that network performance depends on the node density, mobility and the number of files that are being disseminated. Our results point to the existence of data diversity for mobile infostation networks. The achievable throughput increases as the number of files of interest to all users increases. We have also extended the common interest model to the case where nodes have dissimilar interests. Our simulation results show that as mobile nodes change from having identical interests to mutually exclusive interests, the network performance degrades dramatically. We propose an alternative user strategy when nodes have partially overlapping interests and show that the network capacity can be significantly improved by exploiting multiuser diversity inherent in mobile infostation networks. We conclude that data diversity and multiuser diversity exist in noncooperative mobile infostation networks and can be exploited.

In chapter 3, we study the effect of transmit range on the capacity per unit area for four transmission strategies. We show that a stipulated transmit range improves the capacity compared to the rate adaptive Grossglauser-Tse strategy with an unconstrained transmit range by 25%, and outperforms the non-adaptive strategy by 68%. This indicates an optimally operated network involves trading off spatial transmission concurrency for more spectral efficiency on individual links. The capacity per unit area is derived explicitly for four transmission strategies. Numerical results show that the optimal number of neighbors is invariant to node density, and is between 0.6 to 1.2 in our transmission strategies. This result is in contrast to a magic number of 6 to 8 neighbors in multihop networks, where the expected forward progress per hop is maximized. This reflects the different optimization criteria of mobile infostation and multihop ad hoc networks. In addition, the capacity per unit area increases linearly with node density. This is counter-intuitive but can be explained using a rescaling argument drawn from

percolation theory. We also extend our results to practical systems with a specified signal to noise ratio (SIR) threshold. The invariance of the optimal number of neighbors to node density also applies here, and the corresponding packet success rate per unit area is also linearly increasing with node density. The optimal number of neighbors is also weakly dependent on the SIR threshold. It decreases gradually from 2 to 0.5 as the SIR threshold increases from 0 to 30dB.

In chapter 4, we address the effect of node mobility on highway mobile infostation networks. Each node enters a highway segment at a Poisson rate with a random speed drawn from a known but arbitrary distribution. Since nodes have different speed, a node may overtake other nodes or be overtaken as time evolves. Using arguments from renewal reward theory, the long run fraction of time an observer node is connected, and the long run average data rate can be derived and are functions of the observer node speed. We consider both forward traffic scenarios, in which two nodes moving in the same direction have a transient connection when they are within range from each other, and reverse traffic scenarios in which two nodes traveling in opposite directions are connected transiently when they are in range. For node speed that is uniformly distributed, we reveal that the expected fraction of connection time, or expected number of connections in queuing terminology, is independent of the observer node speed in reverse traffic. In forward traffic, on the other hand, the fraction of connection time increases with observer speed. That is, the network performance improves with node mobility, which is unique to the mobile infostation networking paradigm.

In the second part part of the thesis, we focus on generic mobile ad hoc networks, otherwise known as *packet radio networks* or *multihop ad hoc networks*, in which mobile nodes communicate in multihop routing. My focus is on the network layer of the protocol stack, including work on power control and network behavior. These topics are covered in chapter 5 and 6 of this thesis.

In mobile ad hoc networks, it is often more important to optimize for energy efficiency than throughput. In chapter 5, we investigate the effect of transmit range on energy efficiency of packet transmissions. We determine a common range for all nodes such that the average energy expenditure per received packet is minimized. In the first

part of this chapter, we consider stationary networks. We show that energy efficiency depends on various system parameters that includes path loss exponent of the channel, energy dissipation model and network offered load. In particular, when the path loss exponent is large, energy efficiency decreases when the transmit range increases. Hence, the network should be operated at the critical range that just maintains network connectivity. However, when the path loss exponent is small, operating at the critical range yields inferior throughput and energy efficiency. Our results show that energy efficiency is intimately connected to network connectivity. Three network connectivity regimes are identified as the transmit range of all nodes increases. In the second part, we examine the effect of node mobility on energy efficiency. We show that at normal offered load, an optimal transmit range exists such that energy efficiency is maximized. The optimal range turns out to be insensitive to node mobility, and is much larger than the critical range. We show that the energy expenditure can be reduced by 15% to 73% in different mobility scenarios, if the network is operated at the optimal range.

In chapter 6, we examine the network behavior of a routing algorithm for multihop ad hoc networks. Extensive simulations are performed using *ns-2* in a variety of mobility scenarios and offered load regimes. In the literature, performance metrics (goodput, delay and path length) are often obtained through ensemble averaging of many flows. Here we advocate an alternate graphical interpretation of simulation results similar to that used by Holland and Vaidya. Performance metrics of individual monitored flows are plotted instead. By identifying the correlations between performance metrics and system parameters, inter-relationships between them are revealed. For example, we have shown that path length is dependent on system parameters such as mobility, offered load and even the node distribution. These observations often give us insights to the mechanisms that underlie the network behavior. In particular we have resolved a conjecture that goodput improvement under high mobility is due to the load balancing effect. We show that at high mobility, goodput improvement for heavy offered load regimes is a consequence of the reduction of path length in the flows. Furthermore, we have introduced the concept of fraction of congested flows as a new performance metric. This and some other metrics such as fairness can be visualized from our graphs and are

important in characterizing network performance.

In the final part of the thesis, we present a wireless application for a mobile cellular network, an online Dutch auction price setting algorithm. In a Dutch auction, the price of an item decreases incrementally from the starting price at regular intervals. A bidder may buy the item at any time and stop the auction at the current price. Chapter 7 presents an optimal price decrement strategy in a Dutch auction, such that the expected revenue of the auction host is maximized. Properties of the optimal solution and a simple iterative solution methodology are discussed. Numerical studies show that significant gain could be obtained compared with a simple reference strategy.

## 8.3 Communications in Pervasive Sensor Networks

In future pervasive computing environments, our living environment will be filled with sensor nodes. These sensor nodes will have a variety of functionalities to detect different types of signal in our environment. They are connected together to form a *pervasive sensor network*, monitoring our living environment and feeding important information back to the *communication sink* for processing. The processed data are then fed to software applications which take advantage of embedded contextual information such as local microclimate, user location and activity to make intelligent decisions. While there is much research on the development of context-aware applications for pervasive computing, the full potential of these applications can only be realized when inexpensive low-power miniature sensor nodes can be deployed. The Smart Dust project [4, 37] of Berkeley, for instance, focuses on prototyping miniature sensor nodes called *motes*, and explores the limits on their size and power consumption. Although the project has a modest goal of fabricating millimeter scale sensor devices of a cubic volume, it is envisioned that future sensor nodes will be small enough to be freely suspended in air and buoyed by air currents.

The original proposal for Smart Dust [36] advocates the use of active/passive optical communications or multihop wireless communications. The relative merits of optical and wireless communications can be found in [98]. More recently, Kristofer Pister of

UCB, who is the PI of the Smart Dust project, favored multihop wireless communications over optical communications [66, 67]. Multihop wireless transmissions are more energy efficient than single hop transmissions to the communication sink. It can also be implemented much more easily than an optical communication system which requires the proper alignment of mobile transmitter and receiver nodes. To date, the Smart Dust project is a proven concept with real hardware prototypes [5] and real network applications.

As more network applications are being developed on Smart Dust networks, it is expected that the volume of traffic will dramatically increase. Nevertheless, the capacity of multihop networks is not scalable to network size. The results of Gupta and Kumar [24] implies that the achievable throughput of a sensor node to the base station goes to zero in a dense deployment of sensor nodes. To reduce the effective amount of traffic in a sensor network, data fusion techniques have been proposed to aggregate correlated data from proximate sensor nodes.

Our thesis is that by exploiting node mobility of the sensor nodes, we can significantly increase the network capacity to accommodate future increase in traffic. Due to the small size of the Smart Dust motes, sensor nodes exhibit inherent node mobility as they are suspended in the air. The transportation effect of winds and currents on sensor nodes can be used to physically carry the sensor data towards the base station. More generally, mobility of external devices can be exploited. Rather than relying on random mobility to toss the nodes in the right direction, external mobile nodes may be deployed to collect the processed data from local sensor nodes and direct these information back to the base station. These external nodes are synthetic microrobots with wireless communication circuits and sensors on board, and are being developed today in research labs to aid research in sensor networks and ad hoc networks. In USC the microrobots are called Robomotes [83] and in Berkeley they are called synthetic insects [81]. In some more aggressive systems [71], the microrobots may even harvest their own energy, recharge, and deliver energy to other energy depleted sensors.

Besides yielding much higher network capacity, the exploitation of node mobility in sensor node communications also yields higher energy efficiency. Although multihop

wireless communication compares favorably to single hop wireless communication in energy efficiency, the power consumption of is still significant. In Smart Dust motes, the reception of packets consumes energy comparable to that of a packet transmission. In particular, [70] shows that the power consumption in receiving data can even be greater than that of a packet transmission. This is because in a dense sensor network, the transmit range is usually small. A small transmit power suffices in most scenarios whereas a fixed power is expended in packet reception. Extensive multihop forwarding in a dense sensor network leads to significant energy consumption in packet reception. Although data fusion techniques can be used to minimize the amount of local data that needs to be forwarded, it involves heavy signal processing, another major component in the energy budget. On the other hand, in the mobile infostation paradigm, each packet is only relayed via no more than two hops, thus saving precious energy in packet reception. Moreover, since a node transmits at a much lower power in a mobile infostation network [106] than that in a multihop network, significant energy can be saved.

## 8.4 Other Research Directions

In traditional communication paradigms, *node mobility* is undesirable and contributes to performance degradation to wireless networks. This is indeed the case for multihop ad hoc networks [8, 12, 50]. As a result of node mobility, network topology changes dynamically as time evolves. In proactive routing algorithms, frequent topology changes induce a lot of update overhead of routing tables. In reactive routing algorithms, similarly, route maintenance overhead also increases as a result of node mobility. It is somewhat surprising that high node mobility improves the total proportion of time that the node is connected, and thus the data rate [105,106]. That is, the mobile infostation paradigm is robust to node mobility. In applications where users have high mobility, communication using the mobile infostation paradigm is therefore an attractive solution. Our results may have important implications for communications in *highway vehicular networks* [53, 80].

On the other hand, our study on node noncooperative issues in a mobile infostation

network may also find broader application in *peer-to-peer computing networks*. The success of free file sharing networks has catapulted peer-to-peer networks to the limelight in recent years. Nevertheless, the *freeriding problem* is an important issue. Since the service is free, some peers take advantage of the network resources without contributing to it. The proliferation of freeriders in the network inadvertently creates a lot of traffic and place heavy burden to those peers who actively contribute to the network. This freeriding problem has some striking resemblances to the noncooperative content distribution problem presented in chapter 2. Peers in the network are not incentivized to cooperate since a node consumes its network resources (bandwidth) by sharing his files to other peers in the network. In this aspect, the *social contract* defined in the context of noncooperative mobile infostation networks may be useful to induce implicit cooperation between peers in the network without the need of policing. It is desirable to examine the network performance of a peer-to-peer network if all network nodes observe a stipulated social contract.

# References


[1] http://www.isi.edu/nsnam/ns/. Network Simulator Notes and Documentation, the VINT Project, UC Berkeley, LBL, USC/ISI and Xerox PARC.

[2] http://www.monarch.cs.rice.edu/. Rice University Monarch Project, Mobility and Wireless Extensions to ns.

[3] http://www.agorics.com/new.html.

[4] http://robotics/eecs.berkeley.edu/p̃ister/SmartDust/. Smart Dust project page.

[5] http://www.xbow.com/. Crossbow Technology Inc.

[6] S. Agarwal, S. Krishnamurthy, R.H. Katz, and S.K. Dao. Distributed power control in ad-hoc wireless networks. In *Proc. IEEE PIMRC '01*, 2001.

[7] D.P. Bertsekas. *Dynamic Programming: Deterministic and Stochastic Models*. Prentice-Hall, Englewood Cliffs, N.J., 1987.

[8] J. Broch, D.A. Maltz, D.B. Johnson, Y.C. Hu, and J. Jetcheva. A performance comparison of multi-hop wireless ad hoc network routing protocols. In *Proc. Mobicom*, pages 85–97, 1998.

[9] S.M. Cherry. Wi-fi takes new turn with "wireless-g". *IEEE Spectrum Magazine*, pages 12–13, August 2003.

[10] K.P. Chong and H.Z. Stanislaw. *An Introduction to Optimization*. Wiley, New York, 1996.

[11] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

[12] S.R. Das, C.E. Perkins, E.M. Royer, and M.K. Marina. Performance comparison of two on-demand routing protocols for ad hoc networks. *IEEE Personal Communications*, 8(1):16–28, Feb. 2001.

[13] R. Dube, C.D. Rais, K.-Y. Wang, and S.K. Tripathi. Signal stability-based adaptive routing (ssa) for ad hoc mobile networks. *IEEE Personal Communications*, 4(1):36–45, 1997.

[14] T.A. Elbatt, S.V. Krisnamurthy, D. Connors, and S. Dao. Power management for throughput enhancement in wireless ad hoc networks. In *Proc. IEEE ICC '00*, pages 1503–1513, 2000.

[15] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume vol. I. Wiley, New York, 1967.

[16] R. H. Frenkiel, B. R. Badrinath, J. Borras, and R. Yates. The infostations challenge: Balancing cost and ubiquity in delivering wireless data. *IEEE Personal Communications*, 7(2):66–71, April 2000.

[17] R. H. Frenkiel, B. R. Badrinath, J. Borras, and R. Yates. The infostations challenge: Balancing cost and ubiquity in delivering wireless data. *IEEE Personal Communications*, 7(2):66–71, April 2000.

[18] R. G. Gallager. *Information Theory and Reliable Communication*. John Wiley & Sons, New York, NY, 1968.

[19] D. J. Goodman, J. Borras, N. B. Mandayam, and R. D. Yates. Infostations : A new system model for data and messaging services. In *Proceedings of IEEE VTC'97*, volume 2, pages 969–973, May 1997. Phoenix, AZ.

[20] M. Grossglauser and D. Tse. Mobility increases the capacity of ad-hoc wireless networks. In *Proceedings of IEEE INFOCOM '01*, volume 3, pages 1360–1369, 2001.

[21] M. Grossglauser and M. Vetterli. Locating nodes with ease: Mobility diffusion of last encounters in ad hoc networks. In *Proc. IEEE INFOCOM '2003*, 2003.

[22] P. Gupta and P. Kumar. A system and traffic dependent adaptive routing algorithm for ad hoc networks. In *Proceedings of the 36th IEEE Conference on Decision and Control*, volume 3, pages 2375–2380, 1997.

[23] P. Gupta and P.R. Kumar. Critical power for asymptotic connectivity in wireless networks. *Stochastic Analysis, Control, Optimization and Applications: A Volume in Honor of W.H. Fleming*, pages 547–566, 1998.

[24] P. Gupta and P.R. Kumar. The capacity of wireless networks. *IEEE Trans. on Info. Theo.*, 46(2):388–404, 2000.

[25] Z.J. Haas and M.R. Pearlman. The performance of query control schemes for the zone routing protocol. *IEEE Trans. on Net.*, 9(4):427–438, August 2001.

[26] G. Holland and N. Vaidya. Analysis of tcp performance over mobile ad hoc networks. In *Proceedings of the fifth annual ACM/IEEE international conference on Mobile computing and networking*, August 1999.

[27] G. Holland, N. Vaidya, and P. Bahl. A rate-adaptive mac protocol for multihop wireless networks. In *The seventh annual international conference on Mobile computing and networking*, July 2001.

[28] T.C. Hou and V.O.K. Li. Transmission range control in multihop packet radio networks. *IEEE Trans. Commun.*, 34(1):38–44, 1986.

[29] A.L. Iacono and C. Rose. Mine, mine, mine: information theory, infostation networks, and resource sharing. In *Proceedings of IEEE Wireless Communications and Networking Confernce*, volume 3, pages 1541–1546, 2000.

[30] A. Iwata, C.-C. Chiang, G. Pei, M. Gerla, and T.-W. Chen. Scalable routing strategies for ad hoc wireless networks. *IEEE J. Sel. Areas Commun.*, 17(8):1369–1379, August 1999.

[31] R. Jain, A. Puri, and Sengupta. R. Geographical routing using partial information for wireless ad hoc networks. *IEEE Personal Communications*, 8(1):48–57, Feb. 2001.

[32] P. Johanson, T. Larsson, N. Hedman, B. Mielczarek, and M. Degermark. Scenario-based performance analysis of routing protocols for mobile ad hoc networks. In *Proceedings of the ACM/IEEE International Conference on Mobile Computing and Neetworking (Mobicom)*, pages 195–206, August 1999.

[33] D.B. Johnson and D.A. Maltz. Dynamic source routing in ad hoc wireless networks. In *Mobile Computing*, chapter 5, pages 153–181. Kluwer Academic Publishers, 1996.

[34] P. Juang, H. Oki, Y. Wang, M. Martonosi, L.-S. Peh, and D. Rubenstein. Energy-efficient computing for wildlife tracking: Design tradeoffs and early experiences with zebranet. In *ACM SIGARCH Computer Architecture News*, volume 30,5, December 2002.

[35] J. Jubin and J.D. Tornow. The darpa packet radio network protocols. *Proceedings of the IEEE*, 75(1):21–32, 1987.

[36] J.M. Kahn, R.H. Katz, and K.S.J. Pister. Next century challenges: Mobile networking for "smart dust". In *Proceedings of ACM MobiCom*, pages 94–100, Sep. 1991.

[37] J.M. Kahn, R.H. Katz, and S.J. Pister. Emerging challenges: Mobile networking for "smart dust". *Journal of Communications and Networks*, 2(3):188–196, September 2000.

[38] R.E. Kahn. The organization of computer resources into a packet radio network. *IEEE Trans. Commun.*, 25(1):169–178, 1977.

[39] R.E. Kahn. Advances in packet radio technology. *Proceedings of the IEEE*, 66(11):1468–1496, 1978.

[40] P.R. et. al. Karn. Packet radio in the amateur service. *IEEE J. Sel. Areas Commun.*, 3(3):431–439, 1985.

[41] V. Kawadia, S. Narayanaswamy, R. Rozovsky, R.S. Sreenivas, and P.R. Kumar. Protocols for media access control and power control in wireless networks. In *Proc. IEEE Decision and Control '01*, pages 1935–1940, 2001.

[42] J.F.C. Kingman. *Poisson Processes*. Oxford University Press, New York, 1993.

[43] L. Kleinrock. Principles and lessons in packet communications. *Proceedings of the IEEE*, 66(11):1320–1329, 1978.

[44] L. Kleinrock and J. Silvester. Optimum transmission radii for packet radio networks or why six is a magic number. In *Proceedings of the IEEE National Telecommunications Conference*, pages 4.3.1–4.3.5, December 1978.

[45] L. Kleinrock and J. Silvester. Spatial reuse in multihop packet radio networks. *invited paper for Proceedings of the IEEE, Special Issue on Packet Radio Networks*, 75(1):156–167, January 1987.

[46] P. Klemperer. Auction theory: A guide to the literature. *Journal of Economical Surveys*, 13(3):227–284, 1999.

[47] Y.-B. Ko and N.H. Vaidya. Location-aided routing (lar) in mobile ad hoc networks. *Wireless Networks*, 6(4), July 2000.

[48] B.M. Leiner, D.L. Nielson, and F.A. Tobagi. Issues in packet radio network design. *Proceedings of the IEEE*, 75(1):6–20, 1987.

[49] L. Li, V. Bahl, Y.M. Wang, and R. Wattenhofer. Distributed topology control for power efficient operation in multihop wireless ad hoc networks. In *Proc. IEEE INFOCOM '01*, April 2001.

[50] D.A. Maltz, J. Broch, J. Jetcheva, and D.B. Johnson. The effects of on-demand behavior in routing protocols for multihop wireless ad hoc networks. *IEEE J. Sel. Areas Commun.*, 17(8):1439–1453, August 1999.

[51] R. Meester and R. Roy. *Continuum Percolation*. Cambridge University Press, 1996.

[52] P.R. Milgrom and R.J. Weber. A theory of auctions and competitive bidding. *Econometrica*, 50(5):1089–1122, Sep. 1982.

[53] R. Morris, J. Jannotti, F. Kaashoek, J. Li, and D. Decouto. Carnet: A scalable ad hoc wireless network system. In *Proc. ACM SIGOPS European Workshop*, Sep. 2000.

[54] S. Murthy and J.J. Garcia-Luna-Aceves. An efficient routing protocol for wireless neetworks. In *Proc. ACM Mobile Networks and App. J., Special Issue on Routing in Mobile Communication Networks*, pages 183–197, 1996.

[55] R.B. Myerson. Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73, Feb. 1981.

[56] R. Nelson and L. Kleinrock. The spatial capacity of a slotted aloha multihop packet radio network with capture. *IEEE Trans. Commun.*, 32(6):684–694, June 1984.

[57] M.E. Oren and A.C. Williams. On competitive bidding. *Operations Research*, 23:1072–1079, 1975.

[58] M. Papadopouli and H. Schulzrinne. Seven degrees of separation in mobile ad hoc networks. In *Proc. IEEE Globecom '00*, 2000.

[59] M. Papadopouli and H. Schulzrinne. Effects of power conservation, wireless coverage and cooperation on data dissemination among mobile sevices. In *Proc. IEEE MobiHoc '01*, 2001.

[60] V.D. Park and M.S. Corson. A highly adaptive distributed routing algorithm for mobile wireless networks. In *Proc. INFOCOM '97*, April 1997.

[61] M.D. Penrose. A strong law for the longest edge of the minimal spanning tree. *The Annals of Probability*, 27(1):246–260, 1999.

[62] C. Perkins and P. Bhagwat. Highly dynamic destination-sequenced distance-vector routing. In *Proc. ACM Sigcomm 94*, pages 234–244, 1994.

[63] C. Perkins and E.M. Royer. Ad hoc on demand distance vector (aodv) routing. In *Proc. 2nd IEEE Workshop o n Mobile Comp. Sys. and Apps.*, pages 90–100, Feb. 1999.

[64] T. Philips, S. Panwar, and A. Tantawi. Connectivity properties of a packet radio network model. *IEEE Trans. on Info. Theo.*, 35(5), 1989.

[65] P. Piret. On the connectivities of radio networks. *IEEE Trans. on Info. Theo.*, 37(5), 1991.

[66] K. Pister. Smart dust - hardware limits to wireless sensor networks. IEEE ICDCS 2003 Keynote Address.

[67] K. Pister. Smart dust and micro robots. Microsoft Multi-University/Research Laboratory Seminar Series, 5/17/01, http://murl.microsoft, com/.

[68] John G. Proakis. *Digital Communications*. McGraw-Hill International Editions, 1995.

[69] M.B. Pursley, H.B. Russell, and J.S. Wysocarski. Efficient routing in frequency-hop radio networks with partial-band interference. In *Proceedings of IEEE Wireless Communications and Networking Confernce*, volume 1, pages 79–83, 2000.

[70] V. Raghunathan, C. Schurgers, S. Park, and M.B. Srivastava. Energy-aware wireless microsensor networks. *IEEE Signal Processing Magazine*, 19(3):40–50, March 2002.

[71] M. Rahimi, H. Shah, G. Sukhatme, J. Heidemann, and D. Estrin. Energy harvesting in mobile sensor networks. In *Proceedings of the IEEE International Conference on Robotics and Automation*, May 2003.

[72] R. Ramanathan and Rosales-Hail. Topology control of multihop wireless networks using transmit power adjustment. In *Proc. IEEE INFOCOM '00*, pages 404–413, 2000.

[73] S. Ramanathan and M. Steenstrup. A survey of routing techniques for mobile communications networks. *Mobile Networks and Applications*, 1(2), 1996.

[74] G. Riley and W.F. Samuelson. Optimal auctions. *American Economic Review*, 71(3):381–392, 1981.

[75] L.G. Roberts. The evolution of packet switching. *Proceedings of the IEEE*, 66(11):1307–1313, 1978.

[76] S.M. Ross. *Stochastic Processes*. John Wiley & Sons, New York, 1983.

[77] S.M. Ross. *Introduction to Probability Models*. Academic Press, London, 2000.

[78] E.M. Royer, P.M. Melliar-Smith, and L.E. Moser. An analysis of the optimum node density for ad hoc mobile networks. In *Proc. International Conference on Communications*, 2001.

[79] E.M. Royer and C.-K. Toh. A review of current routing protocols for ad hoc mobile wireless networks. *IEEE Personal Communications*, 6(2):46–55, 1999.

[80] M. Rudack, M. Meincke, and M. Lott. On the dynamics of ad hoc networks for inter vehicles communicaitons (ivc). In *Proc. ICWN '02*, 2002.

[81] L. Schenato, X. Deng, and S. Sastry. Flight control system for a micromechanical flying insect: Architecture and implementation. In *IEEE International Conference on Robotics and Automation*, May 2001.

[82] N. Shacham and J. Westcott. Future directions in packet radio architectures and protocols. *Proceedings of the IEEE*, 75(1):83–98, 1987.

[83] G.T. Sibley, M.R. Rahimi, and S. Sukhatme. Robomote: A tiny mobile robot platform for large-scale ad-hoc sensor networks. In *IEEE International Conference on Robotics and Automation*, September 2002.

[84] S. Singh, M. Woo, and C.S. Raghavendra. Power-aware routing in mobile ad hoc networks. In *Proc. ACM/IEEE MOBICOM '98*, October 1998.

[85] T. Small and Z.J. Haas. The shared wireless infostation model - a new ad hoc networking paradigm (or where there is a whale, there is a way). In *Proc. IEEE MobiHoc '03*, 2003.

[86] E. S. Sousa. The performance of a link in a poisson field of interferers. *IEEE Trans. on Info. Theo.*, 38(6):1743–1754, Nov. 1992.

[87] E.S. Sousa and J.A. Silvester. Optimum transmission ranges in a direct-sequence spread-spectrum multihop packet radio network. *IEEE J. Sel. Areas Commun.*, 8(5):762–771, 1990.

[88] Gilbert Strang. *Linear algebra and its applications*. Harcourt, Brace, Jovanovich, San Diego, 1988.

[89] Gordon L. Stuber. *Principles of Mobile Communication*. ress, 1996.

[90] M.W. Subbarao and B.L. Hughes. Optimal transmission ranges and code rates for frequency-hop packet radio networks. *IEEE Trans. Commun.*, 48(4):670–678, April 2000.

[91] H. Takagi and L. Kleinrock. Optimal transmission ranges for randomly distributed packet radio terminals. *IEEE Trans. Commun.*, 32(3):246–257, March 1984.

[92] F.A. Tobagi. Modeling and performance analysis of multihop packet radio networks. *Proceedings of the IEEE*, 75(1):135–155, 1987.

[93] C.-K. Toh. A novel distributed routing protocol to support ad-hoc mobile computing. In *Proc. IEEE Fifteenth Annual International Phoenix Conference on Computers and Communications*, pages 480–486, 1996.

[94] C.-K. Toh. Associativity-based routing for ad-hoc mobile networks. *Wireless Personal Communication Journal, Special issue on mobile networking and computing systems*, March 1997.

[95] W. Vickrey. Counterspeculation, auctions and competitive sealed tenders. *Journal of Finance*, 16:8–37, March 1961.

[96] M. Weiser. "the computer for the twenty-first century". In *Scientific American*, 1991.

[97] R. Yates and M.B. Mandayam. Challenges in low-cost wireless data transmission. *IEEE Signal Processing Magazine*, 17(2):93–102, May 2000.

[98] W.H. Yuen. Communications in pervasive computing systems. submitted for publication.

[99] W.H. Yuen, H.-N. Lee, and T.D. Andersen. A simple and effective cross layer networking system for mobile ad hoc networks. In *Proceedings of IEEE PIMRC*, 2002.

[100] W.H. Yuen and C.W. Sung. On energy efficiency and network connectivity of mobile ad hoc networks. In *to appear, Procceedings of IEEE International Conference in Distributed Computing Systems*, 2003.

[101] W.H. Yuen and R.D. Yates. Optimum transmit range and capacity of mobile infostation networks. In *Procceedings of ACM MobiHoc, poster paper*, 2003.

[102] W.H. Yuen and R.D. Yates. Optimum transmit range and capacity of mobile infostation networks. In *to appear, Procceedings of IEEE GLOBECOM 2003*, 2003.

[103] W.H. Yuen, R.D. Yates, and S.-C. Mau. Exploiting data diversity and multiuser diversity in mobile infostation networks. In *Proceedings of IEEE INFOCOM*, 2003.

[104] W.H. Yuen, R.D. Yates, and S.-C. Mau. Noncooperative content distribution in mobile infostation networks. In *Proceedings of IEEE WCNC*, 2003.

[105] W.H. Yuen, R.D. Yates, and C.W. Sung. Effect of node mobility on highway mobile infostation networks. In *to appear, Proceedings of ACM International Workshop on Modeling, Analysis and Simulations of Wireless and Mobile Systems*, Sep. 2003.

[106] W.H. Yuen, R.D. Yates, and C.W. Sung. Performance evaluation of highway mobile infostation networks. In *to appear, Procceedings of IEEE GLOBECOM 2003*, 2003.

[107] M. Zorzi and S. Pupolin. Optimum transmission ranges in multihop packet radio networks in the presence of fading. *IEEE Trans. Commun.*, 43(7), July 1995.

# Vita

## Wing Ho Andy Yuen

**1995** Bachelor degree in Electrical Engineering, Hong Kong University of Science and Technology, Hong Kong.

**1995-97** Teaching Assistant, Department of Information Engineering, Chinese University of Hong Kong, Hong Kong.

**1997** Master of Philosophy degree in Information Engineering, Chinese University of Hong Kong, Hong Kong.

**1999** Teaching Assistant, Department of Electrical and Computer Engineering, Rutgers University, Piscataway, New Jersey.

**2000** Research Assistant, Chinese University of Hong Kong, Hong Kong.

**2001** Summer Intern, HRL Laboratories, Malibu, California.

**2002** Senior Research Assistant, City University of Hong Kong, Hong Kong.

**2000-03** Graduate Research Assistant, WINLAB, Department of Electrical and Computer Engineering, Rutgers University, Piscataway, New Jersey.

**1997** Wing Ho A. Yuen and Wing Shing Wong, "A Hybrid Bloom Filter Location Update Algorithm for Wireless Cellular Systems," *International Conference on Communications, Proc. ICC '97.*

**1998** Wing Ho A. Yuen and Wing Shing Wong, "A Dynamic Location Area Assignment Algorithm for Mobile Cellular Systems," *International Conference on Communications, Proc. ICC '98.*

**2001** Wing Ho A. Yuen and Wing Shing Wong, "A Hybrid Contention Free Location Update Strategy with Probabilistic Paging for PCS," *Transactions on Vehicular Technology Vol.50 Issue: 1, Jan 2001.*

**2002** Wing Ho Yuen, Heung-no Lee and Timothy D. Andersen, "A Rate Adaptation Scheme and Interference Aware Routing Algorithm for Ad Hoc Networks," *IEEE PIMRC 2002*

**2002** Wing Ho Yuen and Roy D. Yates, "Inter-relationships of Performance Metrics and System Parameters in Mobile Ad Hoc Networks," *IEEE MILCOM 2002*

**2002** Wing Ho Yuen, Chi Wan Sung and Wing Shing Wong, "Optimal Price Decremental Strategy for Dutch Auctions, " *Communications in Information and Systems, Volume 2, Number 4, Dec. 2002.*

**2003** Wing Ho Yuen, Roy D. Yates and Siun-Chuon Mau, "Noncooperative Content Distribution in Mobile Infostation Networks," *IEEE WCNC 2003*

**2003** Wing Ho Yuen, Roy D. Yates and Siun-Chuon Mau, "Exploiting Data Diversity and Multiuser Diversity in Noncooperative Mobile Infostation Networks," *IEEE INFOCOM 2003* (acceptance rate $< 21\%$)

**2003** Wing Ho Yuen and Chi Wan Sung, "On Energy Efficiency and Network Connectivity for Mobile Ad Hoc Networks," *IEEE ICDCS 2003* (acceptance rate $<18\%$)

**2003** Wing Ho Yuen and Roy D. Yates, "Optimum Transmit Range and Capacity of Mobile Infostation Networks," *poster paper in ACM MOBIHOC 2003 and to appear in ACM Mobile Computing and Communications Review*

**2003** Wing Ho Yuen, Roy D. Yates and Chi Wan Sung, "Effect of Node Mobility on Highway Mobile Infostation Networks," *to appear in ACM MSWIM 2003* (acceptance rate $< 15\%$)

**2003** Wing Ho Yuen and Roy D. Yates, "Optimum Transmit Range and Capacity of Mobile Infostation Networks," *to appear in IEEE GLOBECOM 2003*

**2003** Wing Ho Yuen and Roy D. Yates, "Performance Evaluation of Highway Mobile Infostation Networks," *to appear in IEEE GLOBECOM 2003*

**2003** Ph.D. in Electrical and Computer Engineering, Rutgers University, Piscataway, New Jersey.